Nanotechnology

Edited by Günter Schmid

Wiley-VCH

Contents

1
Introduction

Günter Schmid

The term "Nanotechnology" is nowadays commonplace not only in all relevant scientific and technical areas, but also to a considerable extent in the public domain, based on reports in newspapers, on television and, justified or not, in a series of commercially available products with "nano" as part of their names. On the one hand, this development could be considered in a positive sense, indicating nanotechnology as an accepted new technology. On the other hand, it contains some risks that should not be neglected. This is due to the rather complex definition of nanotechnology and nanoscience as a sectional science, involving natural and materials sciences, engineering and medicine. Especially it is the lack of a generally accepted definition of nanotechnology and nanoscience that is responsible for many misunderstandings. The relevant scientific communities have agreed that the term "nano" must always be linked with the appearance of a novel property. If "nano" is restricted just to a length scale, one would preferably speak of "technology on the nanoscale", usually only based on scaling effects ranging from micrometer to nanometer dimensions, without being linked with the appearance of really novel physical or chemical properties. This imprecise view of nanotechnology is frequently misused for products that are linked with the term "nano", but do not really offer a "nano-effect".

The following chapters therefore deal with the principles and fundamentals of nanotechnology, explaining what nanoscience and nanotechnology really means and what it does not mean. Furthermore, this book contains philosophical and ethical aspects, since any new technology opens up questions concerning social consequences. Therefore, first of all, a scientifically unambiguous definition of nanotechnology and nanoscience is discussed in Chapter 2, followed by a series of examples elucidating this definition in various fields, reaching from size effects up to complex biosystems. Chapter 3 deals with the principles of how to generate effective nanosystems. Top-down techniques are completed by bottom-up procedures that are currently becoming increasingly important due to the use of ultimately small building blocks: atoms and molecules. Chapters 4 and 5 consider two kinds of fundamental objects of nanoscience: quantum dots, and wires, rods and tubes, respectively. Those species represent the world of size-determined properties of

Nanotechnology. Volume 1: Principles and Fundamentals. Edited by Günter Schmid

ISBN: 978-3-527-31732-5

manifold materials and so stand for one of the fundamental principles of nanotechnology, in agreement with the definition. Spherical or one-dimensional matter of appropriate size can no longer be described by classical physical laws, but by quantum mechanical rules, indicating the decisive change from the macroscopic or microscopic world to the nanoworld.

An extremely important field of nanoscience and also of nanotechnology is dealing with the intelligent combination of artificial nanoscopic building blocks with biomolecular systems, which can anyway be considered as the most powerful "nanotechnological" inventions that we know. Most building blocks of living cells represent perfect nanosystems, the interplay of which results in the microscopic and macroscopic world of cells. We have learned to learn from Nature and consequently try to develop technologically applicable devices reaching from novel sensor systems up to diagnostic and therapeutic innovations. Chapter 6 gives an insight into this fascinating part of the nanoworld.

Philosophical and ethical questions are discussed in Chapters 7 and 8. What kind of knowledge is produced and communicated by nanotechnology? What is its place in relation to other sciences? These and related problems are discussed in Chapter 7. Studying current and future developments in nanotechnology from the viewpoint of ethics is an essential requirement in order to elaborate rules and concerted actions on how to deal with them in society. Such reflections should accompany any novel technological development, especially nanotechnology, the power of which has already been compared with the beginning of a new genesis.

This is the first of a series of books dealing with the various fields of nanotechnology. In addition to the principles and fundamentals, treated in this volume, information technology, medicine, energy, tools and analytics as well as toxicity will be the subjects of subsequent other books. In all cases, developed fields of nanotechnology and future areas of nanotechnological applications will be described and discussed.

2
The Nature of Nanotechnology

Günter Schmid

2.1
Definition

Numerous definitions of "nanotechnology" exist in the literature. Most of them simply say that nanotechnology considers materials and architectures on the nanoscale. In some of the definitions it is stated that nanotechnology dealings mainly with structures in the region between 1 and 100 nm. In any case, the dimension plays the dominant role. For a more detailed description, however, this is much too simple. First, nanotechnology follows nanoscience, where fundamental effects have been discovered before. In a long course of development, a technology may result from scientific findings in some cases, but not in all cases by a long way. So, to understand nanotechnology, one has to define properly what nanoscience is.

Among these numerous attempts to define nanotechnology, the definition given by the Royal Society and the Royal Academy of Engineering is fairly close to the author's opinion on the matter [1]:

> "*Nanoscience is the study of phenomena and manipulation of materials at atomic, molecular and macromolecular scales, where properties differ significantly from those at a larger scale. Nanotechnologies are the design, characterization, production and application of structures, devices and systems by controlling shape and size at the nanometre scale.*"

A more recent definition, formulated by a team of scientists at the Europäische Akademie zur Erforschung wissenschaftlich-technischer Entwicklungen Bad Neuenahr-Ahrweiler GmbH, focuses still more the real intent of nanoscience and nanotechnology and will therefore be used in the following [2]:

> "*Nanotechnology comprises the emerging application of Nanoscience. Nanoscience deals with functional systems either based on the use of subunits with specific size-dependent properties or of individual or combined functionalized subunits.*"

Nanotechnology. Volume 1: Principles and Fundamentals. Edited by Günter Schmid

ISBN: 978-3-527-31732-5

This is the only existing definition without naming a particular lateral scale. It excludes any kind of simple scaling effect (see later). The decisive aspect is the appearance of novel properties. These can in principle be observed below 1 or above 1000 nm (1 μm). Therefore, the strict limitation to a distinct length scale does not appropriate seem. Scientifically, it would even be absurd to set a limit to size-dependent properties or novel properties of a construct of functionalized subunits. In spite of this deliberate leaving out of a particular lateral scale, the expressions nanoscience and nanotechnology are meaningful for practical use since most of the known "nano-effects" happen on the nanoscale. The limitation of the definition to the nanoscale, however, would degrade nanotechnology to the simple continuation of microtechnology. Microtechnology was and still is overwhelmingly successful by the continuous reduction of materials or tools, aiming not at the creation of novel abilities, but at other advantages. Some examples in the context of the above definition will help us to understand better what it is meant.

Typical *size-dependent* nano-effects that spontaneously occur when a critical dimension is reached are observed when metal particles are downsized. Depending on the kind of change of property, the critical size may vary for the same element.

A very typical and well known nano-effect is observed when gold is downsized. In the bulk state, the beautiful color of gold results from a very fundamental phenomenon, the relativistic effect, based on Einstein's Special Theory of Relativity. One of the basic messages of this theory is that the speed of light in vacuum is an absolute constant everywhere in the universe. It never can be surpassed. If an object were theoretically to be accelerated close to or even exactly to the speed of light, its mass would continuously increase with increasing speed, finally even *ad infinitum*.

Electrons in most atoms are moving around the atomic nucleus with speeds usually far below that of light. However, in very heavy atoms, the high mass of the nuclei accelerates electrons to such an extent that relativistic effects become obvious [3–5]. For instance, such effects are known for the elements lead, tungsten, mercury, platinum and gold. The interesting question is whether relativistic effects influence the physical and chemical properties of such a heavy element. Indeed they do! The acceleration primarily affects the s-electrons which are in the nuclear region, linked with an increase in their mass, reducing the average distance from the nucleus. Consequently, the s-orbitals and their energy are shrinking. As a secondary effect, d-electrons, being farer from the nucleus, become electronically better shielded. Their orbitals are therefore extended and increase energetically. For the p-electrons the effects approximately equilibrate. In bulk metals, the individual orbitals of atoms are extended to electronic bands. The valence band, resulting from the d-orbitals, is energetically increased, whereas the conduction band, formed from the s-orbitals, is lowered. The reduced energy difference between the valence band and conduction band finally allows the low-energy photons of blue light to lift electrons from the valence to the conduction bands. Consequently, gold absorbs blue light and shows the complementary color yellow. Relativistic effects are revealed in various ways: tungsten exhibits an unusually high melting point and mercury is the only metal that is liquid at normal temperatures. Figure 2.1 illustrates the relativistic effect of gold by shrinking the energy difference between the s- and d-orbitals.

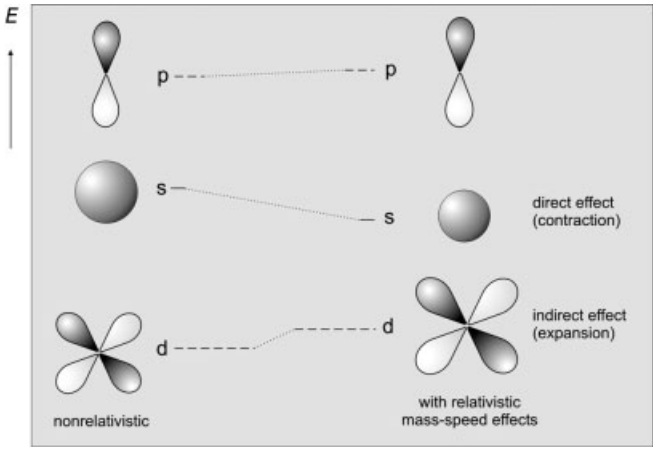


Figure 2.1 Influence of relativistic effects on the energy level of d-, s- and p-orbitals.

Since the relativistic effect is a property of each individual gold atom, it must also be present in nanosized particles, although their color is no longer golden. This metal changes its appearance at about 50 nm to become blue. Further reduction results in purple and finally, at about 15–20 nm, in bright red. This well-understood and long-known effect can be traced back to the existence of a so-called plasmon resonance. The phenomenon is quantitatively described by the Mie theory [6]. Qualitatively, the formation of a plasmon resonance can be explained by a collective electron oscillation with respect to the positive metal core of the particle, caused by the interaction of external electromagnetic radiation (visible light) with the confined electron gas of nanoparticles. The process is illustrated in Figure 2.2.

The energy taken up from light is responsible for the resulting color. This energy depends on, among others, the particle's size and shape and the surrounding medium. With decreasing size the color is shifted to shorter wavelengths. If the particles are not spherical but elongated, two plasmon bands may occur, one for the transverse and the other for the longitudinal resonance. Figure 2.3 shows the UV–visible spectra of spherical 18-nm gold nanoparticles (colloids) with an absorption maximum at 525 nm [7].

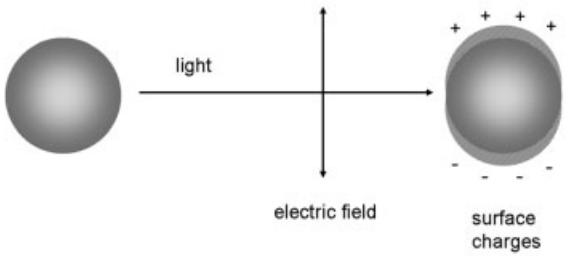


Figure 2.2 Formation of plasma resonance by the interaction of external electromagnetic radiation with the confined electron gas in a metallic nanoparticle.

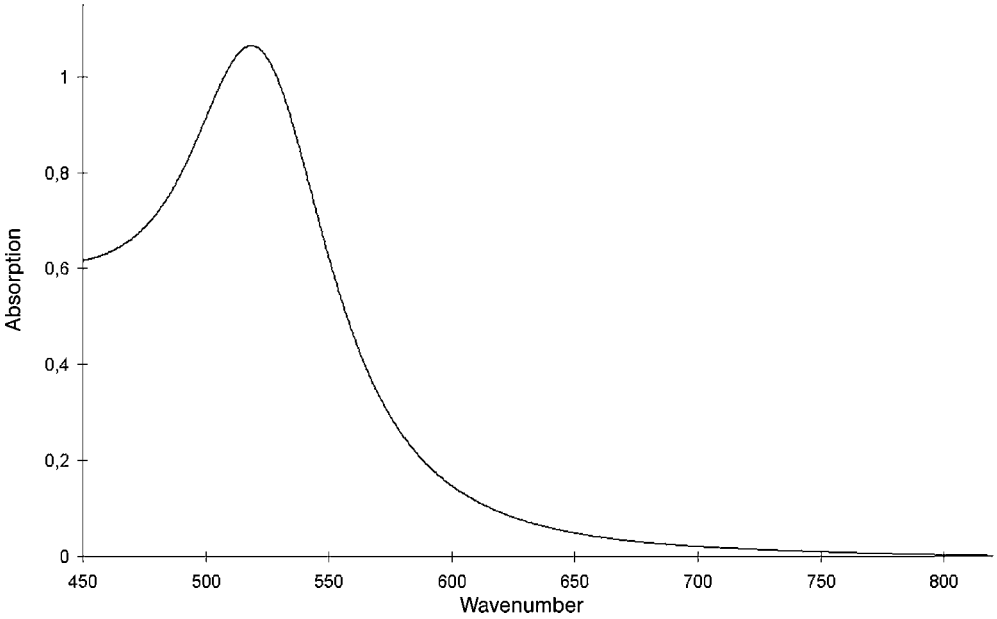


Figure 2.3 UV–visible spectrum of 18-nm gold particles.

Whereas the plasmon resonances of the three metals copper, silver and gold is in the visible region, for most other metals it is in the UV region and so cannot be observed with the naked eye.

The disappearance of the typical color in the case of gold when a critical size is reached and the appearance of blue or red colors simply means that the plasmon resonance superimposes the relativistic effect and so covers the typical color of bulk gold.

The second part of the above definition names the use of individual or combined functionalized subunits. As an *individual functionalized subunit* or building block, molecular switches may serve as an example. A molecular switch is a molecule that exists in two different states which can be adjusted by external stimuli. The two states must display different physical properties, each being stable with long lifetimes. If addressable by electrical or any other stimuli, such molecules could in principle serve as building blocks in future storage systems. For instance, catenanes consist of two interlocked rings equipped with electrochemically active parts (see Figure 2.4). Applied electric potentials cause Coulomb repulsion and make one of the rings move relative to the other, ending in another stable configuration. The sketch in Figure 2.4 elucidates the process [8].

Finally, an example of *combined functionalized subunits*, the last of the three conditions in the definition, can be given. Nature is a perfect nanoarchitect. Many parts of living cells can be considered as a combination of functionalized building blocks, although cells themselves have dimensions in the micro regime. The probably most exciting molecule in Nature is deoxyribonucleic acid (DNA) with its

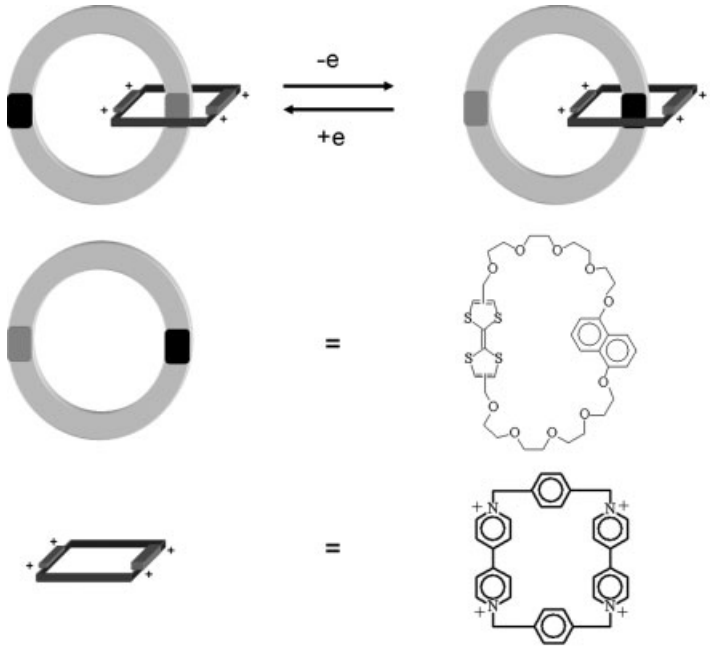


Figure 2.4 Sketch and molecular examples of a catenane-based switching device. The counteranions $[PF_6]^-$ have been omitted.

unique double helical structure. It consists of fairly simple subunits: four different heterocyclic bases, phosphate anions and pentose fragments. It is the shape complementarity of the bases that enables an almost infinite number of combinations to give base pairs, which finally encode the genome of any living system using hydrogen bonds to link complementary bases: thymine (T) combines only with adenine (A), cytosine (C) exclusively with guanine (G). The sugar fragments and the phosphates generate a backbone, holding the base pairs together. Figure 2.5 illustrates the decisive interaction between the four bases. The sequence of bases in one single strand determines the sequence in the other.

In practice, the number of different combinations in a human genome is infinite, considering the realistic length of a DNA double helix consisting of about 3.2 billion base pairs.

The above definition and the following elucidating examples clearly demonstrate that nanoscience and nanotechnology in a strictly scientific manner do not consider simple scaling effects. One speaks of scaling effects if a material is downsized from the macro/microscale even to the nanoscale and the properties, if at all, change continuously, but not spontaneously. In other words, characteristic properties are already present in the micro regime and only change gradually on reaching the nanoscale. A typical example will help to understand easily what a scaling effect is.

The well-known moth-eye effect is a well-developed natural system to avoid light reflection from the eyes of night-active insects [2]. The moth eye is built up from

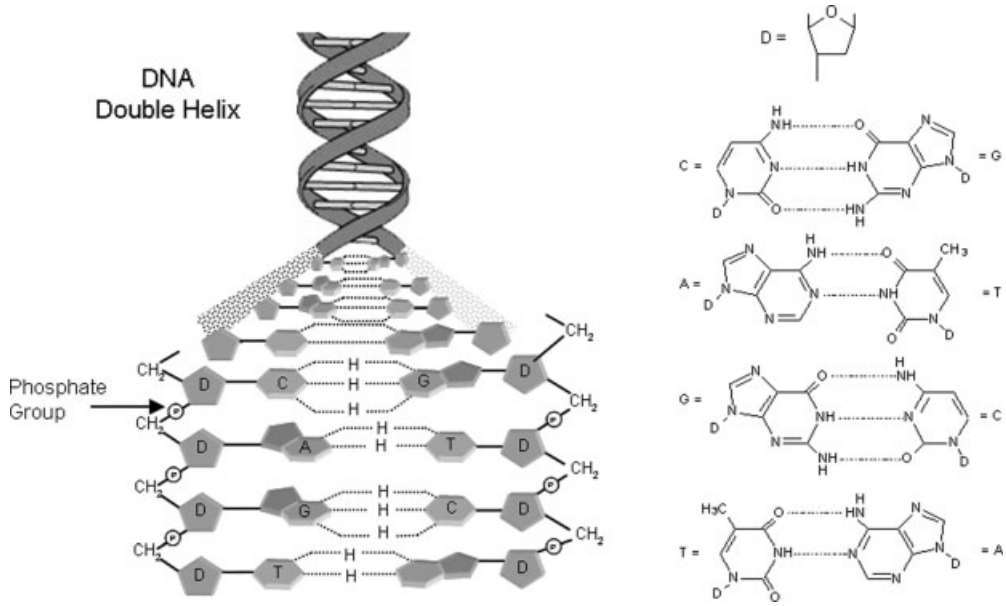


Figure 2.5 Illustration of the base pair interactions in DNA.

hemispheres 200–300 nm in diameter. Since they are smaller than the wavelength of visible light, a continuous increase in the refraction index follows, avoiding the strong reflection that occurs when light hits a flat and optically denser medium. The principle of the moth-eye effect can be used to structure surfaces artificially, for instance those of transparent materials, in order to avoid unintended reflection of light: windows, solar cells, spectacles, and so on. The techniques to nanostructure surfaces are manifold and will not be considered here. Antireflection does not start off from a distinct point. The only condition is that the structure units must be smaller than the wavelength of light. It works with 300-nm units and also with 50-nm building blocks, of course with varying results, but it works. Numerous such scaling effects have been developed into very important techniques. However, they are wrongly called "nanotechnology", since they are based only on scaling effects and not on real nano-effects as the definition demands.

2.2
From Nanoscience to Nanotechnology

Most of the currently known nano-effects are still deeply rooted in nanoscience, that is we cannot speak at all of a technology. In spite of the obvious contradiction, one usually speaks of nanotechnology even if a technology has not yet been realized. In the following, a careful differentiation will be put forward, not just between science and technology, but also with the usual intermediate step, called (nano)engineering. The development of a technique from a scientific finding never happens in a single

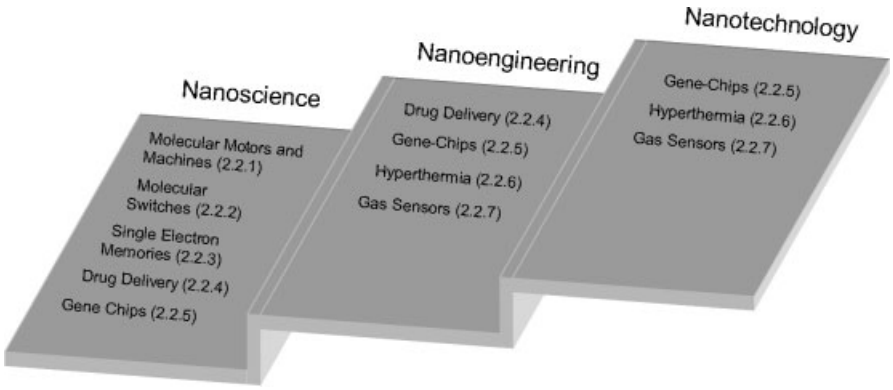


Figure 2.6 Examples of nanoscience, nanoengineering and nanotechnology.

step. Rather, a progressive procedure is necessary to develop a working device successfully. The "proof of principle" and provisional, but working, systems have to be generated before a final and commercially useful technique may result. Figure 2.6 illustrates in a rather simple manner what will be expressed.

Nanoscience, nanoengineering and nanotechnology are represented by three steps following each other. Each step contains several of very many possible examples, which will help to realize how nanoscience develops into nanotechnology. Some of the areas are still only to be found in "nanoscience", others have already developed into the "nanoengineering" step or even into "nanotechnology"; some are present in all three fields, indicating that there is still a further need for basic research to improve or to extend existing technologies. The examples presented for the three steps will be described below in order to elucidate the principles of development from basic research to nanotechnology.

2.2.1
Molecular Motors and Machines

Nature is a perfect nanotechnologist and we are well advised to learn from it. Distinct proteins and protein assemblies are known to perform special motions in response to biological stimuli [9–13]. Such systems are called molecular motors or molecular machines. Numerous attempts have been made during the last two decades to transfer the increasing knowledge of biological systems on a molecular level to devices consisting of completely artificial components or of hybrid systems where biomolecules and technical building blocks interact.

Myosins, kinesins and dyneins are frequently studied natural molecular motors [10–13]. Energetically fuelled by adenosine triphosphate (ATP), these proteins can move back and forth on actin filaments or microtubules transporting substrates. It is not the intension of this chapter to describe these natural molecular machines; rather, it is the discussion of man-made architectures in the sense of nanotechnology. Just one example illustrating Nature's principles will be briefly presented: the transport of actin filaments by myosins.

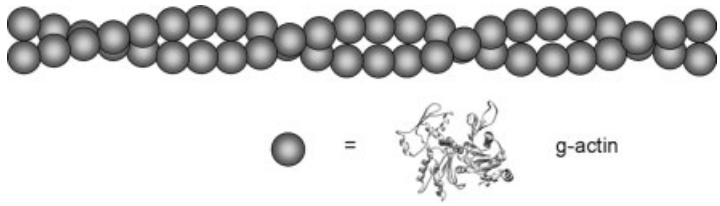


Figure 2.7 Sketch of an actin filament.

Actin filaments consist of double-stranded polymers, composed of globular actin monomers, as is indicated in Figure 2.7.

Eighteen different types of myosins exist, specialized for muscle contraction, signal transduction, vesicle transport, and so on. Myosins are composed of about 200-kDa "heavy chains" and 20-kDa "light chains", wound around the so-called neck domains. One myosin end is marked by the motor domain, as is indicated in Figure 2.8.

Figure 2.9 illustrates the formal transport of actin filaments on a myosin-modified surface. ATP is the energy supplier.

An impressive bio/artificial hybrid machine has been constructed using chaperonin systems. Chaperonins are proteins which can be isolated from *E. coli*, mitochondria and chloroplasts. A series of chaperonin crystal structures have been solved [14–18]. In Nature they make newly formed proteins, folding in their cylindrical cavities with the help of ATP as energy supplier [19]. A detailed description of the working mechanisms of different chaperonins has been published [20]. Instead of acting as host for natural proteins, chaperonins can also be used to capture and to release various nanoparticles with the help of ATP. In Figure 2.10 a primitive sketch of the uptake and the ATP-triggered release of 2–3-nm CdS nanoparticles is presented.

As mentioned above, kinesins are another class of natural molecular motors. They are capable of transporting cargo along intracellular microtubules [21]. In Nature they

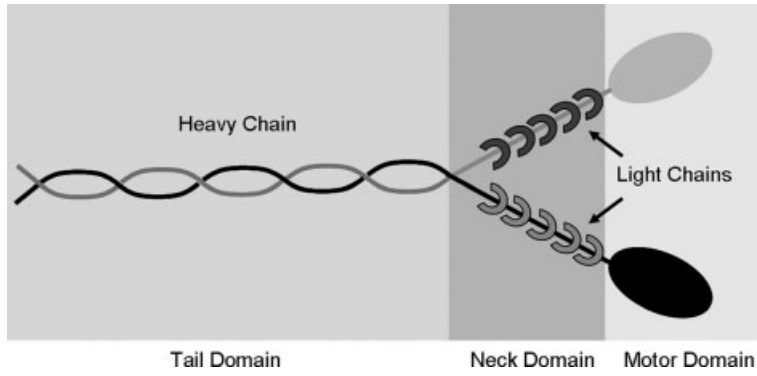


Figure 2.8 A double-headed myosin, consisting of heavy and light chains.

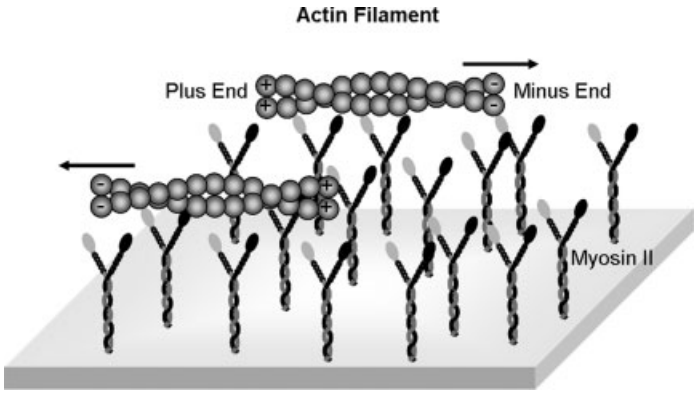


Figure 2.9 Myosin-coated surface transporting actin filaments from plus to minus. ATP serves as the energy supplier.

are part of the transport system for organelles, proteins and mRNAs. Conventional kinesin is composed of two 80-nm long 120-kDa chains, connected to two 64-kDa chains. The heavy chains are rod-like structures with two globular heads, a stalk and fan-like end [9, 22, 23]. One-headed kinesins are also known [24]. The mechanism of motion has been intensively studied using one-headed kinesins [25, 26], but will not be described here. Of course, it is also enabled by the energy of ATP hydrolysis. Microtubules are built up of 8-nm periodic building blocks of heterodimers of the subunits α- and β-tubulin, forming hollow tubes 24 nm in diameter.

Instead of a cargo of natural material such as vesicles or organelles, a recent example impressively demonstrates that artificial nanomaterial can also be transported by kinesin–microtubule systems; 7.6-nm core/shell CdSe/ZnS nanoparticles were functionalized with biotin–avidin. The as-modified quantum dot complexes were then bound to immobilized and fluorescently labeled microtubules. The movement of the loaded kinesin along the microtubules was observed by means of epifluorescence and total internal reflection fluorescence (TIRF) [27]. Figure 2.11 shows a sketch of the microtubule–kinesin–quantum dot hybrid system. The particle

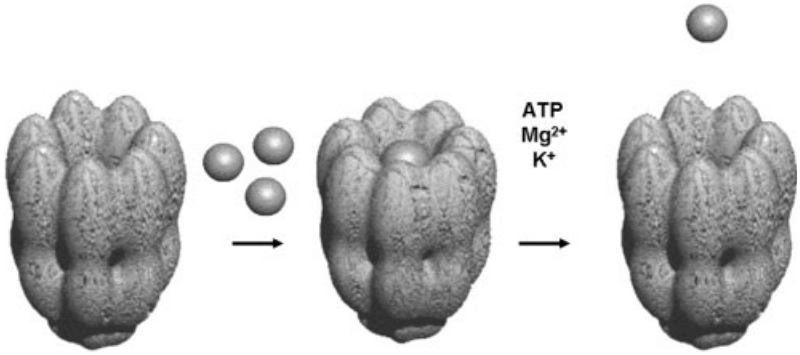


Figure 2.10 Capture and release of CdS nanoparticles by chaperonin.

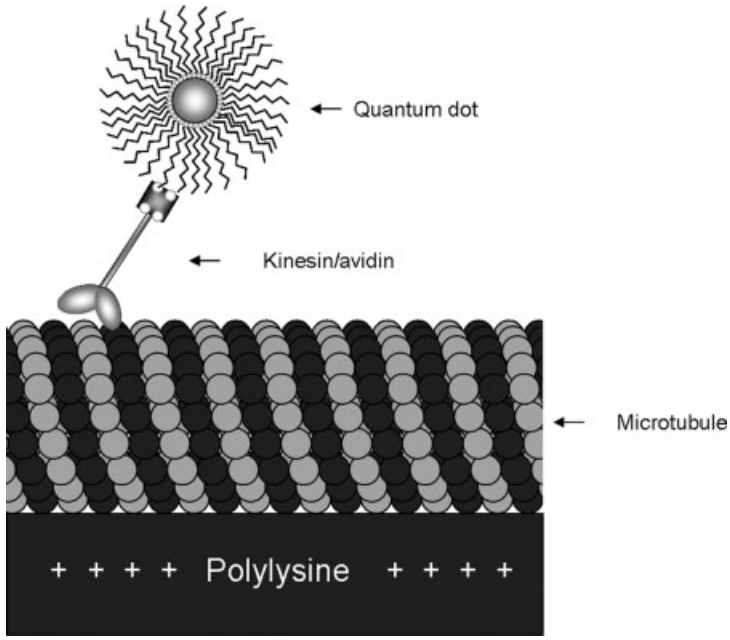


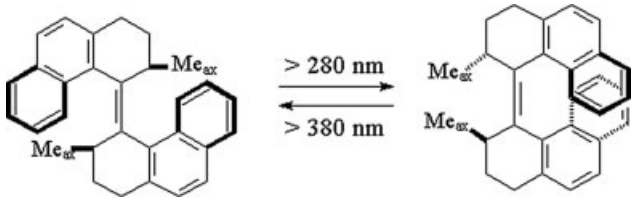
Figure 2.11 A kinesin–avidin–quantum dot hybrid system, moving on the surface of a microtubule.

transport could be visualized over 1200 s, considerably longer than in any comparable experiment before.

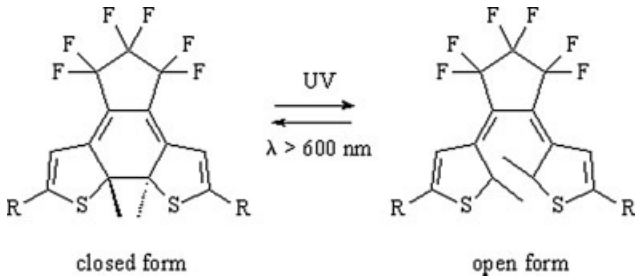
Numerous totally artificial molecular machines with non-biological components have also become known. For instance, the photoisomerization of substituted alkenes is another general method to create molecular motors, provided that the process is reversible and the two isomers are kinetically stable. In the example shown in Scheme 2.1, the four bulky substituents are responsible for an energy barrier at −55 °C between the *trans* and *cis* configurations [28].

Another light-driven system is the so-called "Irie" switch, a molecule with a light-sensitive C−C bond that can be opened or closed by two different frequencies (Scheme 2.2) [29].

Ring opening occurs under UV light and ring closure by light with a wavelength >600 nm. Switching processes induced by light are in practice much more easily



Scheme 2.1 A tetrasubstituted alkene acting as molecular motor by photoinduction.



Scheme 2.2 Light-induced closure and opening of a C–C bond ("Irie" switch).

managed than those using extensive chemistry. Nevertheless, this and comparable objects are still part of basic research.

Rotaxanes consist of two parts: a stiff bar-like part and a ring-shaped part, arranged around the bar. Due to electrostatic interactions, the ring prefers a distinct position. By chemical oxidation of the interacting position in the bar, repulsion results and the ring is shifted to another position. Reduction brings the ring back to the former position. Figure 2.12 shows a rotaxane molecule with the ring in two different positions [30]. Both configurations are stable without any voltage applied.

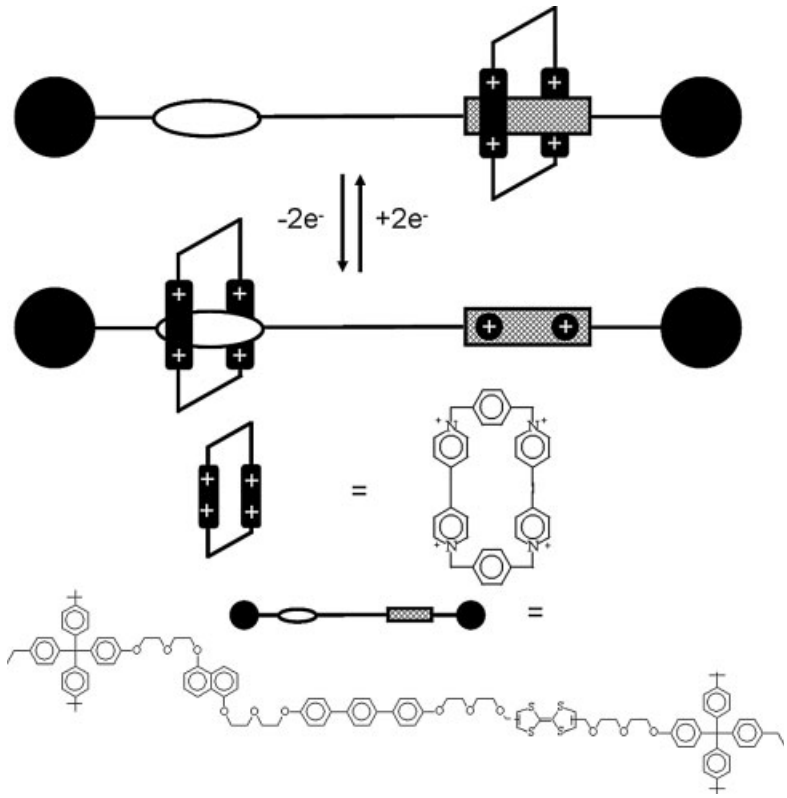


Figure 2.12 Illustration of a rotaxane system.

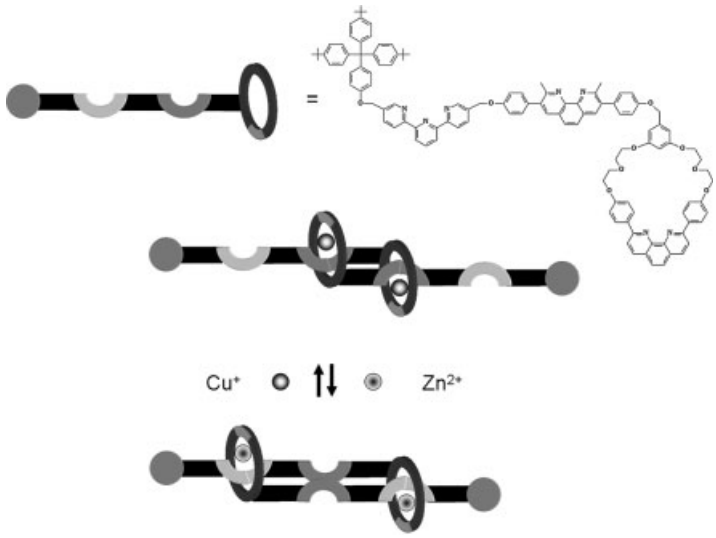


Figure 2.13 A redox-driven "molecular muscle" consisting of two combined rotaxane molecules.

Another elegant example of a rotaxane system is shown in Figure 2.13. A Cu^{+}-containing rotaxane dimer is contracted by the exchange of the Cu^{+} ions by Zn^{2+} ions and vice versa. Except as a molecular switch it can be considered as model of a molecular muscle [31].

A chemically driven rotaxane-like "molecular elevator" has been constructed and is illustrated in Figure 2.14. Protonation and deprotonation of the amine moieties of

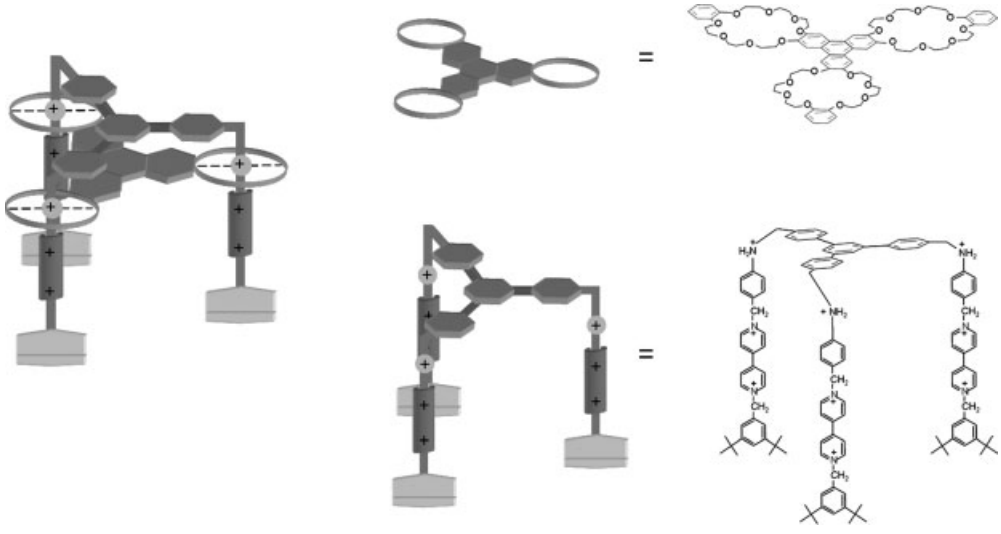


Figure 2.14 A rotaxane-based "molecular shuttle". The platform can be moved up and down by means of addition of acid or base.

part A makes the "platform" B move up and down depending on the NH/NH_2^+ situation in A [32, 33].

It is obvious that the above rotaxane examples are so far not really suited to work in devices, since complex chemistry is necessary to oxidize and to reduce the specific positions. However, the study of such or similar systems is of enormous importance in order to gather experience and to improve continuously the conditions to make such systems applicable.

2.2.2 Molecular Switches

The reasons for the worldwide and intensive search for novel generations of switches and transistors lie in the unavoidable fact that the limits of the present silicon technologies will eventually be reached. The famous Moore's law predicts that between 2010 and 2020 the two-year rhythm of doubling the capacity of computers will find a natural end due to a typical nano-effect [34]: below a not precisely known size barrier, silicon will lose its semiconductor properties and instead it will behave as an insulator. Other technologies which are based on very different nano-effects have to follow. For instance, magnetic data storage systems involving the so-called spintronics [35], magnetic recording systems using nanosized magnetic nanoparticles [36], magnetic domain walls in nanowires [37], and so on, are developing tremendously. All are still far from being of technological relevance in the near future and even the nanoengineering step has not really been reached so far. Hence they are still objects of intense basic research in nanoscience.

The term "molecular switch" is used for molecular systems which are stable in two different states. One state represents 1 and the other state 0. Different states may consist either of two different geometric conformations or of two different electronic states. Both states must be interconvertible by external stimuli. An example of molecular systems existing in two different geometric states has already been introduced in Section 2.1 with a catenane molecule that can in principle be switched by means of electrical pulses. Furthermore, all examples of artificial molecular motors described in Section 2.2.1 are at the same time molecular switches, since they exist in two different but convertible configurations. However, the use of systems, the switching of which is only based on more or less complex chemistry, looks not so much suited for application in future nanoelectronic devices; rather, it is the need to switch systems by light or electric pulses.

Molecular switches, based only on the change of the electronic spin situation in a molecule, are also promising candidates in this respect. Transition metal complexes with d^4–d^7 configurations can exist in either the high-spin (HS) or low-spin (LS) version. High-spin complexes are characterized by a maximum number of unpaired electrons, following Hund's rules. Low-spin complexes have zero or, in the case of an odd number of electrons, one unpaired electron. For instance, if an octahedral complex exists in the HS or the LS configuration depends on the energy gap between the t_{2g} and the e_g orbitals. The separation of the originally equivalent five d orbitals into three t_{2g} and two e_g orbitals (ligand field splitting) is due to the different influence of

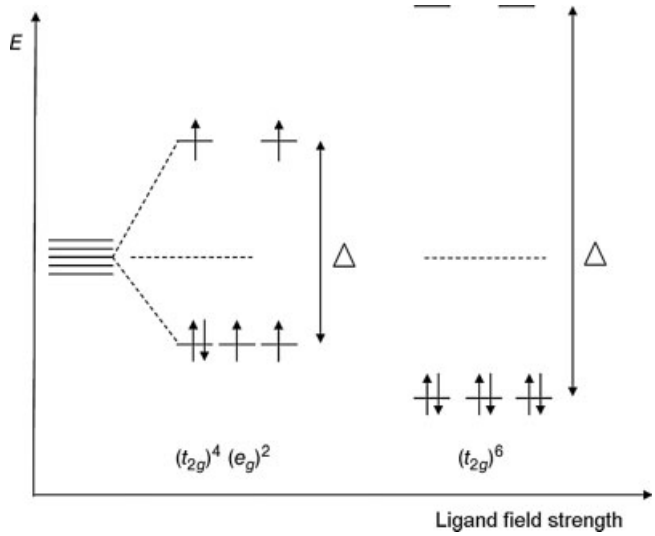


Figure 2.15 High- and low-spin configurations of a d^6 system.

the ligands coordinating the atom or ion along the x, y and z axes (e_g) or between the axes (t_{2g}). Small gaps Δ result in HS configurations and large Δ values in LS complexes. Figure 2.15 shows qualitatively the situation in both cases for a d^6 complex.

The Δ values are dominantly determined by the nature of the ligand molecules coordinating the corresponding transition metal atom or ion. So-called weak ligands (H_2O, halides X^-) cause small ligand field splitting and strong ligands (CN^-, CO, olefins) cause large Δ values. However, situations exist where the energy difference between the HS and LS configurations is small enough to be influenced by stimuli from outside and, consequently, switching between both electronic configurations becomes possible, provided that the transition between the two states is abrupt. Numerous such spin transition complexes have been identified and are of increasing interest with respect to molecular switching systems. It is of special relevance that HS–LS transitions can be induced by different stimuli such as temperature, pressure or light.

An example of a temperature-switchable complex is given in Figure 2.16. The two configurations of the d^7 Co(II) complex can be followed by the magnetic susceptibilities. The HS version has a total spin of $S = 3/2$ and the LS form of $S = 1/2$ [38].

The energetically higher lying states of the HS configuration are reached by increasing the temperature from about 130 to 150 K and vice versa.

The tetranuclear d^6 iron complex shown in Figure 2.17 can be switched by temperature, pressure or light [39]. The four iron centers allow switching over three magnetically different configurations: 3HS/1LS, 2HS/2LS, 1HS/3LS. This special situation allows manifold storage and switching varieties. The chance to switch this complex by light makes it a remarkable candidate for future applications. The

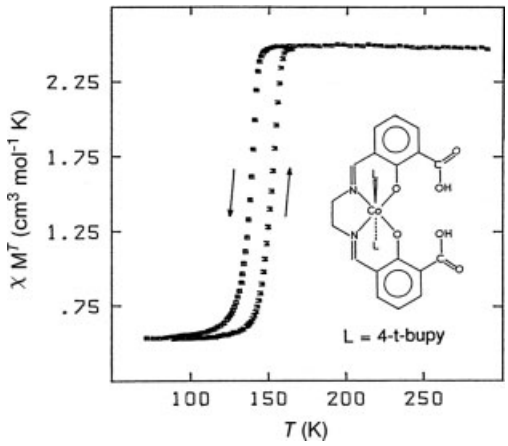
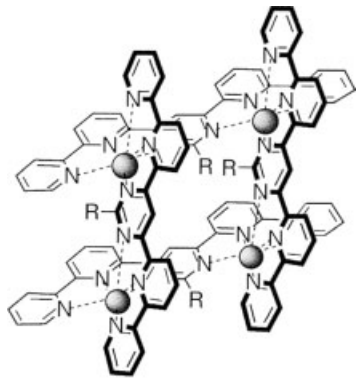


Figure 2.16 Temperature dependence of the magnetic susceptibility of an octahedral Co(II) complex. (© Elsevier, Amsterdam).

light-induced spin-state trapping (LIESST effect) however, could so far only be observed at 4.2 K with $\lambda = 514$ nm, but not the reverse effect.

Since HS and LS configurations are generally linked with a change of the bond length, pressure can be an alternative stimulus.

These few examples of existing molecular switches indicate a development in nanoscience opening up novel alternatives of storage systems on a level that can never be reached with traditional techniques. However, it is also obvious that giant efforts have to be made to reach the nanoengineering step or even the technology level. The route from the discovery of a novel fundamental effect to a product often fails due to unforeseen problems on the practical side. Nevertheless, the finding of fundamental effects in basic research is a presumption to install novel techniques, as history shows.



$[Fe_4L_4](ClO_4)_8$

Figure 2.17 Example of an LS ↔ HS switchable Fe(II) complex.

2.2.3 Single-Electron Memories

As already mentioned, the consequences of Moore's law demand alternatives, available on time. Single-electron memories might be a possible alternative to the present silicon technology if one succeeds in using the quantization of electric charge to handle digital information. Single-electron devices would be associated with enormous advantages compared with available techniques: extremely reduced power consumption, no or extremely reduced heat development, high density arrangements of building blocks and the principle possibility of generating three-dimensional memories. Single-electron memories would constitute the ultimate miniaturization of a memory device and it is worth intensively following up any chance to realize this goal. Single-electron memories, single-electron switches or single-electron transistors all require the transport of individual single electrons in a strictly controlled manner. This can only be realized by quantum mechanical tunneling. A device for single-electron tunneling (SET) must contain a unit that can be charged and discharged by single electrons. Charged and discharged units must be independently stable. To prevent uncontrolled transport of electrons, a charged unit must build up a so-called Coulomb blockade to prevent the transfer of a second, third, and so on, electron. Atoms would be ideal candidates to realize such conditions. The formation of an anion in chemistry is such an event, determined by the electron affinity of the neutral atom. However, atoms are still too small to be individually and routinely handled. Therefore, larger scaled, namely nanosized materials have to be found to overtake this "atomic capability". The ability of a unit to act as an atomic substitute depends decisively on the amount of energy to add an additional electron to an initially uncharged unit [40–44]. This charging energy scales roughly with $1/r$ (r = radius of the nanosized unit). What kind of material fulfils these conditions? – generally, all species that exhibit quantum size behavior. Nanotubes, nanowires and nanoparticles of conducting or semiconducting materials can have such extraordinary facilities. To observe quantum size behavior of a species, its dimensions in one, two or all three dimensions have to be reduced to such an extent that electrons are no longer freely mobile in all dimensions, but are confined to such an extent that they occupy more or less discrete energy levels. In Section 2.1, the phenomenon has already briefly been discussed in connection with the appearance of plasmon resonances. A detailed discussion of quantum size phenomena can be found in Chapter 4. A simplified demonstration of the consequences of electronic confinement of metals is depicted in Figure 2.18.

When reaching situation (c) in Figure 2.18, one speaks of quantum dots, since such particles no longer follow classical physical laws for bulk materials but obey quantum mechanical rules like atoms and molecules do, even at room temperature. Among many different ways to obtain information on whether the quantum dot situation has been reached or not, individual contacting of a particle by two electrodes and to study the current (I)–voltage (U) behavior gives the clearest answer. Figure 2.19 shows the principle of such an experimental arrangement.

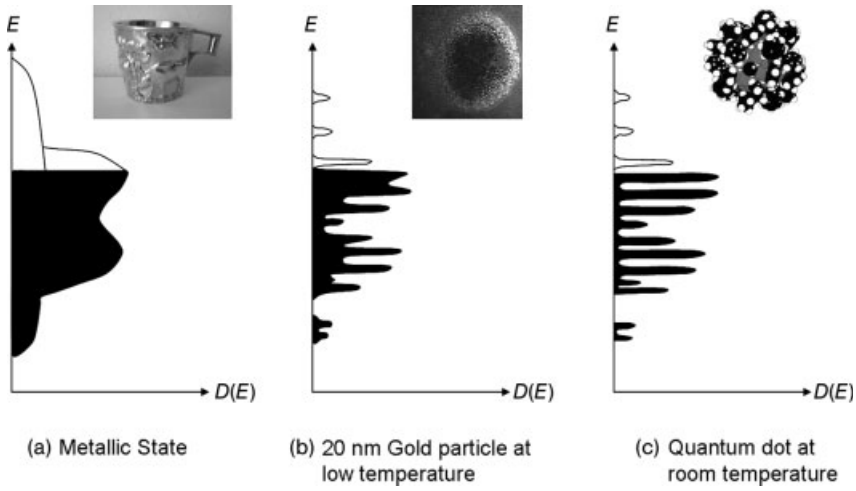


Figure 2.18 The electronic transition from the bulk state to a quantum dot.

SET from one electrode into the nanoparticle leads to an increase in charge by e^- (1.6×10^{-19} C) linked with an increase in the electrostatic energy E_C by $E_C = e^2/2C$, where C is the capacity of the nanoparticle. As can be seen from Figure 2.19, the metal nanoparticle does not directly touch the electrodes, but an insulating envelope (or a respective distance) separates it from the contacts to gain an appropriate capacity of the system. In order to avoid uncontrolled thermal tunneling of electrons, E_C must be much larger than the thermal energy $E_T = k_B T$ (k_B = Boltzmann's constant $= 1.38\times10^{-23}$ J K^{-1}): $e^2/2C \gg k_B T$. The observation of an SET process will only be possible either at very low temperatures or with very small C values. Since $C = \varepsilon\varepsilon_0 A/d$ (ε = dielectric constant, d = electrode distance from metal core, A = surface area of the particle), small C values can be realized by very small particles having a sufficiently thick insulating shell. The charge generated on the particles causes a voltage $U = e/C$, linked with a current $I = U/R_T$ (R_T = tunneling resistance).

The temperature dependence of SET has been convincingly demonstrated by the study of a 17-nm Pd nanoparticle, covered by a shell of $H_2NC_6H_4SO_3Na$ molecules. As can be seen from Figure 2.20, the I–U characteristic at 295 K is a straight line, following Ohm's law. At 4.2 K, however, a well-expressed Coulomb blockade is observed, indicating that between about −55 and +55 mV the current is interrupted due to the existence of an additional electron in the particle, blocking the transport of a second one [45].

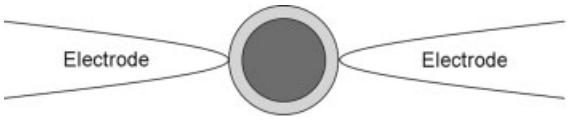


Figure 2.19 Experimental arrangement for the measurement of the *I*–*U* characteristic of a metal nanoparticle.

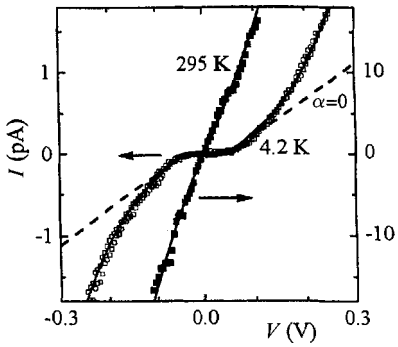


Figure 2.20 *I–U* characteristics of a 17-nm Pd particle at 295 K (Ohm's law behavior) and 4.2 K (Coulomb blockade). (© American Institute of Physics).

Instead of working at very low temperatures, in practice it is more attractive to decrease C sufficiently and to enable Coulomb blockade at room temperature. This aim has indeed been achieved by using the nanocluster $Au_{55}(PPh_3)_{12}Cl_6$ with its metal nucleus of only 1.4 nm and a 0.35-nm thick ligand shell [46, 47]. The *I–U* characteristic, shown in Figure 2.21, has been measured at room temperature. In spite of the high temperature, a Coulomb blockade between about −500 and +500 mV is registered [48].

The electric contacting of a single nanocluster has been performed using a tip of a scanning tunneling microscope (STM) and a conductive substrate on which the particles had been deposited from a very dilute solution.

A deeper insight into the electronic situation in these Au_{55} quantum dots was possible by studying the *I–U* behavior at 7 K [49]. Due to the low temperature, the

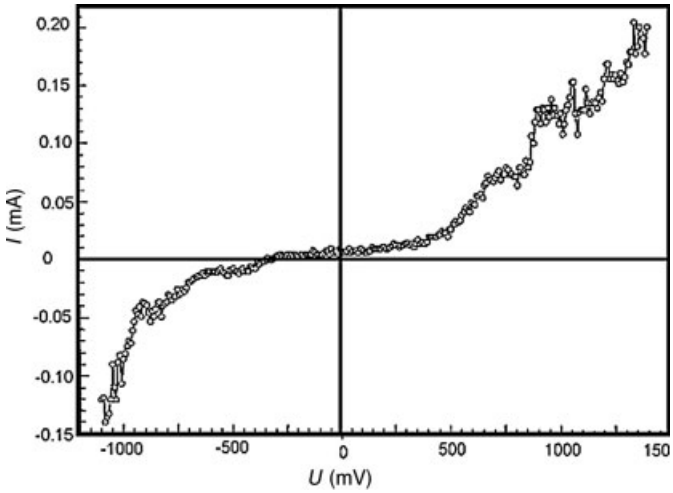


Figure 2.21 Room temperature Coulomb blockade of $Au_{55}(PPh_3)_{12}Cl_6$. (© Springer, Berlin).

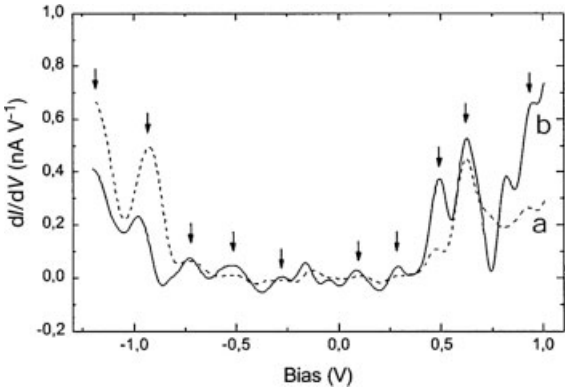


Figure 2.22 d*I*/d*V* diagram for $Au_{55}(PPh_3)_{12}Cl_6$ at 7 K indicating level splitting of 135 meV. (© American Chemical Society).

Coulomb blockade is enlarged; however, the most informative knowledge can be seen from a diagram using d*I*/d*V* values instead of *I* (Figure 2.22). Generally, the blockade then turns into a minimum on the *U* axis. Due to the low temperature, discrete energy levels in the minimum become visible in terms of conductivity oscillations with an average level spacing of 135 meV. The dashed line and the solid line result from two measurements on the same particle, but at different positions, namely above a phenyl ring of PPh_3 and at a position above bare gold atoms. They agree fairly well and so indicate that the result does not depend on the matter between the tip and Au_{55} nucleus.

This result demonstrates the existence of a perfect quantum dot, working at room temperature and representing exactly position (c) in the sketch in Figure 2.18. Figure 2.22 also makes it clear why such working units are sometimes called "artificial atoms".

The Coulomb blockade, based on the transfer of single electrons, represents a perfect single-electron switch and can in principle also be used as a single-electron transistor, as indicated in Figure 2.23.

These fundamental findings make Au_{55} and metal particles of similar size excellent candidates for use in future storage systems. Intensive research and development are still necessary to reach this ultimate goal. The very first steps from the science level to the engineering step are in progress, but nevertheless, quantum dot memories are still deeply involved in basic research.

2.2.4
Drug Delivery

The controlled release of a drug has a significant influence on the therapeutic efficiency. For numerous drugs there exists an area of concentration with the greatest success. Compared with traditional tablets or injections, drug delivery systems have a profile with continuous release, as indicated in Figure 2.24 [50].

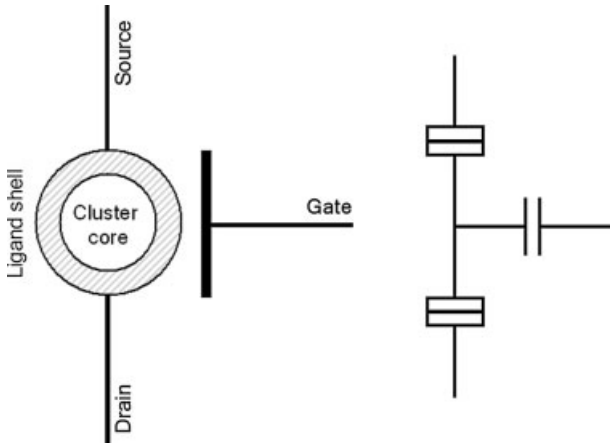


Figure 2.23 Circuit of a single-electron transistor.

Conventional methods are characterized by a rapid increase in release, followed by a fast decay (peak-and-trough cycle). Therefore, there is an urgent need for continuous drug release in the therapeutic area of concentration. The history of drug delivery based on implanted systems goes back to the 1960s [51]. Instruments at that time had considerable dimensions containing electrically driven pumps. In the course of the last two decades, numerous systems have been developed with novel principles [52–56]. Partially they also use novel principles such as release via the skin or the nasal mucous membrane. For longer times of therapy a drug delivery implant of reasonable size and high reliability would be the most effective system. Nanotechnology offers numerous chances to achieve that goal.

Entrapping and encapsulating drugs in nanostructured systems have been developed with remarkable success, releasing a drug uniformly over longer time periods. They are based on guest–host systems. Different kinds of chemical bonds between the guest (drug) and host system determine the speed of release. Hydrogen or van der Waals bonds and also electrostatic interactions can be used to combine the guest and

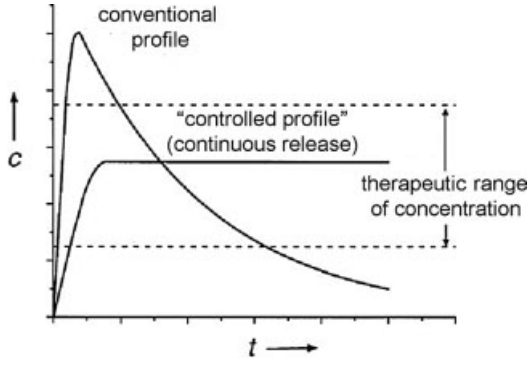


Figure 2.24 Concentration (c)–time (t) profile of a conventional and a controlled drug release.

host to such an extent that slow release is enabled. A host–guest system based on dendrimers as host molecules will be described briefly as an actual example of development. Dendrimers are highly branched molecules in the nanometer size regime. They consist of a core unit from where "branched branches" extend in different directions, forming a three-dimensional architecture bearing end groups of various functionality. Dendrimer structures unavoidably contain cavities inside the skeleton. These are able to take up guest molecules of appropriate size and to release them slowly depending on the surrounding conditions. With an increasing number of branches, the number and geometry of the cavities become variable and also increase [57]. Figure 2.25 shows a formal sketch of a dendrimer molecule taking up and releasing host particles.

Another interesting nano-based drug delivery system uses superparamagnetic iron oxide nanoparticles, usually embedded in a polymer matrix and attached to a drug system. External, high-gradient magnetic fields are applied to transport drug-loaded beads to the corresponding site in the body [58, 59]. Once the system has concentrated in the tumor, the drug is released using different techniques such as increase in temperature, change of pH value or enzymatic activity.

Another method uses superparamagnetic iron oxide nanoparticles, the surface of which is modified by DNA sequences. Those particles easily enter cells using receptor-mediated endocytosis mechanisms in combination with a magnetic field gradient [60]. Having entered the cell, the DNA is liberated and can enter the nucleus. This so-called non-viral transfection is of special interest for gene therapy.

A rapidly growing drug delivery development is based on the use of multifunctional nanoengineered capsules containing various kinds of active compounds. Attempts have been made to solve the general problem of treating only pathological cells and not healthy ones by using, for instance, functionalized polymer capsules having distinct release, permeability and adhesion properties. The inner volume can

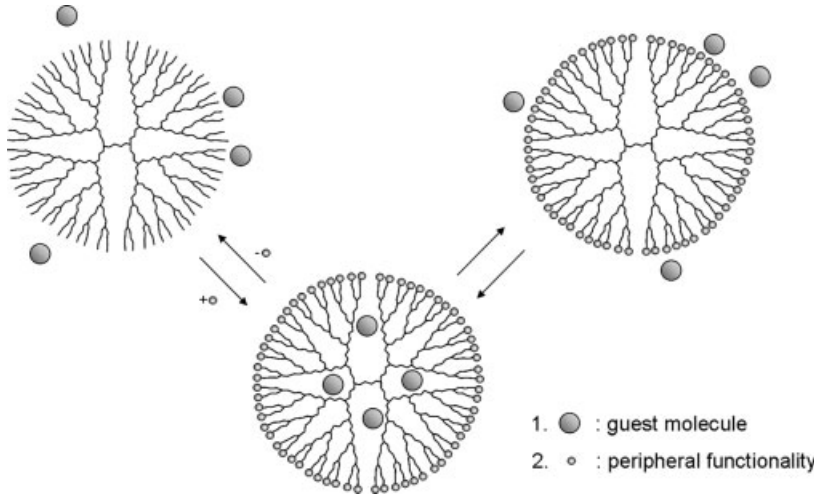


Figure 2.25 Inclusion of guest molecules in cavities of a dendrimer.

be filled with magnetic nanoparticles which allow aimed transport by outer magnetic fields, by modification of the capsule surface with specific receptors to target specifically diseased cells or by generating capsules acting as a nanoreactor producing products which are only toxic for diseased cells and cause selective apoptosis [61, 62].

Self-rupturing microcapsules consist of polyelectrolyte membranes, permeable to water, but not to the drug-containing degradable microgels, filling the holes of the capsules. The hydrolytic degradation of the microgels causes a swelling pressure, rupturing the outer membrane [63].

Polyelectrolyte capsules which degrade at physiological pH values open up novel prospects for drug delivery [64]. Intracellular targets such as nucleic acids or proteins can cause opening of the capsules. Using $CaCO_3$ particles as the carrying material for fluorescein isothiocyanate–dextran (FITC–dextran), assemblies of $CaCO_3$/FITC–dextran are formed by coprecipitation, followed by layer-by-layer polyelectrolyte membrane formation, for instance with poly-L-arginine as the polycation and dextran sulfate as the polyanion. Finally, the $CaCO_3$ particles are removed using buffered EDTA solution.

Finally, laser-induced opening of a polyelectrolyte membrane inside living cells can be mentioned [65].

A completely different drug delivery system has been developed using nanoporous alumina membranes for the controlled long-term release of drugs. Nanoporous alumina membranes with variable pore diameters between 10 and 200 nm are easy to prepare and are used to control the speed of release depending on the pore size [66]. Figure 2.26 shows a sketch of the implantable device and Figure 2.27 illustrates the influence of the speed of release of the same molecule depending on the pore diameter. Instead of a real drug, the system has been developed using crystal violet for the easy determination of concentrations by means of UV–visible spectroscopy. Of course, this system requires individual developments for each drug to optimize the pore size and solubility conditions, for instance with the help of surfactants.

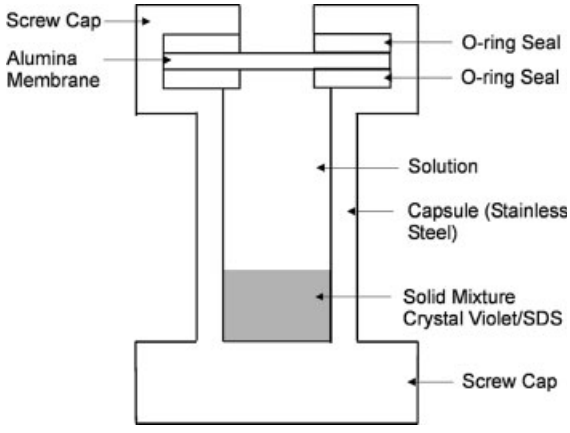


Figure 2.26 Sketch of a drug delivery system using nanoporous alumina membranes to control the speed of release.

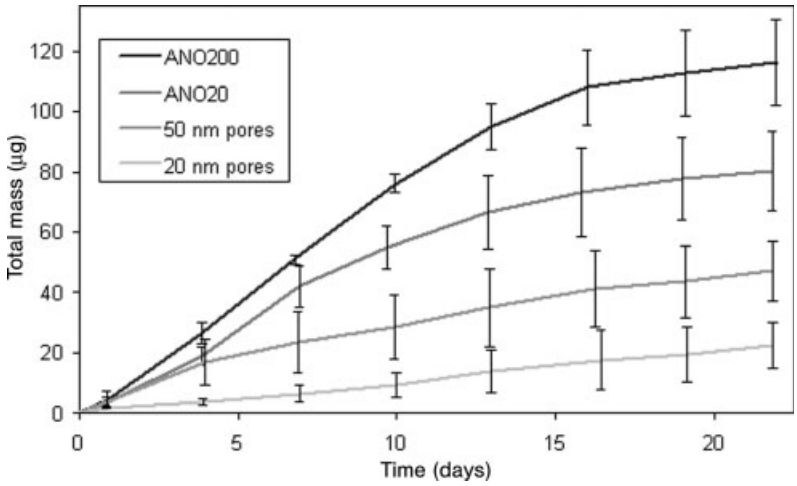


Figure 2.27 Dependence of the release of crystal violet on the pore size.

None of these nano-based techniques have yet reached the "technology standard". Rather, they are under development and have partially reached the "engineering state".

2.2.5 Gene Chips

Progress in the diagnostics of human diseases is of comparable importance to progress in therapy. Nanotechnology is, without doubt, the most promising field where improved developments can be expected. The scientific aim is to develop diagnostics on a molecular level, if possible with routine techniques. Gene chips, also called microarrays, represent the best way to fulfill such dreams. Gene chips are already commercially available, but are still also under investigation and development on the engineering level and also are still objects of intense scientific research. The early diagnosis of a cancer disease is crucial for successful therapy. So, this is a first example with a keyword existing in all three steps in Figure 2.6.

A gene chip consists of a collection of up to 400 000 single-stranded DNA segments (probes), fixed on a glass, silicon or plastic surface. Such DNA microarrays can be used to detect mRNAs which either interact with DNA fragments on the chip or not. Since there can be a huge number of reporters arriving, a microarray experiment can accomplish the equivalent number of genetic tests in parallel. Detection occurs by using organic fluorophore labels. So-called two-channel microarrays contain probes consisting of complementary DNA (cDNA), but also of oligonucleotides. This type of chip is then hybridized with cDNA from two different samples, labeled with two different fluorophores for comparison, for example from cancer cell lines and from a control. Mixing and hybridization on the microarray allow the visualization of the results. Gene chips working on that basis are commercially available and allow cheap and fast diagnosis on a level that was inconceivable one or two decades ago.

DNA microarrays are also used to analyze the sequence of particular genome sequences. Gene chips of this type have, in spite of their enormous contribution in diagnostics, inherent drawbacks due to the fluorophore labeling (in some cases even radioactive labeling is used). Recent advances in nanoscience open the door to increase the sensitivity of DNA detection to an unknown level. DNA–nanoparticle conjugates are powerful tools in this direction.

New developments of microarrays are based on the use of DNA, functionalized with quantum dots. Their excitation and emission properties make them powerful candidates for replacing fluorophore labeling techniques. Gold and silver nanoparticles, but also CdSe and ZnS quantum dots, have been successfully tested.

The surface plasmon resonance of gold nanoparticles and the resulting intense color allow the very sensitive and selective colorimetric detection of corresponding DNA sequences. Au nanoparticles in the size range from about 10 up to 100 nm, typically 10–20 nm, are linked with DNA probe strands by using 3′- or 5′-end mercapto-functionalized species. Due to the preferred formation of strong Au–S bonds, Au–DNA hybrid systems become easily available. Mirkin's group first reported the use of mercaptoalkyloligonucleotide-modified gold nanoparticles [67–71]. They used two samples, each complementary to one half of a target oligonucleotide, so that the formation of a polymeric network was induced by mixing the three species, as indicated in Figure 2.28. The purpose of this experiment was to perform a color change from red to blue, due to the aggregation of the nanoparticles.

Based on the colorimetric nanoparticle approach to DNA detection, microarrays using DNA–gold nanoparticle hybrid systems are being increasingly developed. A three-component system was first described by Mirkin and coworkers [69, 72, 73]. It consists of a glass chip, the surface of which is modified by capture DNA strands to recognize the DNA under investigation. The oligonucleotide-functionalized gold probe and the target DNA complete the system. Intense washing after the combination process results in a selectivity of 10 : 1 for single base-pair mutations. In a final special step, the gold nanoparticles are covered by a silver shell which is simply generated by the catalytic reduction of silver ions on the gold nanoparticles. The capture-strand–target–nanoparticle combination can then be visualized using a flatbed scanner. Due to the presence of silver shells on the gold particles, high

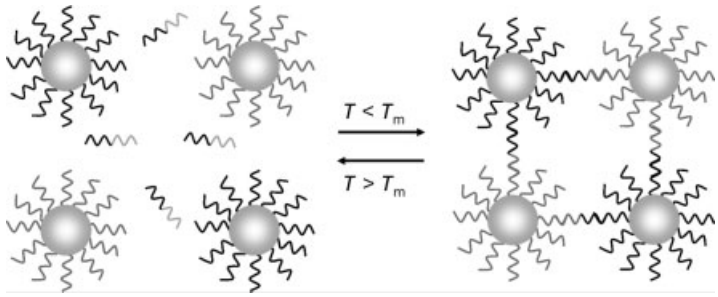


Figure 2.28 Aggregation of oligonucleotide-functionalized gold nanoparticles by means of complementary target DNA.

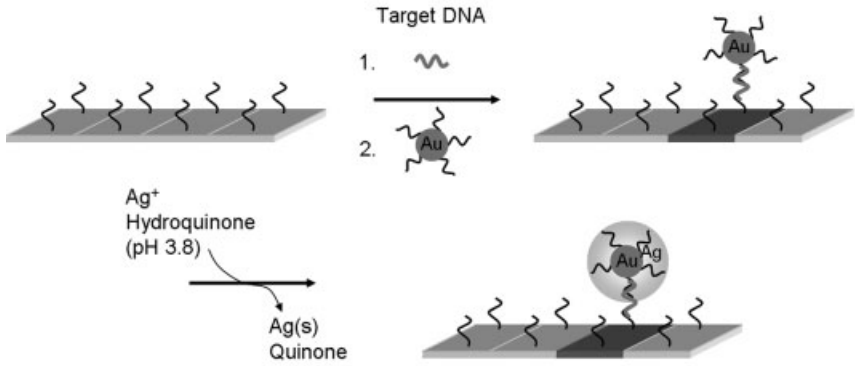


Figure 2.29 Principle of a scanometric DNA assay. A capture oligonucleotide on a surface binds one half of a target molecule and oligonucleotide gold nanoparticles bind the other half. A detectable signal results by the catalytic coverage of the gold nanoparticles by silver.

surface enhanced Raman scattering (SERS) is observed and can also be used for target DNA detection. Figure 2.29 elucidates the various steps.

The improvement of this technique, compared with conventional fluorophore-labeling techniques, is about 100-fold, namely as low as 50 fM. These remarkable nanotechnologically based developments will initiate great progress in diagnostics. As an example, first investigations of Alzheimer's disease (AD) can be mentioned [74].

Alivisatos and coworkers detected single base-pair mutations in DNA by the use of CdSe/ZnS quantum dots in chip-based assays [75]. The detection method was fluorescence microscopy and the detection limit was about 2 nM. This is not yet in the region of the detection limit described above, but it is likely that this technique can be improved considerably since it is known that even individual quantum dots can be detected under ideal conditions [76].

Gene chips based on the use of quantum dots belong to one of the most promising developments in nanotechnology. In an unusually short time, beginning with the very first experience with biomolecule–quantum dot interactions, a development started that has already led to commercially available devices. There are still simultaneous efforts to be made on all three levels. Further improvements of detection limits are still part of nanoscience. At the same time, improvements of routine detection processes are continuing in order to facilitate everyday clinical handling.

2.2.6 Hyperthermia

Hyperthermia, known for several decades, has more or less developed to a level of clinical applications based on nanotechnological attempts and can now be located in the field of "nanoengineering" and also "nanotechnology". It uses the fact that superparamagnetic nanoparticles can be warmed up by external alternate magnetic

fields. As has long been known, tumor cells respond sensitively with apoptosis on temperature increases of only a few degrees (40–44 °C) [77–81].

The superparamagnetic state of a material at room temperature is reached when the thermal energy kT (k = Boltzmann's constant) overcomes the magnetostatic energy of a domain or particle. If the particle or domain is small enough, a hysteresis no longer exists or, in other words, the magnetic unit no longer exhibits the ability to store the larger particle's magnetization orientation; rather, the magnetic moments rotate and so induce superparamagnetic behavior. Typical particle sizes for the transition from ferro- to superparamagnetism are in the range 10–20 nm for oxides. Metal particles have to be downsized to 1–3 nm. A great advantage of superparamagnetic particles is the fact that they can be dispersed in various liquids without any tendency to agglomerate, an important condition for applications in medicine.

Several types of superparamagnetic oxide nanoparticles have been investigated for application in hyperthermia. The most promising candidates are magnetite and maghemite since their biocompatibility has already been shown. The amount of magnetic material to reach the necessary temperature depends on the concentration of the particles in the cells. Direct injection allows larger quantities than intravascular administration or antibody targeting. On the other hand, direct injections into the tumor involve a certain danger of the formation of metastases. In any case, the amount of magnetic nanoparticles necessary for a sufficient temperature increase depends on the magnetic properties of the particles and on the external radio-frequency field. Under then optimum conditions only 0.1 mg per mL of tissue is necessary to induce cell death.

2.2.7 Gas Sensors

Gas sensors have been known and applied since the early 1960s [82, 83]. The fields of application range from industrial and automotive needs (NO_x, NH_3, SO_2, hydrocarbons, etc.) via domestic gas determinations (CO_2, humidity) up to the security sector, where traces of explosives have to be detected. In a working sensor system, the information resulting from the chemical or physical interaction between a gas molecule and the sensor has to be transformed into a measurable signal. Numerous possibilities such as electrochemical, calorimetric, acoustic, chemoresistant and other effects are well established. Chemoresistors typically use metal oxides. These change their electrical resistance when they oxidize or reduce gases to be detected [82–85]. Continuous improvements have been elaborated concerning sensitivity, selectivity, stability, response and recovery time [86]. In connection with nanotechnological developments sensors based on metal oxides and metal nanoparticles have been intensively studied and have reached the state of engineering and even technology (see Figure 2.6) [86].

Sensors based on nanosized *metal oxides* provide both receptor and transducer functions [87]. The receptor must ensure a specific interaction of the sensor's surface with the target analyte. The transducer's task is to transform the molecular information into a measurable change of the electrical resistance. For instance, the conductivity of n-type semiconducting metal oxides increases on contact with reducing

gases, whereas that of p-type oxides decreases. Oxidizing gases cause opposite effects. A frequently used wide-bandgap n-type semiconductor is SnO_2. A qualitative explanation (for details see [86]) of the working principle says that, in the presence of dry air, oxygen is ionosorbed on the oxide surface, depending on temperature as O_2^- (<420 K), as O^- (420–670 K) or as O^{2-} (>870 K). The electrons required for the reduction of O_2 come from the conduction band, so generating an electron-deficient region, the so-called space–charge layer Λ_{air} [88–91]. Λ_{air} depends on the Debye length L_D, a material- and temperature-dependent value. For SnO_2 at 523 K it is about 3 nm [92]. In real systems water is present to some extent, forming hydroxyl groups on the surface and affecting the sensor's properties. The influence of water has been discussed in detail [93].

Reducing gases interact with the ionosorbed oxygen species and are oxidized, for instance $CO \rightarrow CO_2$, which is desorbed. Even traces of reducing gases decrease the number of oxygen species to such an extent that, due to the release of surface-trapped electrons, the increase in conductance becomes measurable. In the case of oxidizing target gases, the process is inverse: additional electrons are removed from the semiconductor, resulting in an increase in Λ_{air}. Hence adsorption of oxidizing gases, for example NO_2 or O_3, causes a decrease in conductance.

The efficiency of a gas sensor depends not only on the material of which it is made, but decisively also on the size of the particles and their arrangement, since the relevant reactions occur on the particles' surface. In an ideal case all existing percolation paths are used, contributing to a maximum change in conductance. The response time depends on the equilibrium between the diffusion rate of the participating gases. Film thickness and porosity are therefore of special relevance for the quality of a sensor [94, 95].

A vital role in this connection is played by the size of the particles forming the macroscopic film. Since the analyte molecule–sensor interaction occurs on the particles surface, their surface:size ratio plays a dominant role. Since the relative proportion of the surface increases with decreasing particle size, smaller particles should be more efficient than larger particles. This can clearly be seen from Figure 2.30 [87].

SnO_2 particles with diameters below about 10 nm exponentially increase the sensor's response. In addition to the increase in surface area, particle radii in the range of the space–charge layer Λ_{air} decrease the Schottky barriers between depleted zones or even lead to an overlap, with the consequence that surface states dominate the electrical properties and so have a decisive influence on the sensor performance. Very small differences in the particle size can have crucial consequences for the sensor's ability. As has been shown for WO_3 nanoparticles, a reduction from 33 to 25 nm increases the sensitivity towards 10 ppm NO_2 at 573 K by a factor of 3–4 [96]. Several similar examples for other metal oxides are known [86].

Metal nanoparticles as building blocks for sensor systems have been under investigation since the late 1990s. Thin films of ligand-protected metal nanoparticles change their conductance when gas molecules are absorbed in the regions between the nanoparticles [97, 98]. Films of 2-nm gold nanoparticles, covered with octanethiol molecules, turned out to change the conductance reversibly if gas molecules such as toluene, 1-propanol or water became part of the interparticle sphere. This principle

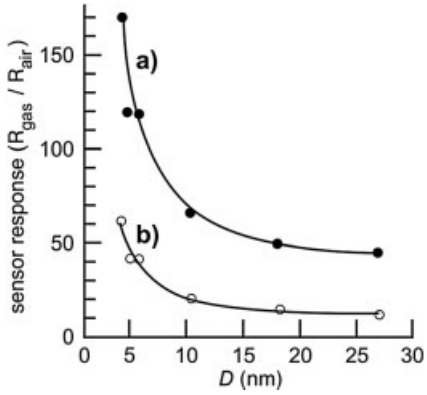


Figure 2.30 Dependence of the sensor response R on the particle diameter D (SnO_2), expressed as R_{gas}/R_{air}. (a) 800 ppm H_2 and (b) 800 ppm CO in air at 573 K. (© Springer, Berlin).

has since been improved by the introduction of specifically functionalized ligand molecules, increasing the sensitivity [99–102]. Self-assembled layers of gold nanoparticles and dendrimer molecules use the ability of dendrimers to host guest molecules in their cavities [101, 103].

Apart from gold, other noble metal nanoparticles have also been tested. Platinum nanoparticles, for instance, partially crosslinked by dithiol molecules, are active sensors towards toluene, H_2O, CO and NH_3 [104–107]. In the case of NH_3, traces down to 100 ppb could be detected. Figure 2.31 informs on the different sensitivities of such a chemoresistor towards different gases.

Whereas charge transport mechanisms between ligand-protected metal nanoparticles have been investigated in detail [86, 108–119], knowledge about the change in conductance due to the influence of various gases is still rather limited.

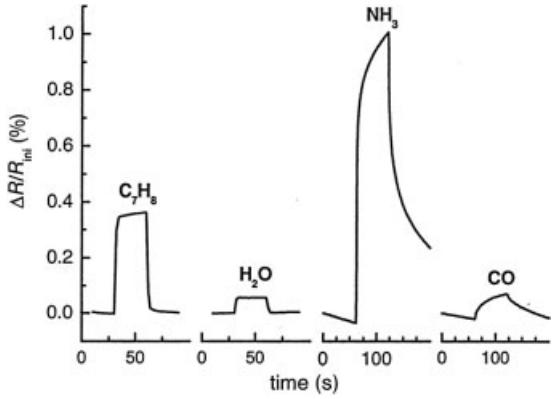


Figure 2.31 Response of a sensor system consisting of partially nonanedithiol-crosslinked Pt nanoparticles towards 400 ppm amounts of different gas molecules. ($\Delta R = R_{gas} - R_{ini}$; R_{ini} = resistance in dry air). (© Elsevier, Amsterdam).

2.3
Technologies on the Nanoscale

2.3.1
Introduction

This section deals with images of "nanotechnology" that do not fulfill the strict definition of nanoscience and nanotechnology, explained in Section 2.1. Why is this necessary? The term "nanotechnology" has now reached such a broad and thereby diffuse meaning that it seems helpful to discuss some examples of those already installed techniques that are falsely integrated into nanotechnology in a broader sense. Indeed, in some cases it is not trivial to decide if an effect and a thereby resulting technique follow the precise definition or not. From experience it can be seen that "wrong" nanotechnology means techniques that are settled on the nanoscale, but without the decisive size-dependent or functionality-determined nano-effect. What is usually meant by this common understanding is "technology(ies) on the nanoscale". The exclusion of those techniques from the scientifically exactly defined techniques is not discrimination. Rather, some of them became very important and others will follow. To conclude this introduction, one should try to differentiate clearly between nanotechnology and technologies on the nanoscale. The latter can also be considered as scaling effects without indicating nano-specific effects.

2.3.2
Structured Surfaces

It has long been known that structured surfaces change the physical and chemical properties of the corresponding material. Two property changes dominate the interest in structured surfaces: (i) change in wettability and (ii) change in optical properties. Both are of enormous importance both in nature and in technique. What kind of structure are we talking about? Let us consider first a natural "technique", that has been copied in many respects: the *wettability behavior*. Barthlott *et al.* have investigated since about 1990 the surface of lotus leaves for its special property of having a permanent clean surface [120–122]. Like all primary parts of plants, lotus leaves are covered by a layer of hydrophobic material. In case of lotus leaves, this layer consists of epicuticular wax crystals. Scanning Electron Microscopy (SEM) investigations additionally showed a structural design that is responsible for the super-hydrophobicity of lotus leaves. This special behavior has subsequently become known as the lotus effect. The SEM images showed that the surface of the leaves consists of a double structure of microsized cells decorated with nanosized waxy crystals.

The physical background for this phenomenon can be seen in the behavior of water droplets on a micro/nanostructured surface. It is important to state that the effect is not dependent on a distinct size of the structure units. As it turned out, the lotus combination of micro- and nanosized units is advantageous, but it is not a condition. Also, the absolute size of the nanostructure units is not decisive to observe the effect, it can only improve or worsen the hydrophobic nature.

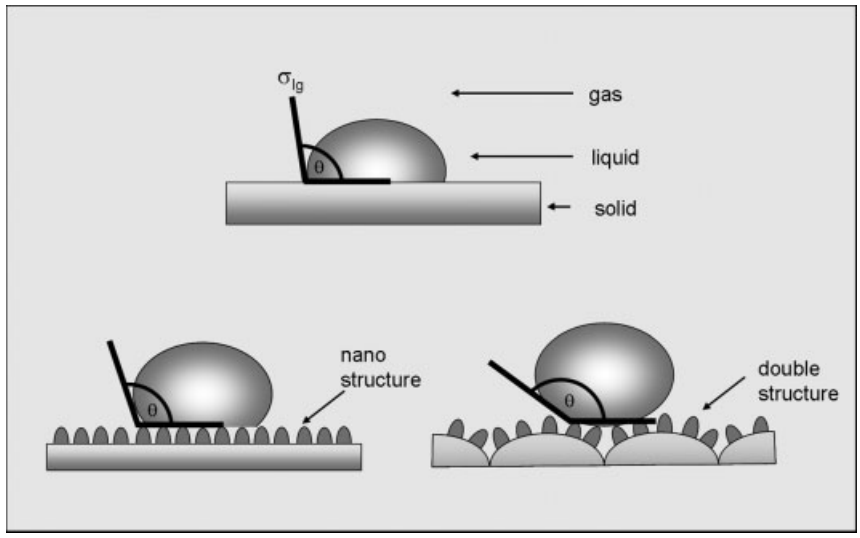


Figure 2.32 Sketch of the gas–liquid–solid three-phase system. (a) Water droplet on a smooth surface resulting in small contact angles θ, (b) on a nanostructured surface with increased θ value and (c) on a bimodal micro/nano structured surface with the largest contact angle.

The wettability generally describes the interaction of a liquid with a solid surface. It is described by the Young equation, $\sigma_{sg} - \sigma_{sl} = \sigma_{lg}\cos\theta$, where σ_{sg} = solid–gas interfacial tension, σ_{sl} = solid–liquid interfacial tension, σ_{lg} = liquid–gas interfacial tension and θ = solid–liquid contact angle [123]. The contact angle θ is the angle between the solid surface and the tangent applied at the surface of the droplet. Figure 2.32 illustrates the situation of a water droplet on a smooth surface, a nanostructured surface and a micro/nanostructured surface.

It is obvious that structured surfaces in any case cause larger contact angles than flat surfaces and so show an increased hydrophobicity. The reason is that the energy to distribute a water droplet on a structured surface extends the gain in energy by additional interactions of water molecules with the surface.

Soon after its recognition, the lotus effect led to the development of artificially micro/nanostructured surfaces. Lithographic techniques, self-assembly processes, controlled deposition, size reduction and replication by physical contact are applicable routes [124]. An elegant replication procedure will be briefly considered. It uses masks consisting of nanoporous alumina films. The advantage is their rather simple fabrication by anodization of aluminum surfaces, the easy adjustability of the pore diameters and the hardness and the temperature stability of alumina [125–129]. Using appropriate imprinting devices, various polymers and metals could be nanostructured [130]. The successful 1 : 1 polymer transfer from the mask to the surface is shown in Figure 2.33 by means of a poly(methyl methacrylate) (PMMA) surface.

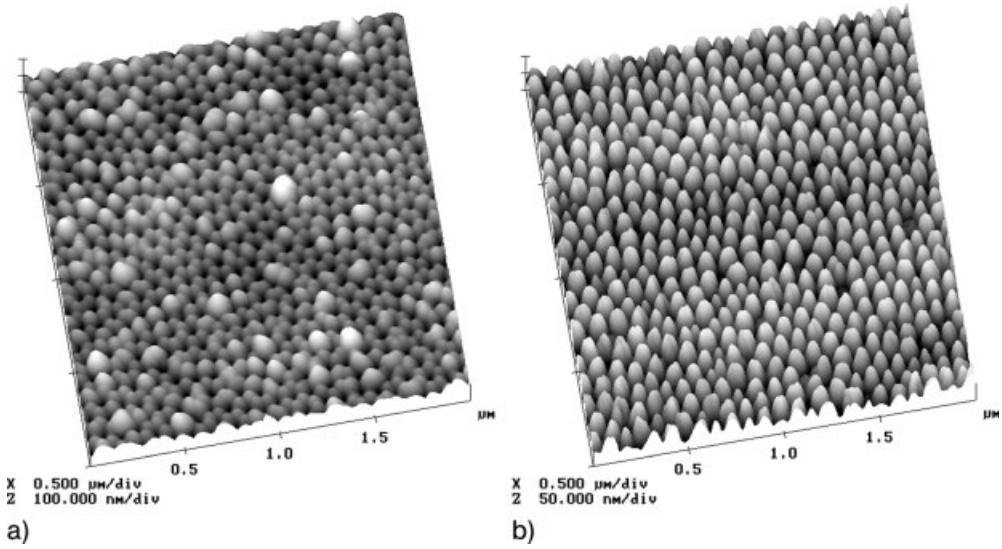


Figure 2.33 Proof of 1 : 1 pattern transfer. (a) AFM image of an alumina surface with 50-nm pores; (b) imprinted PMMA surface indicating some defects from the mask at the same positions.

Polycarbonate (PC) and polytetrafluoroethylene (PTFE) could also be nanostructured with different pore widths. Aluminum, iron, nickel, palladium, platinum, copper, silver and brass are examples of successfully nanostructured metals. Figure 2.34 shows an AFM image of a nanostructured silver surface using a mask with 50-nm pores.

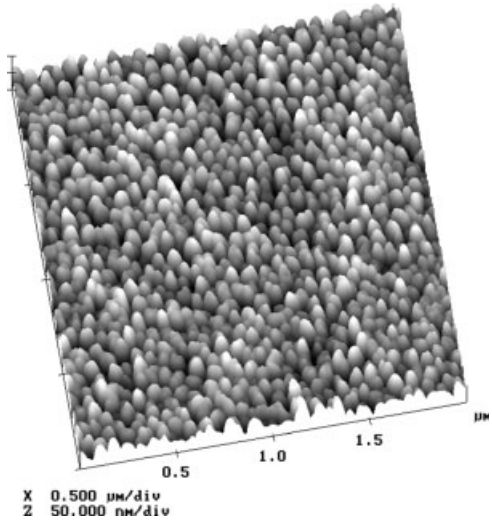


Figure 2.34 AFM image of an imprinted silver surface.

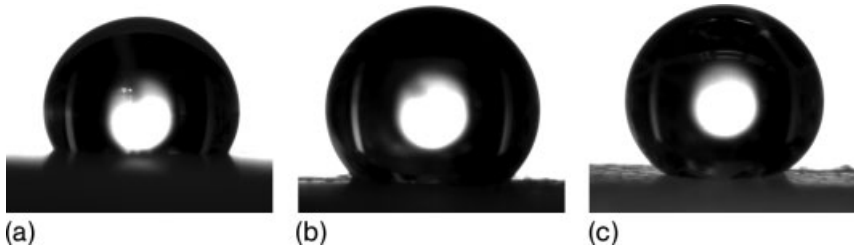


Figure 2.35 Light microscopic images of water droplets on (a) 50-nm, (b) 120-nm and (c) 170-nm structured PTFE surfaces.

By means of PTFE surfaces, imprinted with masks of 50, 120, 170 and 200 nm, it has been demonstrated that each of the surfaces makes the hydrophobicity increase, compared with an untreated surface. However, there is no spontaneous effect to be observed, rather it is a scaling phenomenon. In Figure 2.35, light microscopic images of water droplets on variously nanostructured PTFE surfaces are shown. There is an increasing contact angle to be registered if the pore width and thereby the pillars on the surfaces increase. This continuous development of the contact angles can also be followed from Figure 2.36.

In addition to the wettability properties, a second physical behavior changes with structure: the *light transmission* of transparent materials. In Figure 2.37, the increasing transmission of visible light through PMMA windows with decreasing structure size is demonstrated.

Improvements in the transparency of glasses, linked with a reduction in reflection, has important practical consequences in optical devices.

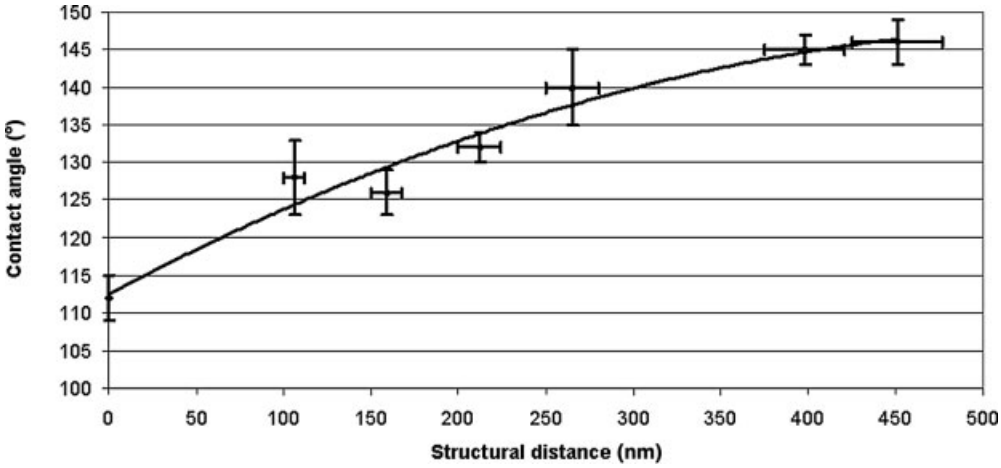


Figure 2.36 Dependence of contact angles of water droplets on PTFE surfaces on the structure size.

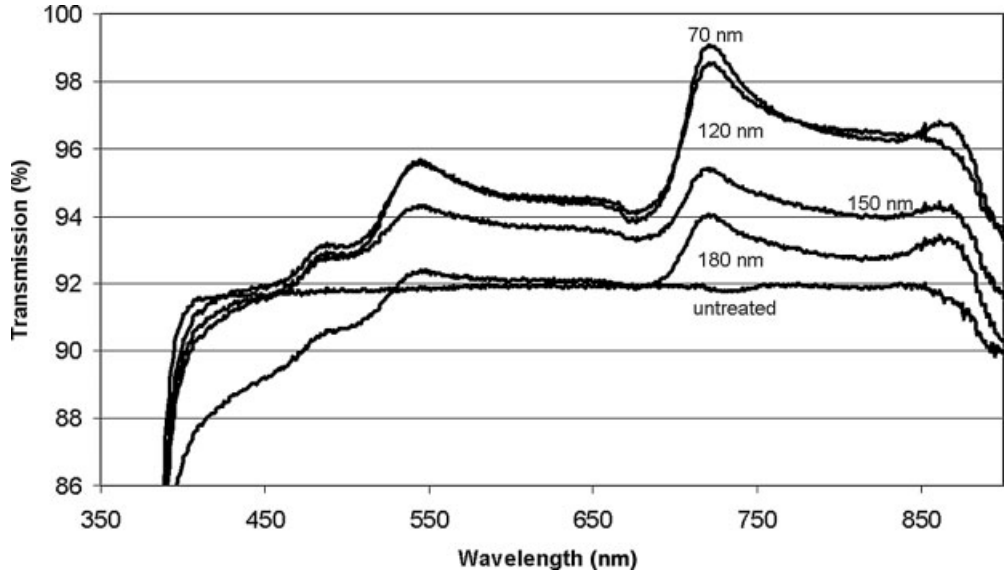


Figure 2.37 UV–visible transmission spectra of PMMA samples structured with 70-, 120-, 150- and 180-nm pillars and of a non-structured sample.

2.4 Final Remarks

Referring to Figure 2.6, only seven of a huge number of possible examples have been selected here to demonstrate the enormous diversity of nanoscience and nanotechnology. Nano-effects can occur everywhere, both in "simple" materials and in complex biological structures. This makes nanoscience and nanotechnology a unique field of research and development. The examples discussed illustrate the universality of this future-determining technology, which in many of the most attractive fields is still at the very beginning. However, it can be predicted that distinct fields that are still part of basic research will develop into techniques which will influence daily life dramatically. Others, usually those of easier and faster research and development, have already become routine techniques.

The selected examples also indicate that basic research is fundamental to developing nanotechnology further. Only by basic research will we discover nano-effects in the different fields named in the definition. However, also in basic science the research assignment should not just be to look for novel nano-effects: physics, chemistry, biology and materials science will also discover relevant effects when working in other fields.

References

1 Nanoscience and nanotechnologies: opportunities and uncertainties, Royal Society and The Royal Academy of Engineering, (2004) 5, *Science Policy Section, The Royal Society, London*.

2 Brune, H., Ernst, H., Grunwald, A., Grünwald, W., Hofmann, H., Krug, H., Janich, P., Mayor, M., Rathgeber, W., Schmid, G., Simon, U., Vogel, V. and Wyrwa, D. (2006) *Wissenschaftsethik und Technikfolgenabschätzung Vol. 27. Nanotechnology – Assessment and Perspectives*, Springer, Berlin.

3 Schwerdtfeger, P. (ed.) (2002) *Relativistic Electronic Structure Theory. Part 1: Fundamentals*, Elsevier, Amsterdam.

4 Schwertdfeger, P. (ed.) (2005) *Relativistic Electronic Structure Theory. Part 2: Applications*, Elsevier, Amsterdam.

5 Hess, B.A. (ed.) (2002) *Relativistic Effects in Heavy-element Chemistry and Physics*, Wiley, New York.

6 Mie, G. (1908) *Annals of Physics*, **25**, 377.

7 Schmid, G. and Giebel, U.unpublished work.

8 Collier, C.P., Mattersteig, G., Wong, E.W., Luo, Y., Beverly, K., Sampaio, J., Raymo, F.M., Stoddart, J.F. and Heath, J.R. (2000) *Science*, **289**, 1172.

9 Schliwa, M. (ed.) (2003) *Molecular Motors*, Wiley-VCH, Weinheim.

10 Tyreman, M.J.A. and Molloy, J.E. (2003) *IEE Proceedings Nanobiotechnology*, **150**, 95.

11 Vallee, R.B. and Hook, P. (2003) *Nature*, **421**, 701.

12 Schliwa, M. and Woehlke, G. (2003) *Nature*, **422**, 759.

13 Ball, P. (2002) *Nanotechnology*, **13**, R15.

14 Braig, K., Otwinowski, Z., Hegde, R., Boisvert, D.C., Joachimiak, A., Horwich, A.L. and Sigler, P.B. (1994) *Nature*, **371**, 578.

15 Ditzel, L., Lowe, J., Stock, D., Stetter, K.-O., Huber, H., Huber, R. and Steinbacher, S. (1998) *Cell*, **93**, 125.

16 Wang, J. and Boisvert, D.C. (2003) *Journal of Molecular Biology*, **327**, 843.

17 Shomura, Y., Yoshida, T., Izuka, R., Maruyama, T., Yohda, M. and Miki, K. (2004) *Journal of Molecular Biology*, **335**, 1265.

18 Shimamura, T., Koike-Takeshita, A., Yokoyama, K., Yoshida, M., Taguchi, H. and Iwatawa, S. (2003) *Acta Crystallographica Section D-Biological Crystallography*, **59**, 1632.

19 Saibil, H.R., Ranson, N.A. (2002) *Trends in Biochemical Sciences*, **27**, 627.

20 Kinbara, K. and Aida, T. (2005) *Chemical Reviews*, **105**, 1377.

21 Vale, R.D., Reese, T.S. and Sheetz, M.P. (1985) *Cell*, **42**, 39.

22 Kozielski, F., Sack, S., Marx, A., Thormahlen, M., Schonbrunn, E., Biou, V., Thompson, A., Mandelkow, E.M. and Mandelkow, E. (1997) *Cell*, **91**, 985.

23 Kikkawa, M., Okada, Y. and Hirokawa, N. (2000) *Cell*, **100**, 241.

24 Kikkawa, M., Sablin, E.P., Okada, Y., Yajima, H., Fletterick, R.J. and Hirokawa, N. (2001) *Nature*, **411**, 439.

25 Hancock, W.O. and Howard, J. (1999) *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 13147.

26 Stewart, R.J., Thaler, J.P. and Goldstein, L.S. (1993) *Proceedings of the National Academy of Sciences of the United States of America*, **90**, 5209.

27 Muthukrishnan, G., Hutchins, B.M., Williams, M.E. and Hancock, W.O. (2006) *Small*, **2**, 626.

28 Koumura, N., Zijistra, R.W.J., van Delden, R.A., Harada, N. and Feringa, B.L. (1999) *Nature*, **401**, 152.

29 Irie, M. (1993) *Molecular Crystals and Liquid Crystals*, **227**, 263.

30 Tseng, H.-R., Vignon, S.A. and Stoddart, J.F. (2003) *Angewandte Chemie-International Edition*, **42**, 1491.

31 Jimenez, M.C., Dietrich-Buchecker, C.O. and Sauvage, J.-P. (2000) *Angewandte Chemie-International Edition*, **39**, 3248.

32 Badjic, J.D., Balzani, V., Credi, A., Silvi, S. and Stoddart, J.F. (2004) *Science* **303**. 1845.

33 Balzani, V. (2005) *Small*, **1**, 278.

34 Moore, G. (1965) *Electronics*, **38**, 114.

35 Wolf, S.A., Treger, D. and Chtchelkanova, A. (2006) *MRS Bulletin*, **31**, 400.

36 Richter, H.J. and Harkness, S.D. IV, (2006) *MRS Bulletin*, **31**, 384.

37 Allenspach, R. and Jubert, P.-O. (2006) *MRS Bulletin*, **31**, 395.

38 Zarembowitch, J. and Kahn, O. (1991) *New Journal of Chemistry*, **15**, 181.

39 Breuning, E., Ruben, M., Lehn, J.-M., Renz, F., Garcia, Y., Knesofontov, V., Gütlich, P., Wegelius, E. and Rissanen, K. (2000) *Angewandte Chemie-International Edition*, **39**, 2504.

40 Feldheim, D.L. and Keating, C.D. (1998) *Chemical Society Reviews*, **27**, 1.

41 Simon, U. (1998) *Advanced Materials*, **10**, 1487.

42 Simon, U. and Schön, G. (2000) in *Handbook of Nanostructured Materials and Nanotechnology* (ed. H.S. Nalwa), Academic Press, New York, Vol. 3, p. 131.

43 Simon, U., Schön, G. and Schmid, G. (1993) *Angewandte Chemie-International Edition in English*, **2**, 250.

44 Simon, U. (2004) in *Nanoparticles. From Theory to Application* (ed. G. Schmid), Wiley-VCH, Weinheim, p. 328.

45 Bezryadin, A., Dekker, C. and Schmid, G. (1977) *Applied Physics Letters*, **71**, 1273.

46 Schmid, G., Boese, R., Pfeil, R., Bandermann, F., Meyer, S., Calis, G.H.M. and van der Velden, J.W.A. (1981) *Chemische Berichte*, **114**, 3634.

47 Schmid, G. (1990) *Inorganic Syntheses, Vol 32*, **7**, 214.

48 Chi, L.F., Hartig, M., Drechsler, T., Schaak, Th., Seidel, C., Fuchs, H. and Schmid, G. (1998) *Applied Physics Letters A*, **66**, 187.

49 Zhang, H., Schmid, G. and Hartmann, U. (2003) *Nano Letters*, **3**, 305.

50 Santini, J.T., Jr. Richards, A.C., Scheidt, R., Cima, M.J. and Langer, R. (2000) *Angewandte Chemie-International Edition*, **39**, 2396.

51 Folkman, J. and Long, D.M. (1964) *Journal of Surgical Research*, **4**, 139.

52 Bar-Shalom, D., Bukh, N. and Larsen, T.K. (1991) *Annals, New York Academy of Sciences*, **618**, 578.

53 Florence, A.T. and Jani, P.U. (1994) *Drug Safety*, **10**, 233.

54 Pereswtoff-Morath, L. (1998) *Advanced Drug Delivery Reviews*, **29**, 185.

55 Davis, S.S. (1999) *Pharmaceutical Science & Technology Today*, **2**, 265.

56 Schmidt, U. (2003) *Spektrum der Wissenschaften*, **10**, 42.

57 Gestermann, S., Hesse, R., Windisch, B. and Vögtle, F. (2000) in *Stimulating Concepts in Chemistry* (eds F. Vögtle, J.F. Stoddart and M. Shibasaki), Wiley-VCH, Weineim. p. 187.

58 Mornet, S., Vasseur, S., Grasset, F. and Duguet, E. (2004) *Journal of Materials Chemistry*, **14**, 2161.

59 Pankhurst, Q.A., Connolly, J., Jones, S.K. and Dobson, J. (2003) *Journal of Physics D*, **36**, 167.

60 Huth, S., Lausier, J., Gersting, S.W., Rudolph, C., Plank, C., Welsch, U. and Rosenecker, J. (2004) *Journal of Gene Medicine*, **6**, 923.

61 Sukhorukov, G.B., Rogach, A.L., Zebli, B., Liedl, T., Skirtach, A.G., Köhler, K., Antipov, A.A., Gaponik, N., Susha, A.S., Winterhalter, M. and Parak, W. (2005) *Small*, **1**, 194.

62 Muños Javier, A., Kreft, O., Piera Alberola, A., Kirchner, C., Zebli, B., Susha, A.S., Horn, E., Kempner, S., Skirtach, A.G., Rogach, A.L., Rädler, J., Sukhorukov, G.B., Benoit, M. and Parak, W.J. (2006) *Small*, **2**, 394.

63 De Geest, B.G., Dégunat, C., Sukhorukov, G.B., Braeckmans, K., De Smedt, S.C. and Demeester, J. (2005) *Advanced Materials*, **17**, 2357.

64 De Geest, B.G., Vandenbroucke, R.E., Guenther, A.M., Sukhorukov, G.B., Hennink, W.E., Sanders, N.N., Demeester, J. and De Smedt, S.C. (2006) *Advanced Materials*, **18**, 1005.

65 Skirtach, A.G., Muños Javier, A., Kreft, O., Köhler, K., Piera Alberola, A., Möhwald, H., Parak, W.J. and Sukhorukov, G.B. (2006) *Angewandte Chemie-International Edition*, **45**, 4612.

66 Kipke, S. and Schmid, G. (2004) *Advanced Functional Materials*, **14**, 1184.

67 Mirkin, C.A., Letsinger, R.L., Mucic, R.C. and Storhoff, J.J. (1996) *Nature*, **382**, 607.

68 Elghanian, R., Storhoff, J.J., Mucic, R.C., Letsinger, R.L. and Mirkin, C.A. (1997) *Science*, **277**, 1078.

69 Rosi, N.L. and Mirkin, C.A. (2005) *Chemical Reviews*, **105**, 1547.

70 Jin, R., Wu, G., Li, Z., Mirkin, C.A. and Schatz, G.C. (2003) *Journal of the American Chemical Society*, **125**, 1643.

71 Storhoff, J.J., Lazarides, A.A., Mucic, R.C., Mirkin, C.A., Letsinger, R.L. and Schatz, G.C. (2000) *Journal of the American Chemical Society*, **122**, 4640.

72 Taton, T.A., Mirkin, C.A. and Letsinger, R.L. (2000) *Science*, **289**, 1757.

73 Cao, Y.W.C., Jin, R. and Mirkin, C.A. (2002) *Science*, **297**, 1536.

74 Georganopoulou, D.G., Chang, L., Nam, J.-M., Thaxton, C.S., Mufson, E.J., Klein, W.L. and Mirkin, C.A. (2005) *Proceedings of the National Academy of Science of USA*, **102**, 2273.

75 Gerion, D., Chen, F.Q., Kannan, B., Fu, A.H., Parak, W.J., Chen, D.J., Majumdar, A. and Alivisatos, A.P. (2003) *Analytical Chemistry*, **75**, 4766.

76 Empedocles, S. and Bawendi, M. (1999) *Accounts of Chemical Research*, **32**, 389.

77 Rosensweig, R.E. (2002) *Journal of Magnetism and Magnetic Materials*, **252**, 370.

78 Jordan, A., Scholz, R., Wust, P., Fahling, H. and Felix, R. (1999) *Journal of Magnetism and Magnetic Materials*, **201**, 413.

79 Jordan, A., Schmidt, W. and Scholz, R. (2000) *Radiation Research*, **154**, 600.

80 Jordan, A., Scholz, R., Maier-Hauff, K., Johannsen, M., Wust, P., Nadobny, P., Schirra, H., Schmidt, H., Deger, S., Loening, S., Lanksch, W. and Felix, R. (2001) *Journal of Magnetism and Magnetic Materials*, **225**, 118.

81 Jordan, A., Rheinlander, T., Waldofner, N. and Scholz, R. (2003) *Journal of Nanoparticle Research*, **5**, 597.

82 Seiyama, T., Kato, A., Fujiishi, K. and Nagatami, M. (1962) *Analytical Chemistry*, **34**, 1502.

83 Seiyama, T. and Kagawa, S. (1966) *Analytical Chemistry*, **38**, 1069.

84 Brattein, W.H. and Bardeen, J. (1953) *Bell System Technical Journal*, **32**, 1.

85 Heiland, G. (1954) *Zeitschrift Fur Physik*, **138**, 549.

86 Franke, M.E., Koplin, T.J. and Simon, U. (2006) *Small*, **2**, 36.

87 Yamazoe, N., Sakai, G. and Shimanoe, K. (2003) *Catalysis Surveys from Asia*, **7**, 63.

88 Samson, S. and Fonstad, C.G. (1973) *Journal of Applied Physics*, **44**, 4618.

89 Jarzebski, Z.M. and Marton, J.P. (1976) *Journal of the Electrochemical Society*, **123**, 299.

90 Maier, J. and Göpel, W. (1988) *Journal of Solid State Chemistry*, **72**, 293.

91 Göpel, W. and Schierbaum, K.D. (1995) *Sensors and Actuators B*, **26–27**, 1.

92 Ogawa, H., Nishikawa, M. and Abe, A. (1982) *Journal of Applied Physics*, **53**, 4448.

93 Bârsan, N. and Weimar, U. (2003) *Journal of Physics-Condensed Matter*, **15**, R813.

94 Sakai, G., Matsunaga, N., Shimanoe, K. and Yamazoe, N. (2001) *Sensors and Actuators B*, **80**, 125.

95 Matsunaga, N., Sakai, G., Shiman, K. and Yamazoe, N. (2003) *Sensors and Actuators B*, **96**, 226.

96 Lu, F., Li, Y., Dong, M. and Wang, X. (2000) *Sensors and Actuators B*, **66**, 225.

97 Wohltjen, H. and Snow, A.W. (1998) *Analytical Chemistry*, **70**, 2856.

98 Snow, A.W. and Wohltjen, H. (2001) US Patent 6 221 673.

99 Evans, S.D., Johnson, S.R., Cheng, Y.L. and Shen, T. (2000) *Journal of Materials Chemistry*, **10**, 183.

100 Han, L., Daniel, D.R., Mayer, M.M. and Zong, C.-J. (2001) *Analytical Chemistry*, **73**, 4441.

101 Krasteva, N., Besnard, I., Guse, B., Bauer, R.E., Muellen, K., Yasuda, A. and Vossmeyer, T. (2002) *Nano Letters*, **2**, 551.

102 Zang, H.-L., Evans, S.D., Henderson, J.R., Miles, R.E. and Shen, T.-H. (2002) *Nanotechnology*, **13**, 439.

103 Vossmeyer, T., Guse, B., Besnard, I., Bauer, R.E., Muellen, K. and Yasuda, A. (2002) *Advanced Materials*, **14**, 238.

104 Joseph, Y., Guse, B., Yasuda, A. and Vossmeyer, T. (2004) *Sensors and Actuators B*, **98**, 188.

105 Simon, U., Flesch, U., Maunz, W., Müller, R. and Plog, C. (1999) *Microporous and Mesoporous Materials*, **21**, 111.

106 Moos, R., Müller, R., Plog, C., Knezevic, A., Leye, H., Irion, E., Braun, T., Marquardt, K.-J. and Binder, K. (2002) *Sensors and Actuators B*, **83**, 181.

107 Franke, M.E., Simon, U., Moos, R., Knezevic, A., Müller, R. and Plog, C. (2003) *Physical Chemistry Chemical Physics*, **5**, 5195.

108 Kreibig, U., Fauth, K., Granquist, C.-G. and Schmid, G. (1990) *Zeitschrift Fur Physikalische Chemie-International Journal of Research in Physical Chemistry & Chemical Physics*, **169**, 11.

109 van Staveren, M.P.J., Brom, H.B. and De Jongh, L.J. (1991) *Physics Reports-Review Section of Physics Letters*, **208**, 1.

110 Simon, U., Schön, G. and Schmid, G. (1993) *Angewandte Chemie-International Edition in English*, **32**, 250.

111 Schön, G. and Simon, U. (1995) *Colloid and Polymer Science*, **273**, 101.

112 Schön, G. and Simon, U. (1995) *Colloid and Polymer Science*, **273**, 202.

113 Brust, M., Betel, D., Shiffrin, D.J. and Kieley, C.J. (1995) *Advanced Materials*, **7**, 795.

114 Andres, R.P., Bielefeld, J.D., Henderson, J.I., Janes, D.B., Kolagunta, V.R., Kubiak, C.P., Mahoney, W.J. and Osifchin, R.G. (1996) *Science*, **273**, 1690.

115 Simon, U. (1998) *Advanced Materials*, **10**, 1487.

116 Torma, V., Schmid, G. and Simon, U. (2001) *ChemPhysChem*, **2**, 321.

117 Torma, V., Vidoni, O., Simon, U. and Schmid, G. (2003) *European Journal of Inorganic Chemistry*, **6**, 1121.

118 Simon, U. (2004), in *Nanoparticles. From Theory to Application* (ed. G. Schmid), Wiley-VCH, Weinheim, p. 328.

119 Schmid, G. and Simon, U. (2005) *Chemical Communications*, 697.

120 Barthlott, W. (1990) in *Scanning Electron Microscopy in Taxonomy and Functional Morphology* (ed. D. Claugher), Clarendon Press, Oxford, p. 69.

121 Barthlott, W. (1993) in *Evolution and Systematics of the Caryophyllales*, (eds H.D. Behnke and T.J. Mabry), Springer, Berlin, p. 75.

122 Barthlott, W. and Neinhuis, C. (1997) *Planta*, **202**, 1.

123 Israelachvili, J. (1995) *Intermolecular and Surface Forces*, 2nd edn, Academic Press, London.

124 Xia, Y., Rogers, J.A., Paul, K.E. and Whitesides, G.M. (1999) *Chemical Reviews*, **99**, 1823.

125 O'Sullivan, J.P. and Wood, G.C. (1970) *Proceedings of the Royal Society, London*, **317**, 511.

126 Thompson, G.E. and Wood, G.C. (1983) *Treatise on Materials, Science and Technology*, **23**, 205.

127 Diggle, J.W., Downie, T.C. and Goulding, C.W. (1969) *Chemical Reviews*, **69**, 365.

128 Masuda, H., Yamada, H., Satoh, M., Asoh, H., Nakao, M. and Tamamura, T. (1997) *Applied Physics Letters*, **71**, 2770.

129 Pelzer, K., Philippot, K., Chaudret, B., Chaudret, W., Meyer-Zaika, W. and Schmid, G. (2003) *Zeitschrift fur Anorganische und Allgemeine Chemie*, **629**, 1217.

130 Schmid, G., Levering, M. and Sawitowski, T. (2007) *Zeitschrift fur Anorganische und Allgemeine Chemie*, **633**, 2147.

3
Top-Down Versus Bottom-Up

Wolfgang J. Parak, Friedrich C. Simmel, and Alexander W. Holleitner

3.1
Introduction

Today, the term "nanotechnology" is a widely used keyword occurring in a variety of different contexts [1, 2]. Semantically, "nano" is an official SI prefix for physical units, which is equivalent to the factor 10^{-9}. This prefactor can be combined with units of physical quantities such as time (nanoseconds), mass (nanograms) or length (nanometers). Historically, nanotechnology has evolved from different scientific fields such as physics, chemistry, molecular biology, microelectronics and material sciences. Generally, nanotechnology aims to study and to manipulate real-world structures with sizes ranging between 1 nm, that is one millionth part of a millimeter, and up to 100 nm. The set of typical "nano"-objects includes colloidal crystals, molecules, DNA-based structures and integrated semiconductor circuits.

We can approach the scale between 1 and 100 nm from different sides, for instance by emphasizing biological and chemical aspects. On the one hand, we can conceptually subdivide complex biological organisms, such as a human body, into smaller and smaller subunits. In a simplified way, the entire body can be classified into organs, each organ is composed out of cells and each cell is a dynamically organized assembly of biological molecules. Arguably, complex molecules, such as lipids, proteins and DNA, are the smallest biologically relevant subunits. These molecules typically have dimensions of a few nanometers up to several tens of nanometers. On the other hand, the smallest integral part of a molecule is the physical bond between individual atoms. The latter have a size of a few tenths of nanometers. Molecules can be very small (such as hydrogen, H_2) and very big (such as polymeric macromolecules). If we leave aside the combination of molecular building blocks to new macromolecules (as in polymerization reactions or in supramolecular chemistry [3]), the typical size range of molecules that can be synthesized from atomic building blocks and their derivatives is again the region from a few nanometers up to several tens of nanometers. It can be concluded that if macroscopic objects such as biological organisms are investigated on a smaller and smaller length scale, one ultimately ends

Nanotechnology. Volume 1: Principles and Fundamentals. Edited by Günter Schmid

ISBN: 978-3-527-31732-5

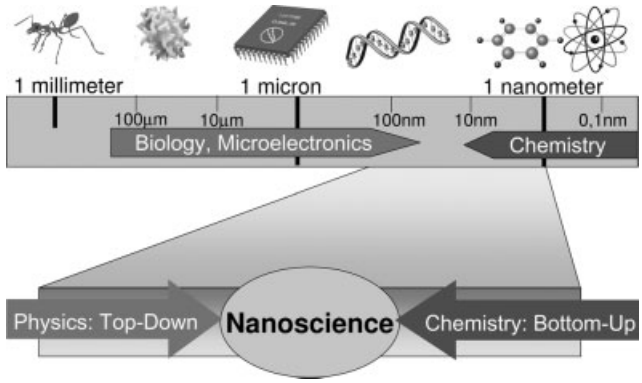


Figure 3.1 Some objects (ant, virus, circuits in computer chip, DNA, benzene, atom) are presented that are characteristic of different size scales ranging from milli- to nanometers. Adapted from a presentation by Professor Dr. J. P. Kotthaus.

up in the nanometer range as the smallest relevant scale for functional subunits. Then again, if materials are built up synthetically from their basic chemical building blocks, one will also first arrive at this length scale. In this sense, "nano" is the size scale where physics, chemistry and biology meet in a natural way (Figure 3.1).

The first scientist who pointed out that "there is plenty of room at the bottom" was Richard Feynman in 1959 [oral presentation given on 29 December 1959 at the Annual Meeting of the American Physical Society at the California Institute of Technology (Caltech)]. He envisioned scientific discoveries and new applications of miniature objects as soon as material systems could be assembled at the atomic scale. To this end, machines and imaging techniques would be necessary which can be controlled at the nanometer or subnanometer scale. Fifty years later, the scanning tunneling microscope and the atomic force microscope are ubiquitous in scientific laboratories, allowing to image structures with atomic resolution [4–12]. As a result, various disciplines within nanotechnology aim towards manufacturing materials for diverse products with new functionalities at the nanoscale.

As we have seen, we can approach the nanoscale from two sides: by making things smaller, that is, by downscaling, and by constructing things from small building blocks, that is, by upscaling. The first method is referred to as the "top-down" and the second as the "bottom-up" approach. The top-down approach follows the general trend of the microelectronic industry towards miniaturization of integrated semiconductor circuits. Modern lithographic techniques allow the patterning of nanoscale structures such as transistor circuits with a precision of only a few nanometers (see http://www.icknowledge.com and publications therein for more information). As we will see in this chapter, the industrial demand for ever smaller electronic circuits has provided several physical tools by which materials can be probed and manipulated at the nanometer scale. In contrast, the bottom-up approach is based on molecular recognition and chemical self-assembly of molecules [13]. In combination with chemical synthesis techniques, the bottom-up approach allows for the assembly

of macromolecular complexes with a size of several nanometers. To visualize the different strategies of these two approaches is the topic of this chapter. We first will give a few simple explanations and arguments. Then we will illustrate the differences of the two strategies in the context of a variety of examples. Finally, we will discuss the limits of both approaches.

3.1.1 Top-Down Strategies

How can we get to smaller and smaller sculptures and objects? The idea of top-down-strategies is to take processes known from the macroscopic world and to adopt them in such a way that they can be used for doing the same thing on a smaller scale. Since ancient times, humans have created artwork and tools by structuring materials. Let us take the artist as an example, who carves and sculpts figures out of blocks of wood or stone, respectively. For this purpose, the artist needs tools, usually a carving knife or chisel, with which he can locally ablate parts from the original piece of material and thus give it its desired shape (Figure 3.2). If we continue with the idea of a sculptor, it is evident that in order to sculpt smaller objects, smaller tools are needed, such as miniature rasps and knives. Chinese artisans have used such tools to carve little pieces of wood and other materials into sculptures in the submillimeter regime, which can only be seen with magnifying glasses [some of these highly impressive sculptures can be seen in the National Taiwan Museum (http://www.ntm.gov.tw)].

If we want to fabricate even smaller structures or objects and study their physical properties, classical mechanical tools such as rasps will no longer work. For that purpose, scanning probe microscopes are powerful instruments for probing and manipulating materials at the nanometer scale and can thus be seen as one of the key inventions for nanotechnology. On the one hand, they allow imaging and probing of the characteristics of nanoscale objects with the highest resolution. Examples include topology, material configuration and electrical, chemical and magnetic properties of the studied objects. On the other hand, scanning probe microscopes allow local manipulation and even shaping of the nanostructures. In a seminal work in 1982, Binnig and Rohrer invented the scanning tunneling microscope (STM) [4], for which they were awarded the Nobel Prize in 1986. In such a microscope, piezoelectric

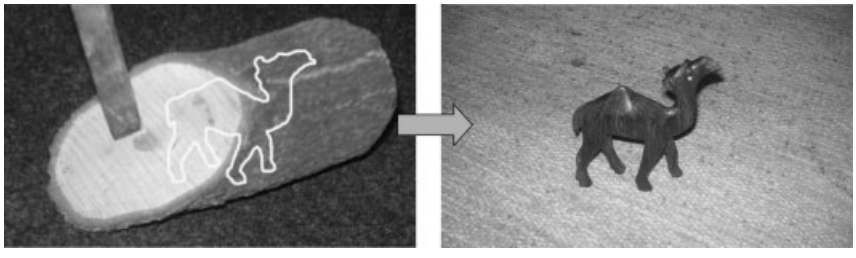


Figure 3.2 An artifact is made out of a piece of wood by carving. The wooden figure is created by locally ablating material from the initial block according to a building plan.

crystals move a scanning tip across the surface of a sample, while the electric current is recorded between the tip and the sample. If the tip is located very close to the surface of the sample, the electric current is composed of tunneling electrons. In the nanoscale quantum world, the wave character of electrons plays an important role. As soon as the distance between the tip and the surface is on the order of the electron wavelength, electrons can tunnel from the tip to the surface. The size of the tunneling current has an exponential dependence on the distance between the tip and surface. Therefore, the point of the tip which is closest to the surface predominantly contributes to the current. In principle, this point can be made up of only one atom, which allows for atomic resolution. The scanning tunneling microscope is sensitive to the electronic density at the surface of the sample, which allows for, for example, imaging of the electronic orbitals of atoms and molecules. The scanning tunneling microscope can operate not only under ultrahigh vacuum conditions, but also at atmospheric pressure and at various temperatures, which makes it a unique imaging and patterning tool for the nanosciences. For instance, scanning tunneling microscopes are utilized to pattern nanostructures by moving single atoms across surfaces, while the corresponding change of the quantum mechanical configuration can be recorded *in situ* [14].

The atomic force microscope (AFM) operates similarly to the scanning tunneling microscope [5]. Here, the force between the scanning tip and the sample surface is extracted by measuring the deflection of the tip towards the sample. Again, the atomic force microscope can be utilized as an instrument to image and to shape materials on the nanometer scale [15].

In experiments in which only a few nanostructures need to be patterned and probed, scanning probe microscopes can be exploited to structure and shape materials at the nanometer scale. To define an array of millions of nanoscale systems in parallel – such as in integrated electronic circuits – the top-down approach of microlithography is the technique of choice (Figure 3.3). Microlithography has been the technological backbone of the semiconductor industry for the last 45 years (see http://www.icknowledge.com and publications therein for more information). The minimum feature size of optically defined patterns depends on the wavelength of the utilized light and also factors that are due to, for example, the shape of the lenses and the quality of the photoresist. In 2003, the typical linewidth of semiconductor circuits fell below 100 nm, that is, the semiconductor industry can be literally seen as being part of the nanotechnologies. For this achievement, argon fluoride excimer lasers are applied with a laser wavelength of 193 nm in the deep ultraviolet region. For lithography in this optical range, tricks such as "optical proximity correction" and "phase shifting" were invented and successfully implemented. On the one hand, the miniaturization of semiconductor circuits is limited by economic costs for the semiconductor industry, since the implementation of new techniques for the realization of ever smaller feature sizes results in ever increasing costs. On the other hand, physical material properties, such as the high absorption level of refractive mirrors at short optical wavelengths, set a natural limit of a few tens of nanometers for the miniaturization process. Since the 1980s, the decline of optical lithography has been predicted as being only a few years away. However, each time optical

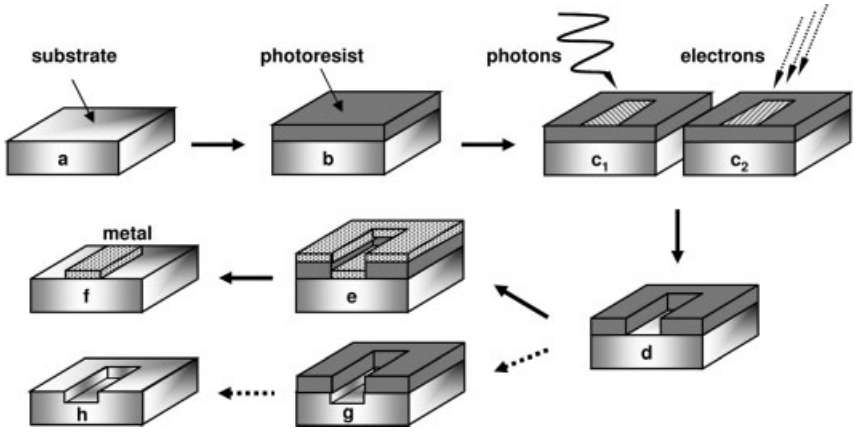


Figure 3.3 Top-down microlithography: in (a) and (b) a so-called photoresist – a liquid, photosensitive polymer – is spin-coated on to the polished surface of a substrate. (c) After hardening of the photoresist, for example, by heating, the substrate with the photoresist layer on top is locally exposed to photons or to electrons. Thereby, patterns can be defined in the photoresist. (d) After a photoresist specific developing process, the exposed (or the unexposed) parts of the original substrate are bared. By follow-up processes, such as metallization (e) or etching (g), the patterns in the photoresist can be translated to the substrate. To this end, the remaining parts of the photoresist are finally removed by a solvent, such as acetone [(f) and (h)]. Thereby, metallic conductor paths, logical circuits or memory cells can be defined.

lithography had reached traditional limits, new techniques extended the economically sustainable lifetime of top-down microlithography.

At present, there are several possible successor top-down nanotechnologies for industry, for example, extreme ultraviolet light lithography (EUV), electron beam lithography with multicolumn processing facilities [see also Figure 3.3(c_2)], the focused ion beam (FIB) technique and the ultraviolet nano-imprinting technique [16]. The implementation of each of these techniques requires enormous technical challenges to be overcome. One of the most promising techniques is that using EUV light with a wavelength of only 13 nm. For this technique, fabrication errors of the "optical" components need to be in the nanometer or subnanometer range. For comparison, state-of-the-art X-ray telescopes, such as the Zeiss XMM Newton telescope, exhibit a granularity of the mirror surfaces of 0.4 nm (see W. Egle, *Mission Impossible: XMM-Newton Proves the Opposite*, Innovation, Carl Zeiss, 2000, *8*, 12–14, ISSN 1431-8059; this file can be downloaded from http://www.zeiss.com). In addition to "state of the art" optical requirements, all metrological components of the EUV technique need to exhibit subnanometer resolution. As a result, the cost of a stepper machine for EUV exposure of photoresists is US$50 million per system, providing a linewidth of 35 nm (see http://www.icknowledge.com for more information).

For medium-sized businesses, the "ultraviolet nano-imprinting technique" seems to be the most promising method to fabricate nanoscale circuits [17]. Here, nanostructures are mechanically imprinted into a photoresist. The stamp with the

nanoscale patterns is made out of fused quartz, a material which is transparent to ultraviolet light. As soon as the stamp has been plunged into the photoresist, a short ultraviolet light pulse causes the photoresist to polymerize along the patterns on the stamp. In line with optical lithography, follow-up processes allow for the definition of nanoscale circuits in various geometries [Figure 3.3(e)–(h)]. The minimum feature size of the nano-imprinting technique is about 10 nm (see the webpage of the Chou group at Princeton University: http://www.princeton.edu/~chouweb/newproject/page3.html). At the same time, the technique is applicable to metals and plastic materials. The costs of an industrial nano-imprinting machine are less than US$1 million. However, the throughput of nano-imprinting machines is much lower than that for stepper machines.

What we have illustrated above is the top-down approach. We take tools and methods known in the macroscopic world and scale them down to structure materials on increasingly smaller length scales. As explained above, this downscaling process is not simply a reduction in size of the tools – we cannot make a nanoscale carving knife that directly resembles a macroscopic knife to take away material on the nanometer scale. In most cases, the relative importance of the forces and interactions involved changes with decreasing system size – and this also has to be considered when constructing the tools themselves.

3.1.2 Bottom-Up Strategies

The antipode of the top-down approach is the so-called bottom-up technique. Here a complex structure is assembled from small building blocks. These building blocks possess specific binding capabilities – often termed "molecular recognition properties" – which allow them to arrange automatically in the correct way. Self-assembly is an essential component of bottom-up approaches [18]. The ultimate examples of molecular recognition are biological receptor–ligand pairs: molecules that recognize and bind to each other with very high specificity. Prominent examples of such pairs are antibodies and their corresponding antigens and complementary strands of deoxyribonucleic acid (DNA) [19].

We can visualize the bottom-up assembly of materials with the example of LEGO building blocks, a common toy for children. Again we take the example of a small sculpture as also used to describe top-down approaches. LEGO building blocks can have different functions, as symbolized by their color (Figure 3.4). Furthermore, there are building blocks of different size and each building block has a defined number of binding sites, realized here as knobs. In this way the blocks can only be attached to each other in a defined way. This can be seen as "molecular" recognition. However, we should point out that the example of the LEGO blocks fails to describe self-assembly. There is still a helping hand needed to assemble the individual bricks to form the complete structure – the assembly process is not spontaneous (cf. the self-assembly models of Whitesides [20]). The example of LEGO blocks hold the basic concept of a bottom-up strategy: the construction of a new material by assembling basic small building blocks. If instead of LEGO blocks nanoscopic building blocks are

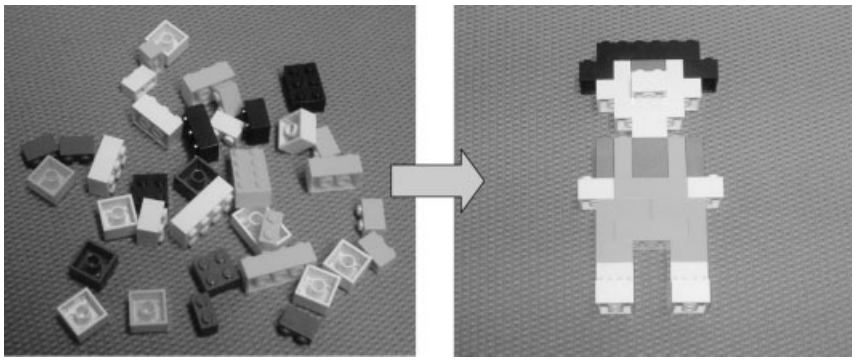


Figure 3.4 LEGO blocks can symbolize different functionalities (colors) and they can be assembled with their knobs in a defined way. Assembly of all building blocks (bricks) leads to the formation of the desired structure.

used, structures on the nanometer scale can be assembled in a very analogous way [21–24].

Generally, bottom-up assembly techniques seek to fabricate composite materials comprising nanoscale objects which are spatially ordered via molecular recognition. The prime examples of the technique are self-assembled monolayers (SAMs) of molecules [25]. A substrate is immersed in a dilute solution of a surface-active organic material that adsorbs on the surface and organizes via a self-assembly process. The result is a highly ordered and well-packed molecular monolayer. The method can be extended towards layer-by-layer (LBL) assembly, by which polymer light-emitting devices (LEDs) have already been fabricated [17]. The self-assembly technique also allows positioning of single molecules between two metal electrodes and subsequently into an experimental circuit. By this set-up, quantum mechanical transport characteristics of single molecules, such as photochromic switching behavior, can be studied in order to build electronic devices with new functionalities [26].

There are several combinations and variations of the SAM technique. It can be combined with nano-imprinting methods, the atomic force microscope or the focused ion beam technique, allowing the fabrication of geometric patterns of molecules with variable wetting properties, chemical functionality and/or topological characteristics [17]. Since most of the processes are performed in solution, electrical fields may be utilized to assemble charged compounds in a directed way, for example, for the creation of functionalized sensor electrodes. Electrodes modified with negatively charged gold nanoparticles and positively charged host molecules have proven highly sensitive for the detection of, for example, adrenaline (as the guest molecule) [27]. Another very promising field is the DNA-directed assembly of network materials [28]. Here, molecular recognition reactions between single DNA strands are translated into aggregate formation of nano-objects such as nanoparticles [21, 22]. Self-assembly can also be used to construct simple nanomechanical devices. For instance, the molecular recognition between DNA molecules has been exploited to build “nanotweezers” [29]. Further, networks of materials can be

synthesized on top of so-called block copolymer templates [30]. A typical application of this technique is the formation of networks of metallic nanowires, that is, metals are vapor deposited on to a preformed template matrix made out of copolymers. The polymer networks can be used as two- or even three-dimensional templates. Generally, there is a wide range of materials which can be engineered by bottom-up techniques. The research involved has led to numerous sensing, electronic and optoelectronic interfaces in addition to devices.

3.2
First Example: Nanotweezers

Among the fundamental tools for mechanical work are tweezers or fingers. The basic function of tweezers is to hold and release things. In principle, tweezers comprise two fingers, which can close to hold something and open to release it again. So far only a few functional nanofingers are available. Although a variety of tools exist which can hold nanometer-sized objects, the problem is to release them again. This problem is of fundamental rather than of technological nature. The interested reader is referred to the excellent articles by Nobel Prize winner Richard Smalley, who describes the problem of nonspecific adhesion as the problem of "sticky fingers" [31, 32]. The basic module of a nanofinger is a nanometer-sized hinge, that is, a structure which can repeatedly open and close. In this section, we will show examples of such nanofingers or hinges based on top-down microlithography and on bottom-up self-assembly of biological molecules.

3.2.1
Top-Down Nanotweezers

Lithographically defined tweezers have been realized as a so-called micro-electromechanical system (MEMS). Similar micron-sized mechanical resonators have already been applied in inkjet printers and in airbags of cars. MEMS resonators have also triggered major scientific interest, since the reduction in size of an electromechanical resonator down to the nanometer regime would allow researchers to study, for example, mechanical effects in the quantum realm [33–36] or the coupling of the eigenmode of such a nanoscale resonator to the electromagnetic field of a photon [37, 38]. Furthermore, the small mass of nanoelectromechanical devices makes them extremely sensitive adsorptive sensors, since the mass of adsorbed molecules can significantly change the eigenfrequency of a nanoscale mechanical device [39, 40].

For lifting and manipulation of objects on the nanoscale, it is necessary to use opposing forces to seize and hold a floating or a freely suspended nanostructure, such as a single protein in liquids or an individual nanowire within a three-dimensional electrical circuit. Kim and Lieber reported the first nanoscale tweezers in operation [42]. They attached two carbon nanotubes to metal electrodes on opposite sides of a micron-thick glass needle, in order to manipulate nanoscale clusters and nanowires with a size of about 500 nm. The tweezers were closed electrostatically using a voltage

difference between the nanotubes. Generally, nanotweezers can act as manipulators, sensors and injectors. For instance, they can investigate the interaction between nanomaterials and they can measure the electrical conductance of nanostructures by employing the two probes of the tweezer as electrodes. The tweezers combine a mechanical degree of control of a two probe sensing instrument with the spatial resolution of a scanning probe microscope. To date, scanning tunneling microscopy and atomic force microscopy have evolved to a level where individual atoms on surfaces can be arranged into the shape of nanoscale letters [43] and clusters of atoms can be pushed into tiny junctions to make quantum devices [44]. In order to move a nanoscale object with a single probe, there needs to be sufficient adhesion to the probe to ensure that the object can be picked up and then transported to its new destination. To deposit the object at the new location, the adhesion at the final site needs to be larger than the adhesion to the single probe. One way to overcome the problem is to use tweezers – nanoelectromechanical probes, which are attached to the tips of an atomic force microscope [45].

In most cases nanotweezers are closed and opened electrostatically using a voltage difference between the two nanoprobes. The electrical field and the potential difference between the probes may cause problems, especially when biological samples are investigated. Here, mechanical manipulation of the nanotweezers is favorable. Recently, piezo-driven tweezers have been developed with a size of about 1 μm that can grab, hold and transport bacteria in addition to micron-sized particles in liquids [46]. The experiment demonstrates that lithographically defined tweezers are a promising alternative to optical tweezers commonly applied for the manipulation of biological samples [47]. The latter utilize the effect that a focused laser beam provides an attractive force for dielectric objects.

One of the smallest tweezers devices – with a gap of only 25 nm – was fabricated by a technique which combines conventional silicon microlithography with electron beam deposition of carbon [48]. How carbon nanofingers can be fabricated by this technique is shown schematically in Figure 3.5 [41]. The fabrication process follows a classical top-down approach by microstructuring silicon with lithographic techniques [49]. To this end, silicon is partially covered by a protective polymer layer (see Figure 3.3 for top-down lithography). Then, the unprotected areas are etched away [Figure 3.5(a)–(c)]. The process uses a heterostructure with an intermediate silicon oxide layer, which can be selectively etched by hydrofluoric acid. As depicted in Figure 3.5(d), the technique enables freely suspended silicon fingers to be defined. Finally, a carbon nanofinger is defined on top of the freely suspended structure by focusing an electron beam on the silicon nanostructure in the presence of residual gases [Figure 3.5(e)] [41]. As a result, the electron beam-deposited carbon builds up to form the carbon nanofingers [Figure 3.5(f)]. The functioning of the resulting device [Figure 3.6(a) and (b)] is very intuitive. It looks like tweezers known in everyday life, only orders of magnitude smaller. By applying a voltage to the inner two, freely suspended silicon electrodes, the gap between the carbon nanofingers can be opened and closed. Most importantly, the carbon nanofingers are nonconductive. No voltage difference is applied between the nanofinger tips, making the device ideal for application with such fragile structures as organic objects.

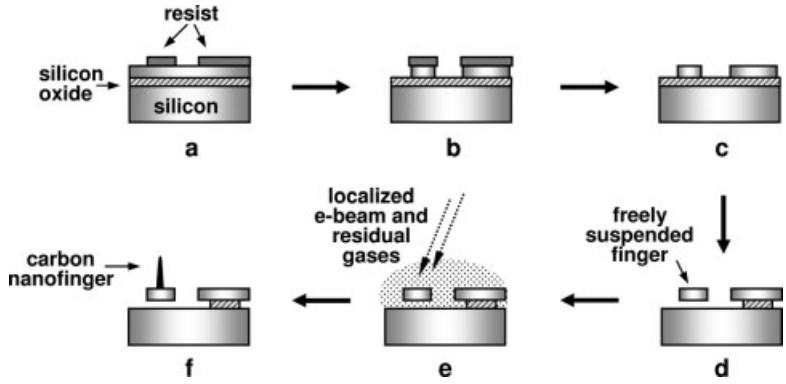


Figure 3.5 Fabrication of a freely suspended carbon nanofinger: (a) the photo- or e-beam resist is prepatterned by microlithographic steps on top of a multilayered silicon-on-insulator (SOI) wafer. The latter consists of a silicon top layer, an intermediate silicon oxide layer and a silicon substrate. (b), (c) In accordance with the pattern, the silicon top layer is removed either by wet-chemical or by reactive ion etching. (d) By the use of hydrofluoric acid, the silicon oxide layer is partially removed. Thereby, the smaller features of the prestructured silicon top layer become freely suspended. (e), (f) If an electron beam is focused on specific locations of the structure in the presence of residual gases, an electron beam-deposited carbon nanofinger can be formed [41].

As already mentioned, the release of objects picked up by the tweezers constitutes a severe conceptual problem. An object can only be released when it is deposited at a site to which its adhesion is higher than the adhesion to the nanofingers. This concept is based on hierarchical binding forces. Although this concept would allow for repeatedly picking up objects from one site A, transporting them to site B and releasing them there (in the case that $F_A < F_F < F_B$, where F_A, F_F and F_B represent the binding forces of the object to site A, to the nanofinger and to site B, respectively), it would fail in the reverse direction to transport objects from B to A (as F_B is larger than F_F, the holding force of the nanofinger would not be strong enough to detach the object from binding site B). Alternatively, the attachment of the object to the

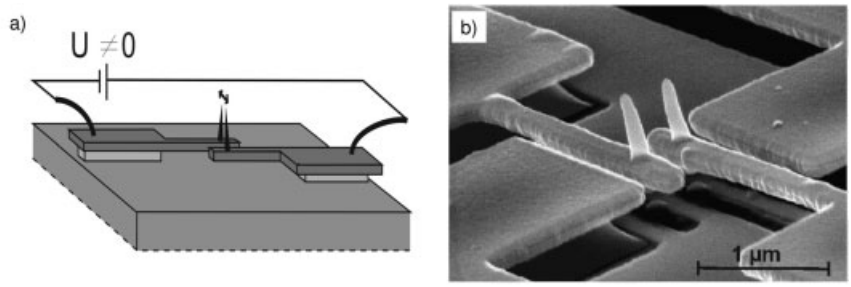


Figure 3.6 (a) By applying an electric voltage to the conductive silicon arms, the "fingers" move. (b) Electrically operated nanotweezers. The structure is made out of silicon with lithographic techniques. The tips on the silicon arms have subsequently been grown with electron beam deposition (EBD). The figure is based on the original work of Dr. Christine Meyer, Dr. Bert Lorenz and Professor Dr. K. Karrai and is reproduced with their permission [41].

nanofinger might be modulated by a biological molecule, for example, an antibody whose conformation can be modulated by light. In this way, the binding properties of the antibody could be remotely controlled by switching on or off the light and thus objects could be held and released. However, so far this concept has not been realized practically.

3.2.2 Bottom-Up Nanotweezers

In recent years, the unique biochemical and mechanical properties of DNA have been increasingly utilized for non-biological applications, for example, to realize tiny nanomechanical devices. Many of these devices exploit the different rigidity of single- and double-stranded DNA and can be switched back and forth between several structures by the addition or removal of DNA "fuel" strands. One of the prototype devices in this field are DNA nanotweezers [50]. Their operational principle is shown in Figure 3.7. In the open state, the DNA tweezers consist of three strands of DNA.

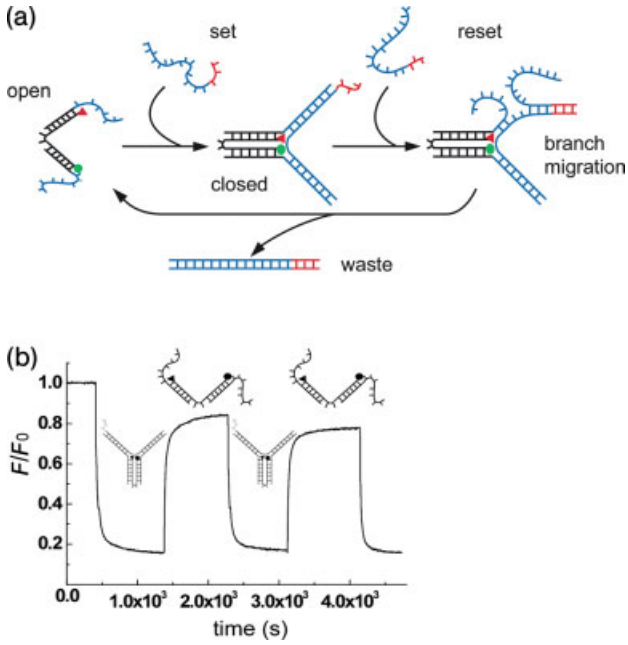


Figure 3.7 DNA nanotweezers. (a) Principle of operation. In the open state, three strands are hybridized together to form a molecular structure resembling tweezers. Due to hybridization with a "fuel" strand, the arms of the tweezers can be closed. Another DNA strand can be used to remove the fuel strand and restore the original open state via branch migration. (b) FRET experiment. The movement of the tweezers can be followed in fluorescence measurements. To this end, the arms of the tweezers are fluorescently labeled. One of the labels quenches the fluorescence of the other due to fluorescence resonance energy transfer. As this effect is strongly distance dependent, the motion of the arms is accompanied by a change in fluorescence intensity.

One central DNA strand is hybridized to two other strands in such a way that together they form two roughly 6-nm long, rigid double-stranded "arms" connected by a short, single-stranded flexible "hinge". In the open state, each of the arms still has single-stranded extensions available for hybridization. The addition of a long "fuel" strand which is complementary to these extensions can then be used to close the tweezer structure, that is, the two arms are forced together by the hybridization with the fuel strand. The device can be switched back to its original configuration with a biochemical "trick". In the closed state, a short single-stranded section of the device is deliberately left unhybridized. These nucleotides serve as an attachment point for an "anti-fuel" strand, which is exactly complementary to the fuel strand. A biochemical process known as "branch migration" unzips the structure, when fuel and anti-fuel try to bind with each other. After completion of this process, a waste "duplex" is ejected and the DNA tweezers have returned to their open configuration again. The device can be operated cyclically by the alternate addition of set and reset strands. The transition between the different states of the device can be characterized in fluorescence measurements, utilizing the distance dependent quenching of fluorophores due to fluorescence resonance energy transfer (FRET).

The same operation principle has since been used in many other DNA devices [51]. A simple variation of the DNA tweezers is the DNA actuator device [52]. Here, instead of two single-stranded extensions, the arms of the tweezers are connected by a single-stranded loop. Depending on the sequence of the fuel strands, the device can either be closed similarly to the tweezers or stretched into an elongated conformation. Whereas the intermediate configuration is a rather floppy structure (like the open tweezers), the closed and the stretched configurations are much more rigid. While these devices resemble macroscopic tweezers in shape, they cannot actually be used to "grab" nano-objects. However, such a function may be achieved by the incorporation of so-called "aptamers" into DNA devices [53]. These are special DNA or RNA sequences with a high binding affinity for other molecules, for example, proteins.

3.3 Second Example: Nanomotors

Another important mechanical element is the motor. A motor is a device that can generate periodic movements and carry a load with it. Similarly to the previous section, we will show two examples of nanomotors. The version using the top-down approach is again based on microlithography and follows the ideas of micromechanical engineering. The version using the bottom-up approach is based on functional organic molecules.

3.3.1 Top-Down Nanomotors

Nanomotors can also be created using refined lithographic techniques. As the strategies employed are somehow similar to those described in Section 3.2.1, we

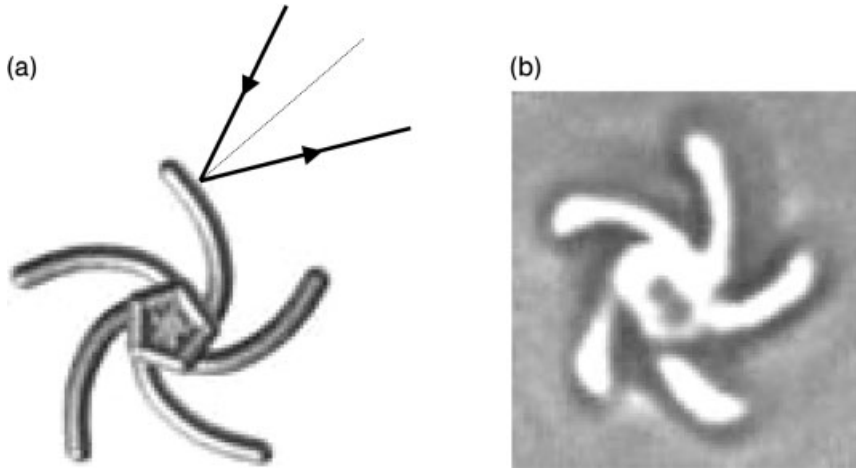


Figure 3.8 (a) A micrometer-sized rotor is driven by the momentum transferred due to reflection of an incident light beam. (b) Phase contrast image of the rotor in solution. Adopted from an paper by the group of Professor Dr. P. Ormos and is used with his permission. Details can be found in the original article [55, 56].

will only give two brief examples in this section. As with macroscopic motors, nanomotors also periodically convert input energy into mechanical work. The energy to drive the motor can originate from different sources, such as light, electric fields or chemical gradients.

A simple light-driven propeller has been demonstrated by Ormos's group. The propeller itself is created by selectively illuminating the parts of a light curing resin that resemble the shape of the propeller. The resin is cured at the illuminated regions due to photo-polymerization. Dissolution of the noncured parts of the resin finally results in a freestanding propeller (Figure 3.8) [54–57]. The propeller can be driven by incident light into rotation. When light is reflected at the arms of the propeller, momentum is transferred which causes the rotation. For a detailed description of the underlying physics we refer to the original article by Galajda and Ormos [55]. In order to achieve controlled rotation of the propeller, either a freestanding propeller is trapped in the focus of laser tweezers [55, 56] or a propeller bound to an axis on top of a substrate is driven by the light origination from an integrated waveguide [54]. So far the smallest propellers created in this way still have a size of a few micrometers.

An even smaller synthetic motor has been created by Zettl's group by scaling down MEMS technology to the nanometer scale [58]. The axis of this motor is formed by a carbon nanotube which is fixed between two anchor electrodes (Figure 3.9). A metal plate fixed to the carbon nanotube acts as a rotor. The outer shell of the nanotube beneath the rotor has been detached from the inner shells of the nanotube by shear forces so that a rotational bearing has been formed. The rotor is driven by an oscillating electric field between three additional electrodes. In practice, the rotor was

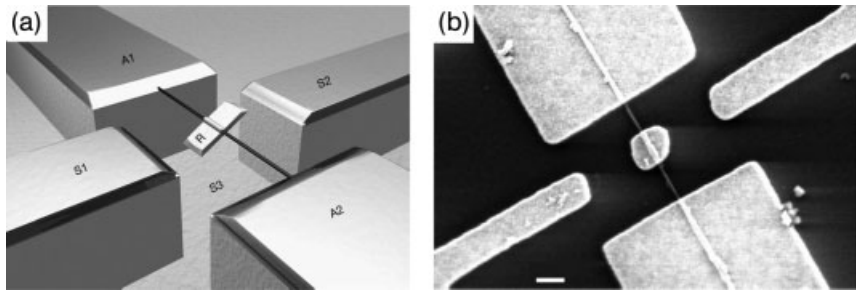


Figure 3.9 (a) Sketch of the nanomotor. A metal plate rotor (R) has been attached to a multi-walled carbon nanotube of which the outer shell beneath the metal plate can rotate freely around the inner shells and which is anchored on two electrodes (A1, A2). The motor is driven by electric fields applied from three stator electrodes, two on the SiO_2 surface (S1, S2) and one buried beneath the surface (S3). (b) Scanning electron microscope image of the nanomotor. Image reproduced with permission from the original article by Zettl's group [58].

realized by first depositing a multi-walled carbon nanotube on a silicon oxide substrate, followed by patterning the rotor and the electrodes using electron beam lithography and by etching down the silicon oxide below the rotor. More details can be found in the original breakthrough article by Zettl *et al.* [58].

3.3.2 Bottom-Up Nanomotors

Motors play an important role not only in our technical macroscopic life, but also on a molecular level in any biological organism. Molecular motors, for example, help bacteria to move and to transport cargo inside cells, and they are also the basis for contracting and expanding our muscles [59]. Since the field of nanotechnology emerged, scientists have thought about how to harness molecular motors for the construction of artificial machines. For example, the molecular motor kinesin has been used to transport colloidal quantum dots along the tracks of microtubules [60]. Such a concept might eventually lead to a scenario where building blocks could be transported along rails to their designated positions. Another example is the membrane-bound protein ATP synthase [61]. This rotary motor is driven by a proton gradient and is used in cells for the synthesis of ATP [61]. In the reverse direction the motor uses the energy of ATP hydrolysis to create a proton gradient. The rotation of the motor could be used to propel an actin filament that had been attached to the upper subunit of the motor with ATP as fuel [62]. Such a concept might eventually be used as a nanopropeller for moving small vesicles. While these concepts are based on using naturally existing molecular motors, synthetic molecular motors have also been chemically synthesized [63]. In this section we will briefly describe the concept and realization of such a motor that has been demonstrated by Gaub's group [64].

Molecules can often assume different structural arrangements called "conformations". Azobenzene, for example, can be reversibly switched upon illumi-

nation at two different excitation wavelengths between an extended *trans* and a shorter *cis* conformation [65]. In order to obtain longer effective length changes upon switching between the two different conformations, many azobenzene molecules can be linearly connected to one long polyazobenzene molecule. Single force spectroscopy techniques allow for attaching single molecules between the tip of an atomic force microscopy (AFM) cantilever and a substrate that is placed on top of a piezo-actuator [66]. In a periodic cycle, a single polyazobenzene molecule fixed between the tip and piezo of an AFM can now be elongated and shortened by illumination, which can be used to lift a load (as realized by the pulling force of the piezo-actuator) (Figure 3.10). In this way, in each cycle light energy is converted into mechanical energy [64]. Although the estimated efficiency of the motor is relatively low (total mechanical energy output/total optical energy input $\approx 10^{-4}$), it nevertheless demonstrates nicely how a light-driven artificial nanomotor can look.

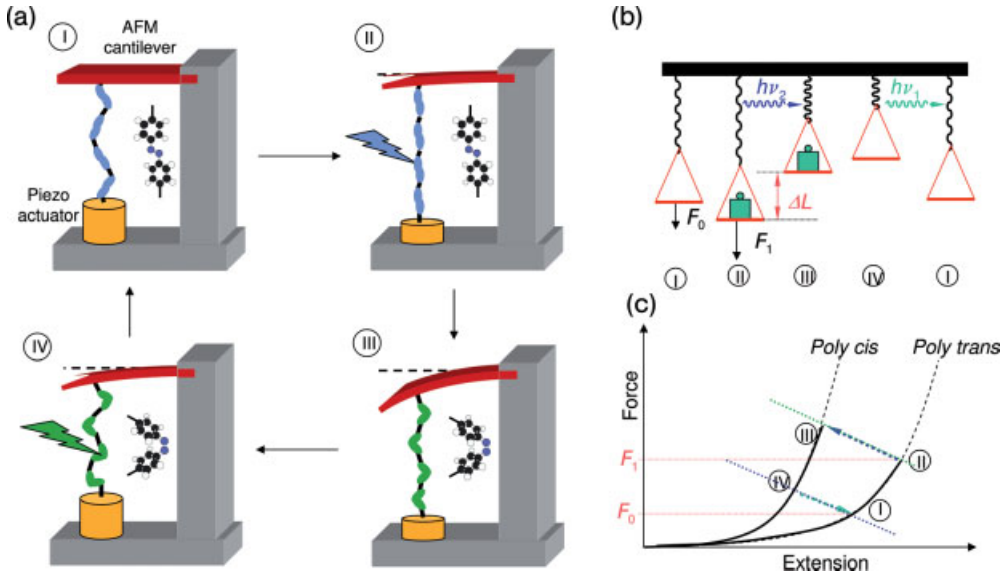


Figure 3.10 (a) A single polyazobenzene molecule is fixed between the tip of an AFM cantilever and a glass substrate mounted on top of a piezo-actuator. In a periodic cycle first the polyazobenzene is stretched by applying a load by retracting the piezo-actuator (I ⇒ II). Then, upon a light flash $h\nu_2$ the polyazobenzene is brought from the stretched *trans* to the shortened *cis* configuration (II ⇒ III). The load is now released by moving the piezo-cantilever upwards and thus the polyazobenzene can elongate (III ⇒ IV). With another light flash $h\nu_1$ the polyazobenzene is driven back to the extended *trans* configuration (IV ⇒ I). (b) Visualization of the load that is applied to the polyazobenzene by moving the piezo-actuator. (c) The cyclic process can be described in a force–extension diagram, where *F* is the force applied to the polyazobenzene by stretching it between the AFM cantilever and the glass substrate on top of the piezo and the extension describes the effective length of the polyazobenzene. Adapted from a presentation by Professor Dr. H. Gaub; details can be found in the original publication by Gaub's group [64].

3.4 Third Example: Patterning

One basic requirement for many technologies and applications is the ability to form controlled patterns on a surface. This can be exemplified with the components and techniques needed to fabricate an electronic circuit [67]. In order to make circuits, one essentially has to master three different steps. First, active elements such as transistors must be realized which are able to process information [68]. Second, the active elements have to be arranged into a functional geometry [28]. Third, the active elements have to be connected by wires [69]. Arranging elements on a surface is basically the same as forming a pattern. Another example is controlled cell attachment. If a surface is patterned partly with molecules that promote cell adhesion and partly with molecules that repel cells, cells only will grow on the desired parts of the surface. This is very important, for example, for bioelectronic interfaces, where cells have to be guided in such a way that they adhere on top of the active electronic elements, but not to other parts of the surface [70].

In this section we will briefly describe two examples for making small surface patterns. Soft lithography is a top-down approach, whereby surfaces can be structured with lithographically generated stamps. Following a self-assembly strategy, two-dimensional lattices of biological molecules, in particular DNA, can be formed in a bottom-up approach.

3.4.1 Soft Lithography

Soft lithography is an increasingly popular application of traditional lithographic techniques for the ordered deposition of "soft" materials such as small molecules, biomacromolecules or even live cells on surfaces. A variety of different techniques can be summarized under the term "soft lithography", for example, microcontact printing, replica molding and micromolding in capillaries [71, 72]. Most of these techniques utilize the soft, rubber-like material polydimethylsiloxane (PDMS). A schematic depiction of a typical soft lithographic process is shown in Figure 3.11. Usually the desired structure is first patterned into a "hard" material using conventional lithography. In Figure 3.11(a)–(c) a pattern defined on a mask is transferred into a negative resist on a silicon substrate. The structured resist film can be used as a template to transfer the patter into PDMS. To this end [Figure 3.11(d)–(f)], polymer precursors are poured over the template, followed by a curing step. After complete polymerization of the PDMS, the resulting "soft" structure can be peeled off from the substrate. There are many different applications for the structured PDMS material. In microcontact stamping, for example, the PDMS is used as a stamp with which molecules can be transferred on to a substrate in a pattern. Examples are shown in Figure 3.11(g) and (h), where fluorescently labeled proteins and also live cells have been patterned into an array of squares. PDMS-based structures are not only utilized for patterning, but have also been extensively used for the definition of microfluidic systems [73].

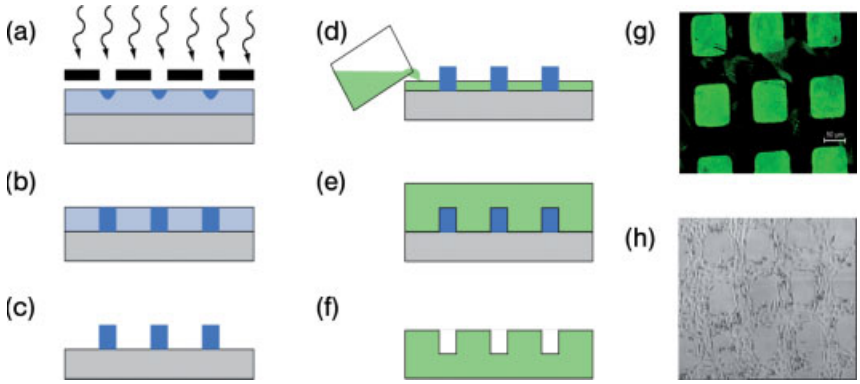


Figure 3.11 Soft lithography and microcontact printing of molecules. (a)–(c) Using conventional photolithography, a pattern is defined in a negative photoresist (blue) on a silicon wafer (gray). After exposure and development of the resist, a relief pattern remains on the surface. (d)–(f) This pattern can be used as a "master template" to create flexible microstamps from PDMS. Polymer precursors are poured over the master and cured. The resulting stamp can be peeled off the master and then used for soft lithographic processes such as stamping. (g) A square pattern of fluorescently labeled proteins stamped on a glass surface. (h) Live cells only reside outside of the square pattern which has been stamped with hydrophobic molecules which do not promote cell adhesion.

3.4.2
Two-Dimensional DNA Lattices

Whereas soft lithography can be mainly regarded as a top-down method to arrange molecules into a particular pattern, a different approach relies on molecular recognition events and self-assembly. As mentioned before, interactions between complementary strands of DNA can be utilized for the "programmable" assembly of molecular structures. Starting with the synthesis of a DNA molecule with the topology of a cube by Chen and Seeman in 1991 [74], DNA has been used to construct molecules with the structure of a truncated octahedron [75], DNA catenanes [76], tetrahedra and octahedra [77, 78] and other geometric objects. DNA has also been used to construct two-dimensional molecular lattices from DNA branched junctions [19, 79–81]. One of the central building blocks of such lattices is the so-called "double-crossover" (DX) construction. In a DX molecule, two double-stranded DNA molecules exchange strands twice, creating an entity which can be regarded as the molecular analogue of a LEGO building block [Figure 3.12(a)]. The rigid DX molecule has dimensions on the order of 4×15 nm. Depending on the sequence of single-stranded "sticky ends" at the corners of the DX molecules, they can arrange into large two-dimensional lattices [Figure 3.12(b) and (c)]. In fact, the assembly of such DX "tiles" closely resembles a computational process – regarding DNA self-assembly as the execution of a molecular "program" can actually be exploited to realize more complex molecular structures than with any other self-assembly technique [82]. When these networks fold back to themselves, they can also form DNA "nanotubes" [83–85]. Recently, such networks have even been utilized to arrange nanoparticles and proteins in two dimensions

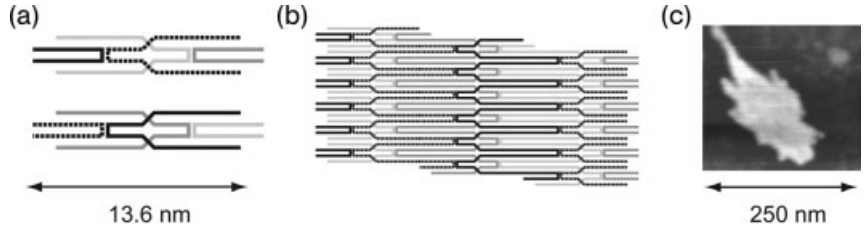


Figure 3.12 DX assembly. (a) The basic building blocks – DX molecules – are composed of two DNA double strands which exchange strands twice, creating a structure held together by two DNA crossovers. (b) Depending on the "sticky ends" at the corners of the DX tiles, a variety of two-dimensional lattices may be formed in a self-assembly process. (c) A small patch of a 2D DX lattice imaged with an atomic force microscope under air. Functionalization of these lattices allows for the arrangement of nanoparticles or proteins into sequence-programmed patterns.

[28, 86–89] and the impressively powerful "DNA origami" scheme [82] will soon allow the arrangement of these nanoparticles into arbitrary geometries and patterns.

3.5
Fourth Example: Quantum Dots[1)]

In Chapter 4, the physical principles of quantum dots are described. Quantum dots are arguably the ultimate examples of nanometer structures. In this section, we will show that we can manufacture quantum dots both with top-down (Section 3.5.2) and with bottom-up techniques (Section 3.5.4). Depending of the manufacturing process used, the properties of the quantum dots vary and we will explain this in particular with respect to their optical properties. Also the possible applications vary among the different types of quantum dots.

3.5.1
Different Methods for Making Quantum Dots

Here we will give a brief overview of the most popular methods used to fabricate quantum dots in practice. Lithographically defined quantum dots are the classical example of the application of a top-down method, whereby the quantum dot is created by locally etching away parts of the raw material. Colloidal quantum dots, on the other hand, are an example of a bottom-up approach, in which they are assembled from small building blocks (in this case surfactant-stabilized atoms).

The ultimate technique for the fabrication of quantum dots should be able to produce significant amounts of sample, with such high control of quantum dot size, shape and monodispersity that single-particle properties are not averaged out by sample inhomogeneity. So far, ensembles of quantum dots produced by the best

1) This paragraph has been adopted from a previous edition and the authors acknowledge their former coauthors Dr. Liberato Manna, Dr. Daniele Gerion and Prof. Dr. Paul Alivisatos 90.

available techniques still show a behavior deriving from a distribution of sizes, but this field is evolving very rapidly. In this section we give a short survey of the most popular fabrication approaches. Different techniques lead to different typologies of quantum dots. The confinement can be obtained in several different ways and in addition the quantum dot itself can have a peculiar geometry, it can be embedded into a matrix or grown on a substrate or it can be a "free" nanoparticle. Each of these cases is strictly related to the preparative approach chosen.

3.5.2
Lithographically Defined Quantum Dots

Lithographically defined quantum dots are formed by isolating a small region of a two-dimensional electron system (see Chapter 4) via tunneling barriers from its environment. Such two-dimensional electron systems (2DES) or 2D electron gases (2DEGs) can be found in metal–oxide–semiconductor field effect transistors (MOSFETs) or in the so-called semiconductor heterostructures [91, 92]. Heterostructures are composed of several thin layers of different semiconductor materials grown on top of each other, using molecular beam epitaxy (MBE). The layer sequence can be chosen in such a way that all free charge carriers are confined to a thin slice of the crystal, forming essentially a two-dimensional electron system. A superstructure derived from the periodic repetition of this sequence of layers is also called a "multiple quantum well". One of the most widely investigated systems is, for instance, the aluminum gallium arsenide/gallium arsenide (AlGaAs/GaAs) quantum well. AlGaAs has the same lattice constant as GaAs but a wider bandgap, whose exact value depends on the aluminum content of the layer. Therefore, electrons in the GaAs layer are confined to this layer and form a two-dimensional electron gas.

Quantum dot systems can be generated in a lateral or in a vertical arrangement, as shown in Figure 3.13. In the lateral geometry, the 2DEG is locally electrostatically depleted by applying negative voltages to electrodes deposited on top of the crystal. We can understand this effect with the following argument. Let us assume that we apply a negative voltage to the metal electrodes above the two-dimensional electron gas. Due to electrostatic repulsions, electrons will be repelled by the electric field of the electrodes and the region of the 2DEG below the electrodes will be depleted of electrons. A charge-depleted region behaves like an insulator. Therefore, by applying an electric field with metal electrodes of an appropriate shape, it is possible to create an island of charges insulated from the rest of the 2DEG. If small enough, the island within the 2DEG behaves as a quantum dot. In the vertical geometry, a small pillar of the 2DEG is isolated by etching away the heterostructure around it. In such an arrangement charge carriers are again confined in all three dimensions.

Most of the electron transport measurements on quantum dots that have been performed to date have used the two types of quantum dots as samples that we have just described. The lateral arrangement offers a relatively high degree of freedom for the design of the structure, as this is determined by the choice of the electrode geometry. In addition, it is possible to fabricate and study "artificial molecules" [93–99] composed of several quantum dots linked together. With the

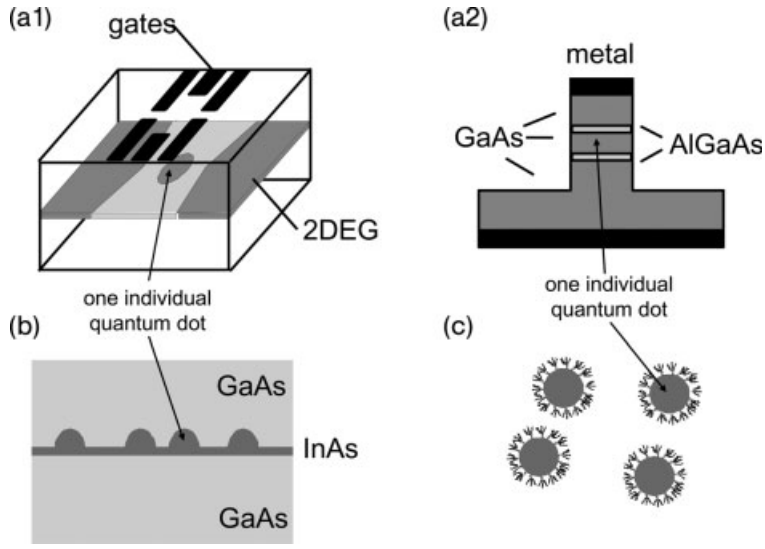


Figure 3.13 Three different types of quantum dots. (a1) A lithographically defined quantum dot in a lateral arrangement can be formed by electrostatic depletion of a two-dimensional electron gas (2DEG, shown in dark gray) via gate electrodes. The 2DEG is formed typically 20–100 nm below the surface of a semiconductor heterostructure (usually GaAs/AlGaAs). Application of negative voltages to metal gates on top of the heterostructure depletes the 2DEG below the gates (shown in light gray) and cuts out a small electron island from the 2DEG. Electrons can still tunnel on to and from the island. Electrical contact to the 2DEG is realized through ohmic contacts (not shown). (a2) A vertical quantum dot can be formed in a double barrier heterostructure. A narrow pillar is etched out of a GaAs/AlGaAs/GaAs/AlGaAs/GaAs heterostructure. The AlGaAs layers (light gray) form tunnel barriers that isolate the central GaAs region from the contact region. This central GaAs region behaves now as a quantum dot (shown in dark gray). Electrical contact is made via metal contacts (depicted in black) on top of the pillar and below the heterostructure. (b) Self-assembled quantum dots: molecular beam epitaxy (MBE) growth of InAs (dark gray) on GaAs (light gray) first leads to the formation of an extended layer of InAs (the wetting layer) and then to the formation of small InAs islands. Single electrons or electron–hole pairs (excitons) can be confined in these InAs quantum dots, either electrically or optically. (c) Colloidal quantum dots: these colloidal particles, having a diameter of only a few nanometers, are formed using wet chemistry and can be produced for most of the type II–VI, III–V, IV–VI and some type IV semiconductors. The surface of colloidal quantum dots is coated with a layer of surfactant molecules that prevents aggregation of the particles.

vertical arrangement, structures with very few electrons can be realized [100]. Recently, several research efforts have been focused on the investigation of many-body phenomena in these quantum dot systems. Relevant examples are, for instance, the study of the Kondo effect [101–104] and the design and control of coherent quantum states with the ultimate goal of quantum information processing [105, 106].

A remarkable advantage of lithographically defined quantum dots is that their electrical connection to the "macro-world" is straightforward. The manufacturing processes are similar to those used in chip fabrication and in principle such structures could be embedded within conventional electronic circuits. However, as the geometry

of these quantum dots is determined lithographically, it is restricted by the usual size and resolution limits of lithographic techniques. Even by using electron beam lithography for the creation of the quantum dots, it is not possible to tailor their size with nanometer precision. Lithographically fabricated quantum dots are typically larger than 10 nm and so relatively low lateral confining energies can be achieved.

3.5.3 Epitaxially Self-Assembled Quantum Dots

A breakthrough in the field of epitaxially grown nanostructures was the discovery of epitaxial growth regimes that favored the formation of nanometer-sized islands of semiconductor materials on suitable substrates (Figure 3.13). These islands, exhibiting quantum dot behavior, are obtained naturally by epitaxially growing a thin layer of a low bandgap material over a higher bandgap material, using MBE or MOCVD techniques [107–110]. The respective crystal faces in contact must have a significant lattice mismatch (1–8%), as in the case of InAs on GaAs [111, 112] and Ge on Si [113]. During the growth, a strained film, called the "wetting layer", initially forms. The maximum thickness of this layer is related to the difference in the lattice constants between the two materials. Past this critical thickness, a 2D → 3D transition in the growth regime is observed, with the spontaneous formation of an array of nanometer-sized islands (Stranski–Krastanov regime), leading to partial release of the strain. If the growth is not interrupted at this step, misfit dislocations form because the energy of formation of these defects becomes smaller than the elastic energy accumulated in the strained film. The formation of dislocations in highly strained epilayers (when the lattice mismatch is of the order of 10% or more) before the formation of islands limits the range of possible substrate–island materials. The shape of the islands can be controlled by the growth conditions. Usually, the islands have a truncated pyramidal shape, but it is also possible to form, for example, ring-shaped quantum dots [114]. The final step consists in the growth, on the top of the islands, of several layers of the substrate material, so that the dots are completely buried and interfaces are passivated. The relative alignment of the bandgaps creates a confining potential for charge carriers that accumulate inside the quantum dots. In addition, strain fields in the proximity of the island–substrate interface, due to the lattice mismatch between the two materials, create potentials that modify the bandgap of the quantum dots at the bottom of the island. Holes are more likely to be localized in this region, as they are heavier than the electrons.

Self-assembled quantum dots can have a diameter as small as a few nanometers and so very pronounced quantum size effects can be observed in these systems. Self-assembled quantum dots have predominantly been characterized using optical or capacitance spectroscopy in a regime where they contain only a small number of charge carriers. Measurements on ensembles still suffer from inhomogeneous broadening of the spectroscopic features. However, in recent years it has been possible to look at only a few or even single self-assembled quantum dots at a time by reducing the number of quantum dots by mesa-etching [115] or by using confocal microscopy techniques [116]. Photoluminescence from single self-assembled quantum dots is a highly efficient process, characterized by several narrow emission lines,

related to different exciton states in the dots, and is reminiscent of the emission from atoms. As mentioned already for lithographically defined quantum dots, many parallels can be drawn between atoms and quantum dots [115, 117–119]. For these reasons, quantum dots have gained the nickname "artificial atoms". Current research efforts are devoted to quantum dot ordering and positioning and also to the reduction of the quantum dot size distribution. In contrast to the case of lithographically defined quantum dots, it is challenging to make electrical contact to self-assembled quantum dots and therefore most of the possible applications can be found in optics. One of the major goals of research on self-assembled quantum dots is the fabrication of non-classical light sources from single dots. Another is to use them as light-addressable storage devices.

3.5.4 Colloidal Quantum Dots

Colloidal quantum dots are remarkably different from the quantum dot systems mentioned above as they are chemically synthesized using wet chemistry and are free-standing nanoparticles or nanocrystals grown in solution (Figure 3.13) [120]. In this case, colloidal quantum dots are just a subgroup of a broader class of materials that can be synthesized at the nanoscale using wet chemical methods. In the fabrication of colloidal nanocrystals, the reaction chamber is a reactor containing a liquid mixture of compounds that control the nucleation and the growth. In a general synthesis of quantum dots in solution, each of the atomic species that will form the nanocrystal building blocks is introduced in the reactor as a precursor. A precursor is a molecule or a complex containing one or more atomic species required for growing the nanocrystal. Once the precursors have been introduced into the reaction flask they decompose, forming new reactive species (the monomers) that will cause the nucleation and the growth of the nanocrystals. The energy required to decompose the precursors is provided by the liquid in the reactor, either by thermal collisions or by a chemical reaction between the liquid medium and the precursors or by a combination of the two mechanisms [121].

The key parameter in the controlled growth of colloidal nanocrystals is the presence of one or more molecular species in the reactor, broadly termed here "surfactants". A surfactant is a molecule that is dynamically adsorbed on the surface of the growing quantum dot under the reaction conditions. It must be mobile enough to provide access for the addition of monomer units, but stable enough to prevent the aggregation of nanocrystals. The choice of surfactants varies from case to case: a molecule that binds too strongly to the surface of the quantum dot is not suitable, as it would not allow the nanocrystal to grow. On the other hand, a weakly coordinating molecule would yield large particles or aggregates [122]. Some examples of suitable surfactants include alkanethiols, phosphines, phosphine oxides, phosphates, phosphonates, amides, amines, carboxylic acids and nitrogen-containing aromatics. If the growth of nanocrystals is carried out at high temperatures (e.g., at 200–400 °C) then the surfactant molecules must be stable under such conditions in order to be a suitable candidate for controlling the growth.

At low temperatures, or more generally when the growth is stopped, the surfactants are more strongly bound to the surface of the nanocrystals and provide their solubility in a wide range of solvents. This coating allows for great synthetic flexibility in that it can be exchanged with another coating of organic molecules having different functional groups or polarity. In addition, the surfactants can be temporarily removed and an epitaxial layer of another material with different electronic, optical or magnetic properties can be grown on the initial nanocrystal [123, 124].

By controlling the mixture of surfactant molecules that are present during the generation and the time of growth of the quantum dots, excellent control of their size and shape is possible [121, 125–127] (Figure 3.14). As described in Chapter 4, the wavelength of emission of quantum dots depends on their size. This can be nicely seen by observing the fluorescence light of solutions of colloidal quantum dots with different size [128, 129] (Figures 3.14 and 3.15).

In contrast to organic fluorophores, colloidal quantum dots have a continuous absorption spectrum for energies higher than their bandgap, a symmetric emission

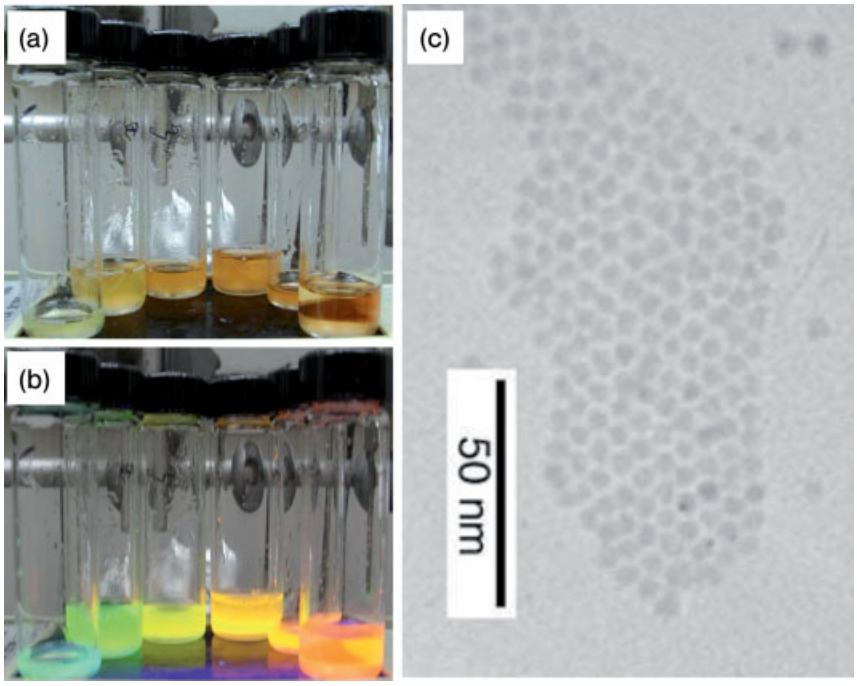


Figure 3.14 Colloidal CdSe quantum dots of different size dissolved in chloroform. The size of the quantum dots increases from the left to the right vial. (a) Photograph of the solutions. (b) Photograph of the solutions upon UV illumination from below. The different colors of fluorescence can be seen. (c) Each quantum dot corresponds to a spherical CdSe nanoparticle that is dispersed in the solvent. When a drop of this solution is put on a grid, the solvent evaporates and the quantum dots can be imaged with transmission electron microscopy (TEM). The authors are grateful to Stefan Kudera for recording the TEM image.

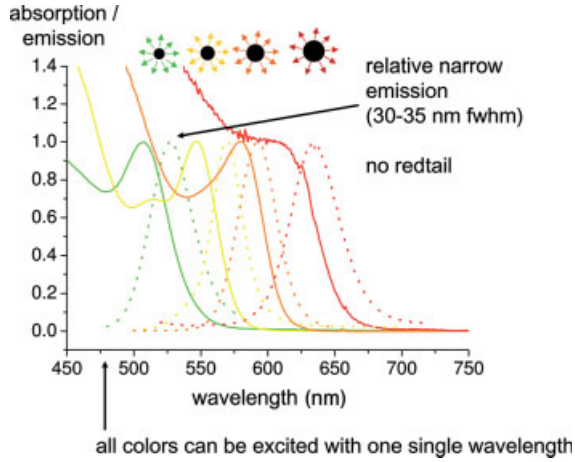


Figure 3.15 Absorption (solid lines) and emission spectra (dotted lines) of colloidal CdSe quantum dots of different sizes. The absorption peaks of green/yellow/orange/red fluorescent nanocrystals of 2.3/4.0/3.8/4.6 nm diameter are at 507/547/580/605 nm and the fluorescence peaks are at 528/57/592/637 nm.

spectrum without a red tail and, most importantly, reduced photobleaching [130, 131] (Figure 3.15).

Since colloidal nanocrystals are dispersed in solution, they are not bound to any solid support as is the case for the other two quantum dots systems described above. Therefore, they can be produced in large quantities in a reaction flask and later they can be transferred to any desired substrate or object. It is possible, for example, to coat their surface with biological molecules such as proteins or oligonucleotides. Many biological molecules perform tasks of molecular recognition with extremely high efficiency. This means that ligand molecules bind with very high specificity to certain receptor molecules, similarly to a key-and-lock system. If a colloidal quantum dot is tagged with ligand molecules, it specifically binds to all the positions where a receptor molecule is present. In this way it has been possible, for example, to make small groupings of colloidal quantum dots mediated by molecular recognition [22, 23, 132] and to label specific compartments of cell with different types of quantum dots [133–135].

Although colloidal quantum dots are rather difficult to connect electrically, a few electron transport experiments have been reported. In these experiments, nanocrystals were used as the active material in devices that behave as single electron transistors [68, 136].

3.6
Perspectives and Limits of Top-Down and Bottom-Up Approaches

In 1965, Gordon Moore, co-founder of Intel Corporation, predicted that the number of transistors on a computer chip would double about every 18 months [137]. The

exponential law, also known as Moore's first law, has described the development of integrated circuits surprisingly well for decades. As the market for information technology continues to grow, the demand for computer hardware instigates more and more sophisticated top-down techniques to build more densely packed transistor circuits. Moore's second law states that the implementation of a next generation of integrated circuits at minimum cost will be exponentially more expensive. Until all the constraints finally limit the growth of the semiconductor top-down industry, scientists and engineers assume that nanotechnology will give answers to most of the technological challenges. For instance, as soon as the feature size of the semiconductor transistors reaches the level at which quantum phenomena are important, different concepts for the assembly need to be considered. One possibility is the bottom-up approach, which is based on molecular recognition and chemical self-assembly of molecules [13]. In combination with chemical synthesis techniques, the bottom-up approach allows the assembly of macromolecular complexes with new functionalities. Assuming Moore's laws apply [137], the final limit for optical top-down lithography is likely to be reached in less than a decade.

There are limitations also for the bottom-up assembly of complex nanostructures. We can illustrate this by the example of assembling nanoparticles with DNA to groupings of particles. Although building blocks exist in which each nanoparticle is modified with an exactly defined number of binding sites, still no controlled assemblies of more than around five particles exist. There are two fundamental technological problems: nonspecific adsorption and "floppyness" of biological molecules. Nonspecific adsorption causes particles to stick together, although they are not supposed to be connected. Many biological molecules that can be used as "glue" for the assembly of particle groupings, such as proteins and DNA, tend to adsorb nonspecifically on the surface of nanoparticles. Although covalent attachment of these molecules dominates over nonspecific adsorption and thus connection of the particles via their designated binding sites, there is a non-negligible amount of nonspecific interaction between the particles. For a larger particle grouping, just a single nonspecific connection can destroy the build-up of the whole grouping. One major task in this direction is to improve the surface chemistry of particles in order to obtain inert surfaces to which biological molecules do not adsorb nonspecifically. Biological molecules are intrinsically "soft" compared with inorganic materials. This implies that the connection between particles that are connected via biological molecules will always retain a certain degree of flexibility. In particular, the attachment of the "glue", that is, the biological molecule, to the particle surface is not rigid. Therefore, it will be almost impossible to form large, nonperiodic three-dimensional stiff structures with static geometry of particle groupings connected via biological molecules. Further, particle assemblies involving biological molecules as glue will be always limited in stability. Biological molecules are bound to their natural environment and cannot withstand many artificial conditions, such a high temperatures.

Although both top-down and bottom-up strategies have clear limitations, we are still far away from having reached them. In fact, the intrinsics may finally be overcome by combining both approaches. Still, Feynman's statement is true: "there is plenty of room at the bottom!"

References

1 Whitesides, G.M. (2005) Nanoscience, nanotechnology and chemistry. *Small*, **1**, 172–179.

2 Whitesides, G.M. (1998) Nanotechnology: art of the possible. *Technology Review*, **101**, 84–87.

3 Lehn, J.M. (2004) Supramolecular chemistry: from molecular information towards self-organization and complex matter. *Reports on Progress in Physics*, **67**, 249–265.

4 Binnig, G., *et al*. (1982) Surface studies by scanning tunneling microscopy. *Physical Review Letters*, **49**, 57–61.

5 Binnig, G., Quate, C.F. and Gerber, C. (1986) Atomic force microscope. *Physical Review Letters*, **56**, 930–933.

6 Gimzewski, J.K. and Joachim, C. (1999) Nanoscale science of single molecules using local probes. *Science*, **283**, 1683–1688.

7 Lieber, C.M., Liu, J. and Sheehan, P.E. (1996) Understanding and manipulating inorganic materials with scanning probe microscopes. *Angewandte Chemie International Edition in English*, **35**, 687–704.

8 Poggi, M.A., *et al*. (2004) Scanning probe microscopy. *Analytical Chemistry*, **76**, 3429–3443.

9 Wiesendanger, R. (1997) Scanning-probe-based science and technology. *Proceedings of the National Academy of Sciences of the United States of America*, **94**, 12749–12750.

10 Wiesendanger, R. (1994) Contributions of scanning probe microscopy and spectroscopy to the investigation and fabrication of nanometer-scale structures. *Journal of Vacuum Science & Technology B*, **12**, 515–529.

11 Friedbacher, G. and Fuchs, H. (1999) Classification of scanning probe microscopies (technical report). *Pure and Applied Chemistry*, **71**, 1337–1357.

12 Quate, C.F. (1997) Scanning probes as a lithography tool for nanostructures. *Surface Science*, **386**, 259–264.

13 Wilbur, J.L. and Whitesides, G.M. (1999) Self-assembly and self-assembled monolayers in micro- and nanofabrication, in *Nanotechnology*, (ed. G. Timp), Springer, New York.

14 Crommie, M.F., Lutz, C.P. and Eigler, D.M. (1993) Imaging standing waves in a 2-dimensional electron-gas. *Nature*, **363**, 524–527.

15 Irmer, B., *et al*. (1998) Josephson junctions defined by a nanoplough. *Applied Physics Letters*, **73**, 2051–2053.

16 Tennant, D.M. (1999) Limits of conventional lithography, in Nanotechnology, in *Nanotechnology*, (ed. G. Timp), Springer, New York.

17 Di Ventra, M., Evoy, S. and Heflin, J.R. (2004) *Introduction to Nanoscale Science and Technology*, Kluwer, Dordrecht.

18 Whitesides, G.M. and Grzybowski, B. (2002) Self-assembly at all scales. *Science*, **295**, 2418–2421.

19 Winfree, E., *et al*. (1998) Design and self-assembly of two-dimensional DNA crystals. *Nature*, **394**, 539–544.

20 Bowden, N., *et al*. (1997) Self-assembly of mesoscale objects in ordered two-dimensional arrays. *Science*, **276**, 233–234.

21 Alivisatos, A.P., *et al*. (1996) Organization of "nanocrystal molecules" using DNA. *Nature*, **382**, 609–611.

22 Mirkin, C.A., *et al*. (1996) A DNA-based method for rationally assembling nanoparticles into macroscopic materials. *Nature*, **382**, 607–609.

23 Zanchet, D., *et al*. (2002) Electrophoretic and structural studies of DNA-directed Au nanoparticle groupings. *Journal of Physical Chemistry B*, **106**, 11758–11763.

24 Sperling, R.A., *et al*. (2006) Electrophoretic separation of nanoparticles with a discrete number of functional groups. *Advanced Functional Materials*, **16**, 943–948.

25 Love, J.C., *et al*. (2005) Self-assembled monolayers of thiolates on metals as a

form of nanotechnology. *Chemical Reviews*, **105**, 1103–1169.

26 Reed, M.A., *et al.* (1997) Conductance of a molecular junction. *Science*, **278**, 252–254.

27 Lahav, M., Shipway, A.N. and Willner, I. (1999) Au-nanoparticle–bis-bipyridinium cyclophane superstructures: assembly, characterization and sensoric applications. *Journal of the Chemical Society-Perkin Transactions 2*, 1925–1931.

28 Zheng, J., *et al.* (2006) Two-dimensional nanoparticle arrays show the organizational power of robust DNA motifs. *Nano Letters*, **6**, 1502–1504.

29 Simmel, F.C. and Yurke, B. (2004) DNA-based nanodevices, in *Encyclopedia of Nanoscience and Nanotechnology*, (ed. H.S. Nalwa), American Scientific Publishers, Stevenson Ranch, 495–504.

30 Chan, V.Z-H., *et al.* (1999) Ordered bicontinuous nanoporous and nanorelief ceramic films from self assembling polymer precursors. *Science*, **286**, 1716–1719.

31 Smalley, R.E. (2001) Chemie, Liebe und dicke Finger. *Spektrum der Wissenschaft Spezial Nanotechnologie 2*, November 2001, 66–67.

32 Smalley, R.E. (September, 2001) Of chemistry, love and nanobots. *Scientific American*, 76–77.

33 Cleland, A.N. and Roukes, M.L. (1996) Fabrication of high frequency nanometer scale mechanical resonators from bulk Si crystals. *Applied Physics Letters*, **69**, 2653–2655.

34 Carr, D.W. and Craighead, H.G. (1997) Fabrication of nanoelectromechanical systems in single crystal silicon using silicon on insulator substrates and electron beam lithography. *Journal of Vacuum Science & Technology B*, **15**, 2760–2763.

35 Craighead, H.G. (2000) Nanoelectromechanical systems. *Science*, **290**, 1532–1535.

36 Knobel, R.G. and Cleland, A.N. (2003) Nanometre-scale displacement sensing using a single electron transistor. *Nature*, **424**, 291–293.

37 Karrai, K. (2006) Photonics – a cooling light breeze. *Nature*, **444**, 41–42.

38 Höhberger Metzger, C. and Karrai, K. (2004) Cavity cooling of a microlever. *Nature*, **432**, 1002–1005.

39 Ilic, B., Yang, Y. and Craighead, H.G. (2004) Virus detection using nanoelectromechanical devices. *Applied Physics Letters*, **85**, 2604–2606.

40 Ekinci, K.L., Huang, X.M.H. and Roukes, M.L. (2004) Ultrasensitive nanoelectromechanical mass detection. *Applied Physics Letters*, **84**, 4469–4471.

41 Meyer, C., Lorenz, H. and Karrai, K. (2003) Optical detection of quasi-static actuation of nanoelectromechanical systems. *Applied Physics Letters*, **83**, 2420–2422.

42 Kim, P. and Lieber, C.M. (1999) Nanotube nanotweezers. *Science*, **286**, 2148–2150.

43 Eigler, D.M. and Schweizer, E.K. (1990) Positioning single atoms with a scanning tunneling microscope. *Nature*, **344**, 524–526.

44 Carlsson, S.B., *et al.* (1999) Mechanical tuning of tunnel gaps for the assembly of single-electron transistors. *Applied Physics Letters*, **75**, 1461–1463.

45 Akita, S., *et al.* (2001) Nanotweezers consisting of carbon nanotubes operating in an atomic force microscope. *Applied Physics Letters*, **79**, 1691–1693.

46 Jericho, S.K., *et al.* (2004) Micro-electro-mechanical systems microtweezers for the manipulation of bacteria and small particles. *Review of Scientific Instruments*, **75**, 1280–1282.

47 Ashkin, A., *et al.* (1986) Observation of a single-beam gradient force optical trap for dielectric particles. *Optics Letters*, **11**, 288–290.

48 Boggild, P., *et al.* (2001) Fabrication and actuation of customized nanotweezers with a 25 nm gap. *Nanotechnology*, **12**, 331–335.

49 Blick, R.H., *et al.* (2002) Nanostructured silicon for studying fundamental aspects

of nanomechanics. *Journal of Physics-Condensed Matter*, **14**, R905–R945.

50 Yurke, B., *et al*. (2000) A DNA-fuelled molecular machine made of DNA. *Nature*, **406**, 605–608.

51 Simmel, F.C. and Dittmer, W.U. (2005) DNA nanodevices. *Small*, **1**, 284–299.

52 Simmel, F.C. and Yurke, B. (2002) A DNA-based molecular device switchable between three distinct mechanical states. *Applied Physics Letters*, **80**, 883–885.

53 Dittmer, W.U., Reuter, A. and Simmel, F.C. (2004) A DNA-based machine that can cyclically bind and release thrombin. *Angewandte Chemie-International Edition*, **43**, 3550–3553.

54 Kelemen, L., Valkai, S. and Ormos, P. (2006) Integrated optical motor. *Applied Optics*, **45**, 2777–2780.

55 Galajda, P. and Ormos, P. (2002) Rotors produced and driven in laser tweezers with reversed direction of rotation. *Applied Physics Letters*, **80**, 4653–4655.

56 Galajda, P. and Ormos, P. (2002) Rotation of microscopic propellers in laser tweezers. *Journal of Optics B: Quantum and Semiclassical Optics*, **4**, S78–S81.

57 Galajda, P. and Ormos, P. (2001) Complex micromachines produced and driven by light. *Applied Physics Letters*, **78**, 249–251.

58 Fennimore, A.M., *et al*. (2003) Rotational actuators based on carbon nanotubes. *Nature*, **424**, 408–410.

59 Schliwa M. (ed.) (2002) *Molecular Motors*, Wiley-VCH, Weinheim.

60 Muthukrishnan, G., *et al*. (2006) Transport of semiconductor nanocrystals by kinesin molecular motors. *Small*, **2**, 626–630.

61 Stock, D., Leslie, A.G.W. and Walker, J.E. (1999) Molecular architecture of the rotary motor in ATP synthase. *Science*, **286**, 1700–1705.

62 Sambongi, Y., *et al*. (1999) Mechanical rotation of the c subunit oligomer in ATP synthase (F0F1): direct observation. *Science*, **286**, 1722–1724.

63 Fletcher, S.P., *et al*. (2005) A reversible, unidirectional molecular rotary motor driven by chemical energy. *Science*, **310**, 80–82.

64 Hugel, T., *et al*. (2002) Single-molecule optomechanical cycle. *Science*, **296**, 1103–1106.

65 Hartley, G.S. (1937) The *cis*-form of azobenzene. *Nature*, **140**, 281–281.

66 Grandbois, M., *et al*. (1999) How strong is a covalent bond? *Science*, **283**, 1727–1730.

67 Eichen, Y., *et al*. (1998) Self-assembly of nanoelectronic components and circuits using biological templates. *ACTA Polymerica*, **49**, 663–670.

68 Klein, D.L., *et al*. (1997) A single-electron transistor made from a cadmium selenide nanocrystal. *Nature*, **389**, 699–701.

69 Braun, E., *et al*. (1998) DNA-templated assembly and electrode attachement of a conducting silver wire. *Nature*, **391**, 775–778.

70 Offenhäusser, A., Rühe, J. and Knoll, W. (1995) Neuronal cells cultured on modified microelectronic device surfaces. *Journal of Vacuum Science & Technology A-Vacuum Surfaces and, Films* **13**, 2606–2612.

71 Xia, Y.N. and Whitesides, G.M. (1998) Soft lithography. *Annual Review of Materials Science*, **28**, 153.

72 Quake, S.R. and Scherer, A. (2000) From micro- to nanofabrication with soft materials. *Science*, **290**, 1536–1540.

73 Unger, M.A., *et al*. (2000) Monolithic microfabricated valves and pumps by multilayer soft lithography. *Science*, **288**, 113–116.

74 Chen, J.H. and Seeman, N.C. (1991) Synthesis from DNA of a molecule with the connectivity of a cube. *Nature*, **350**, 631–633.

75 Zhang, Y.W. and Seeman, N.C. (1994) Construction of a DNA-truncated octahedron. *Journal of the American Chemical Society*, **116**, 1661–1669.

76 Mao, C., Sun, W. and Seeman, N.C. 1997 Assembly of borromean rings from DNA. *Nature* **386** 137–138.

77 Goodman, R.P., *et al*. (2005) Rapid chiral assembly of rigid DNA building blocks for

molecular nanofabrication. *Science*, **310**, 1661–1665.

78 Shih, W.M., Quispe, J.D. and Joyce, G.F. (2004) A 1.7-kilobase single-stranded DNA that folds into a nanoscale octahedron. *Nature*, **427**, 618–621.

79 LaBean, T.H., *et al.* (2000) Construction, analysis, ligation and self-assembly of DNA triple crossover complexes. *Journal of the American Chemical Society*, **122**, 1848–1860.

80 Liu, D., *et al.* (2004) Tensegrity: construction of rigid DNA triangles with flexible four-arm DNA junctions. *Journal of the American Chemical Society*, **126**, 2324–2325.

81 Liu, H.P., *et al.* (2006) Approaching the limit: can one DNA oligonucleotide assemble into large nanostructures? *Angewandte Chemie-International Edition*, **45**. 1942,

82 Rothemund, P.W.K. (2006) Folding DNA to create nanoscale shapes and patterns. *Nature*, **440**, 297–302.

83 Liu, D., *et al.* (2004) DNA nanotubes self-assembled from triple-crossover tiles as templates for conductive nanowires. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 717–722.

84 Mitchell, J.C., *et al.* (2004) Self-assembly of chiral DNA nanotubes. *Journal of the American Chemical Society*, **126**, 16342–16343.

85 Rothemund, P.W.K., *et al.* (2004) Design and characterization of programmable DNA nanotubes. *Journal of the American Chemical Society*, **126**, 16344–16352.

86 Le, J.D., *et al.* (2004) DNA-templated self-assembly of metallic nanocomponent arrays on a surface. *Nano Letters*, **4**, 2343–2347.

87 Pinto, Y.Y., *et al.* (2005) Sequence-encoded self-assembly of multiple-nanocomponent arrays by 2D DNA scaffolding. *Nano Letters*, **5**, 2399–2402.

88 Malo, J., *et al.* (2005) Engineering a 2D protein-DNA crystal. *Angewandte Chemie-International Edition*, **44**, 3057–3061.

89 Liu, Y., *et al.* (2005) Protein nanoarrays – aptamer-directed self-assembly of protein arrays on a DNA nanostructure. *Angewandte Chemie-International Edition*, **44**, 4333–4338.

90 Parak, W.J. *et al.* (2004) Quantum dots, in *Nanoparticles – From Theory to Application*, (ed. G. Schmid), Wiley-VCH, Weinheim, pp. 4–49.

91 Davies, J.H. (1998) *The Physics of Low-dimensional Semiconductors*, Cambridge University Press, Cambridge.

92 Ando, T., Fowler, A.B. and Stern, F. (1982) Electronic properties of two-dimensional systems. *Reviews of Modern Physics*, **54**, 437–672.

93 Fuhrer, A., *et al.* (2001) Energy spectra of quantum rings. *Nature*, **413**, 822–825.

94 Blick, R.H., *et al.* (1998) Complex broadband millimeter wave response of a double quantum dot: Rabi oscillations in an artificial molecule. *Physical Review Letters*, **81**, 689–692.

95 Kemerink, M. and Molenkamp, L.W. (1994) Stochastic Coulomb blockade in a double quantum dot. *Applied Physics Letters*, **65**, 1012–1014.

96 Waugh, F.R., *et al.* (1995) Single-electron charging in double and triple quantum dots with tunable coupling. *Physical Review Letters*, **75**, 705–708.

97 Hofmann, F. and Wharam, D.A. (1995) Investigation of the Coulomb blockade in a parallel quantum dot geometry. *Adv. Solid State Phys.*, **35**, 197–214 (e.g. at http://www.springerlink.com.contents/92863331-121165wl).

98 Blick, R.H., *et al.* (1996) Single-electron tunneling through a double quantum dot: the artificial molecule. *Physical Review B-Condensed Matter*, **53**, 7899–7902.

99 Bayer, M., *et al.* (1998) Optical modes in photonic molecules. *Physical Review Letters*, **81**, 2582–2585.

100 Tarucha, S., *et al.* (1996) Shell filling and spin effects in a few electron quantum dot. *Physical Review Letters*, **77**, 3613–3616.

101 Goldhaber-Gordon, D., *et al.* (1998) Kondo effect in a single-electron transistor. *Nature*, **391**, 156–159.

102 Cronenwett, S.M., Oosterkamp, T.H. and Kouwenhoven, L.P. (1998) *Science*, **281**, 540.

103 Simmel, F., *et al.* (1999) Anomalous Kondo effect in a quantum dot at nonzero bias. *Physical Review Letters*, **83**, 804–807.

104 Schmid, J., *et al.* (2000) Absence of odd–even parity behavior for Kondo resonances in quantum dots. *Physical Review Letters*, **84**, 5824–5827.

105 Petta, J.R., *et al.* (2005) Coherent manipulation of coupled electron spins in semiconductor quantum dots. *Science*, **309**, 2180–2184.

106 Koppens, F.H.L., *et al.* (2006) Driven coherent oscillations of a single electron spin in a quantum dot. *Nature*, **442**, 766–771.

107 Petroff, P.M., Lorke, A. and Imamoglu, A. (May, 2001) Epitaxially self-assembled quantum dots. *Physics Today, Physics Today*, **54**, *5*, 46–52.

108 Cho, A.Y. (1999) How molecular beam epitaxy (MBE) began and its projections into the future. *Journal of Crystal Growth*, **202**, 1–7.

109 Fafard, S., *et al.* (1994) 0-Dimensional-induced optical properties in self-assembled quantum dots. *Superlattices and Microstructures*, **16**, 303–309.

110 Petroff, P.M. and DenBaars, S.P. (1994) MBE and MOCVD growth and properties of self-assembling quantum dot arrays in III–V semiconductor structures. *Superlattices and Microstructures*, **15**, 15–21.

111 Leon, R., *et al.* (1995) Spatially resolved visible luminescence of self-assembled semiconductor. *Science*, **267**, 1966–1968.

112 Luyken, R.J., *et al.* (1999) The dynamics of tunneling into self-assembled InAs dots. *Applied Physics Letters*, **74**, 2486–2488.

113 Paul, D.J. (1999) Silicon–germanium strained layer materials in microelectronics. *Advanced Materials*, **11**, 191–204.

114 Garcia, J.M., *et al.* (1997) Intermixing and shape changes during the formation of InAs self-assembled quantum dots. *Applied Physics Letters*, **71**, 2014–2016.

115 Bayer, M., *et al.* (2000) Hidden symmetries in the energy levels of excitonic "artificial atoms". *Nature*, **405**, 923–926.

116 Warburton, R.J., *et al.* (2000) Optical emission from a charge-tunable quantum ring. *Nature*, **405**, 926–929.

117 Tarucha, S. (1998) Transport in quantum dots: observation of atomlike properties. *MRS Bulletin*, **23**, 49–53.

118 Lorke, A. and Luyken, R.J. (1998) Many-particle ground states and excitations in nanometer-size quantum structures. *Journal of Physics B: Condensed Matter*, **256–258**, 424–430.

119 Gammon, D. (2000) Semiconductor physics: electrons in artificial atoms. *Nature*, **405**, 899–900.

120 Alivisatos, A.P. (1996) Semiconductor clusters, nanocrystals and quantum dots. *Science*, **271**, 933–937.

121 Murray, C.B., Norris, D.J. and Bawendi, M.G. (1993) Synthesis and characterization of nearly monodisperse CdE (E = S, Se, Te) semiconductor nanocrystallites. *Journal of the American Chemical Society*, **115**, 8706–8715.

122 Peng, X., Wickham, J. and Alivisatos, A.P. (1998) Kinetics of II–VI and III–V colloidal semiconductor nanocrystal growth: "focusing" of size distributions. *Journal of the American Chemical Society*, **120**, 5343–5344.

123 Dabbousi, B.O., *et al.* (1997) (CdSe)ZnS core–shell quantum dots: synthesis and characterization of a size series of highly luminescent nanocrystallites. *Journal of Physical Chemistry B*, **101**, 9463–9475.

124 Peng, X., *et al.* (1997) Epitaxial growth of highly luminescent CdSe/CdS core/shell nanocrystals with photostability and electronic accessibility. *Journal of the American Chemical Society*, **119**, 7019–7029.

125 Peng, X., *et al.* (2000) Shape control of CdSe nanocrystals. *Nature*, **404**, 59–61.

126 Puntes, V.F., Krishnan, K. and Alivisatos, A.P. (2002) Synthesis of colloidal cobalt nanoparticles with controlled size and shapes. *Topics in Catalysis*, **19**, 145–148.

127 Kudera, S., *et al.* (2006) Synthesis and perspectives of complex crystaline nanostructures. *Phys. Status Solidi C*, **203**, 1329–1336.

128 Alivisatos, A.P.(August, 1995) Semiconductor nanocrystals. *MRS Bulletin*, 23–32.

129 Alivisatos, A.P. (1996) Perspectives on the physical chemistry of semiconductor nanocrystals. *Journal of Physical Chemistry A*, **100**, 13226–13239.

130 Wu, M.X., *et al.* (2003) Immunofluorescent labeling of cancer marker Her2 and other cellular targets with semiconductor quantum dots – corrigenda. *Nature Biotechnology*, **21**, 452.

131 Bruchez, M.P. (2005) Turning all the lights on: quantum dots in cellular assays. *Current Opinion in Chemical Biology*, **9**, 533–537.

132 Loweth, C.J., *et al.* (1999) DNA-based assembly of gold nanocrystals. *Angewandte Chemie-International Edition*, **38**, 1808–1812.

133 Bruchez, M.J., *et al.* (1998) Semiconductor nanocrystals as fluorescent biological labels. *Science*, **281**, 2013–2016.

134 Chan, W.C.W. and Nie, S. (1998) Quantum dot bioconjugates for ultrasensitive nonisotopic detection. *Science*, **281**, 2016–2018.

135 Dubertret, B., *et al.* (2002) *In vivo* imaging of quantum dots encapsulated in phospholipid micelles. *Science*, **298**, 1759–1762.

136 Klein, D.L., *et al.* (1996) An approach to electrical studies of single nanocrystals. *Applied Physics Letters*, **68**, 2574–2576.

137 Moore, G.E. (1965) Cramming more components onto integrated circuits. *Electronics*, **38**, 8.

4
Fundamental Principles of Quantum Dots[1)]

Wolfgang J. Parak, Liberato Manna, and Thomas Nann

4.1
Introduction and Outline

4.1.1
Nanoscale Science and Technology

In the last decade new directions of modern research, broadly defined as "nanoscale science and technology" have emerged [2, 3]. These new trends involve the ability to fabricate, characterize and manipulate artificial structures, whose features are controlled at the lower nanometer scale. They embrace areas of research as diverse as engineering, physics, chemistry, materials science and molecular biology. Research in this direction has been triggered by the recent availability of revolutionary instruments and approaches that allow the investigation of material properties with a resolution close to the atomic level. Strongly connected to such technological advances are pioneering studies that have revealed new physical properties of matter at a level which is intermediate between the atomic and molecular level and bulk.

Materials science and technology is a rapidly evolving field and is currently making the most significant contributions to research in nanoscale science. It is driven by the desire to fabricate materials with novel or improved properties. Such properties can be, for instance, strength, electrical and thermal conductivity, optical response, elasticity and wear resistance. Research is also evolving towards materials that are designed to perform more complex and efficient tasks. Examples include materials with a higher rate of decomposition of pollutants, a selective and sensitive response towards a given biomolecule, an improved conversion of light into current and more efficient energy storage. For such and more complex tasks to be realized, novel materials have to be based on several components whose spatial organization is engineered at the molecular level. This class of materials can be defined as

1) This chapter has been partly adapted from a previous version which included contributions also from Dr. Daniele Gerion, Dr. Friedrich Simmel and Professor Dr. Paul Alivisatos [1].

Nanotechnology. Volume 1: Principles and Fundamentals. Edited by Günter Schmid

ISBN: 978-3-527-31732-5

"nano-composites". They are made of assembled nanosized objects or molecules. Their macroscopic behavior arises from the combination of the novel properties of the individual building blocks and their mutual interaction.

In electronics, the design and assembly of functional materials and devices based on nanoscale building blocks can be seen as the natural, inevitable evolution of the trend towards miniaturization. The microelectronics industry, for instance, is fabricating integrated circuits and storage media whose basic units are approaching the size of few tens of nanometers. For computers, "smaller" goes along with higher computational power at lower cost and with greater portability. However, this race towards higher performance is driving current silicon-based electronics to the limits of its capability [4–7]. The design of each new generation of smaller and faster devices involves more sophisticated and expensive processing steps and requires the solution of new sets of problems, such as heat dissipation and device failure. If the trend towards further miniaturization persists, silicon technology will soon reach limits at which these problems become insurmountable. In addition, scientists have found that device characteristics in very small components are strongly altered by quantum mechanical effects. In many cases, these effects will undermine the classical principles on which most of today's electronic components are based. For these reasons, alternative materials and approaches are currently being explored for novel electronic components, in which the laws of quantum mechanics regulate their functioning in a predictable way. Perhaps in the near future a new generation of computers will rely on fundamental processing units that are made of only a few atoms.

Fortunately, the advent of new methods for the controlled production of nanoscale materials has provided new tools that can be adapted for this purpose. New terms such as nanotubes, nanowires and quantum dots are now common jargon in scientific publications. These objects are among the smallest, man-made units that display physical and chemical properties which make them promising candidates as fundamental building blocks of novel transistors. The advantages envisaged here are higher device versatility, faster switching speed, lower power dissipation and the possibility of packing many more transistors on a single chip. Prototypes of these new single nanotransistors are nowadays fabricated and studied in research laboratories and are still far from commercialization. How millions of such components could be arranged and interconnected in complex architectures and at low cost still remains a formidable task to solve.

With a completely different objective, the pharmaceutical and biomedical industries try to synthesize large supramolecular assemblies and artificial devices that mimic the complex mechanisms of nature or that can potentially be used for more efficient diagnoses and better cures for diseases. Examples in this direction are nanocapsules such as liposomes, embodying drugs that can be selectively released in living organs or bioconjugate assemblies of biomolecules and magnetic (or fluorescent) nanoparticles that might provide faster and more selective analysis of biotissues in addition to less invasive cures for several types of diseases. These prototype systems might one day evolve into more complex nanomachines, with highly sophisticated functional features, able to carry out complicated tasks at the cellular level into a living body.

Nanoscience and nanotechnology will definitely have a strong impact on many aspects of future society. The scientific community envisages that nanotechnology will strongly permeate key areas such as information technology and telecommunications, medicine, energy production and storage and transportation. This chapter, however, is not meant as a survey of the present state and future developments of nanoscale science and technology and the list of examples mentioned above is far from complete. Here, we simply want to stress the point that any development in nanoscience must necessarily follow the understanding of the physical laws that govern matter at the nanoscale and how the interplay of the various physical properties of a nanoscopic system translates into a novel behavior or into a new physical property. In this sense, this chapter will serve as an overview of basic physical rules governing nanoscale materials, with particular emphasis on quantum dots, including their various physical realizations and their possible applications. Quantum dots are the ultimate example of a solid in which all dimensions shrink down to a few nanometers. Moreover, semiconductor quantum dots are probably the most studied nanoscale systems.

The outline of this chapter is as follows. In Section 4.2 we will try to explain with a few examples why the behavior of nanoscale materials can be very different from that of their bulk and from their atomic counterparts and how quantum mechanics can help us in rationalizing this. Following this discussion, we will give a definition of "quantum dot". In Section 4.3, we follow a bottom-up approach and give a simplified picture of a solid as being a very large molecule, where the energy levels of each individual atomic component have merged to form bands. The electronic structure of a quantum dot, being intermediate between the two extreme cases of single atoms and the bulk, will then be an easier concept to grasp. In Section 4.4, we will use the model of a free electron gas and the concept of quantum confinement to explain what happens to a solid when its dimensions shrink one by one. This will lead us to a more accurate definition of quantum wells, quantum wires and quantum dots. In Section 4.5, we will examine in more detail the electronic structure of quantum dots, although we will try to keep the level of the discussion relatively simple.

4.2
Nanoscale Materials and Quantum Mechanics

4.2.1
Nanoscale Materials are Intermediates Between Atomic and Bulk Matter

Nanoscale materials frequently show a behavior which is intermediate between that of a macroscopic solid and that of an atomic or molecular system. Consider, for instance, the case of an inorganic crystal composed of very few atoms. Its properties will be different from those of a single atom, but we cannot imagine that they will be the same as those of a bulk solid. The number of atoms on its surface, for instance, is a significant fraction of the total number of atoms and therefore will have a large influence on the overall properties of the crystal. We can easily imagine that this

crystal might have higher chemical reactivity than the corresponding bulk solid and that it will probably melt at lower temperatures. Consider now the example of a carbon nanotube, which can be thought of as a sheet of graphite wrapped in such a way that the carbon atoms on one edge of the sheet are covalently bound to the atoms on the opposite edge of the sheet. Unlike its individual components, a carbon nanotube is chemically extremely stable because the valences of all its carbon atoms are saturated. Moreover, we guess that carbon nanotubes can be good conductors because electrons can move freely along these tiny, wire-like structures. Once again, we see that such nanoscopic objects can have properties which do not belong to the realm of their larger (bulk) or smaller (atoms) counterparts. However, there are many additional properties specific to such systems which cannot be understood by such a simple reasoning. These properties are related to the sometimes counterintuitive behavior that charge carriers (electrons and holes) can exhibit when they are forced to dwell in such structures. These properties can only be explained by the laws of quantum mechanics.

4.2.2 Quantum Mechanics

A fundamental aspect of quantum mechanics is the particle–wave duality, introduced by de Broglie, according to which any particle can be associated with a matter wave whose wavelength is inversely proportional to the particle's linear momentum. Whenever the size of a physical system becomes comparable to the wavelength of the particles that interact with such a system, the behavior of the particles is best described by the rules of quantum mechanics [8]. All the information we need about the particle is obtained by solving its Schrödinger equation. The solutions of this equation represent the possible physical states in which the system can be found. Fortunately, quantum mechanics is not required to describe the movement of objects in the macroscopic world. The wavelength associated with a macroscopic object is in fact much smaller than the object's size and therefore the trajectory of such an object can be derived with the principles of classical mechanics. Things change, for instance, in the case of electrons orbiting around a nucleus, since their associated wavelength is of the same order of magnitude as the electron–nucleus distance.

We can use the concept of particle–wave duality to give a simple explanation of the behavior of carriers in a semiconductor nanocrystal. In a bulk inorganic semiconductor, conduction band electrons (and valence band holes) are free to move throughout the crystal and their motion can be described satisfactorily by a linear combination of plane waves whose wavelength is generally of the order of nanometers. This means that, whenever the size of a semiconductor solid becomes comparable to these wavelengths, a free carrier confined in this structure will behave as a particle in a potential box [9]. The solutions of the Schrödinger equation are standing waves confined in the potential well and the energies associated with two distinct wavefunctions are, in general, different and discontinuous. This means that the particle energies cannot take on any arbitrary value and the system exhibits a

discrete energy level spectrum. Transitions between any two levels are seen, for instance, as discrete peaks in the optical spectra. The system is then also referred to as "quantum confined". If all the dimensions of a semiconductor crystal shrink down to a few nanometers, the resulting system is called a "quantum dot" and will be the subject of our discussion throughout this chapter. The main point here is that in order to rationalize (or predict) the physical properties of nanoscale materials, such as their electrical and thermal conductivity or their absorption and emission spectra, we need first to determine their energy level structure.

For quantum confined systems such as quantum dots, the calculation of the energy structure is traditionally carried out using two alternative approaches. One approach was just outlined above. We take a bulk solid and we study the evolution of its band structure as its dimensions shrink down to few nanometers. This method will be described in more detail later (Section 4.4). Alternatively, we can start from the individual electronic states of single isolated atoms as shown in Section 4.3, and then study how the energy levels evolve as atoms come closer and start to interact with each other.

4.3
From Atoms to Molecules and Quantum Dots

The chemical approach towards quantum dots resembles the "bottom-up" strategy by which molecules are formed by chemical reactions between individual atoms or smaller molecules. Molecules are stable assemblies of an exactly defined finite number of atoms, whereas solids are stable assemblies of atoms or molecules with quasi-infinite dimensions. If the assembly of units within a solid does not follow translational symmetry, the solid is called amorphous. If, on the other hand, the units are repeated regularly, the solid is called crystalline. Since quantum dots have finite dimensions and are regular assemblies of atoms, such nano-objects are regarded as large molecules from a chemist's point of view, whereas physicists see them usually as small crystals.

The electronic properties of quantum dots can now be described and calculated by linear combinations of atomic orbitals (LCAO method or Hückel theory) and other approximations [10]. Here, the starting point is an atom, whereas the physical approach in Section 4.4 starts with an infinite wavefunction. It will be shown that the results of both approaches are basically the same.

The fundamental idea of quantum theory was developed at the beginning of the last century and affirms that particles have wave-like properties and vice versa. In about 1923, de Broglie suggested his famous momentum–wavelength relationship (4.3)[2)] by combining Einstein's relativistic energy (4.1) with the energy of a photon (4.2):

2) (1) = (2) $\Rightarrow mc^2 = h\nu$ (h = Plank's constant, c = speed of light).
Momentum of a photon $p = mc \Rightarrow pc = h\nu$.
Wavelength of a photon $\lambda = c/\nu \Rightarrow p = h/\lambda$
$\Rightarrow \lambda = h/p$.

$$E = mc^2 \tag{4.1}$$

$$E = h\nu \tag{4.2}$$

$$\lambda = h/p \tag{4.3}$$

The left-hand term (wavelength) in Equation 4.3 represents the wave nature of a particle, whereas the right-hand term (momentum) represents the particle-nature of a wave. Equation 4.3 can be written as[3)]

$$\lambda = h/p = h/\sqrt{2m(E-V)} \tag{4.4}$$

where m is the mass of the particle and E and V are its total and potential energy, respectively. Combining Equation 4.4 with the classical three-dimensional wave equation:

$$\nabla^2\psi(x,y,z) = -(2\pi/\lambda)^2\,\psi(x,y,z) \tag{4.5}$$

where Ψ is the wavefunction results in

$$\nabla^2\psi = -[2\pi\sqrt{2m(E-V)}/h]^2\,\psi = -8\pi^2 m/h^2(E-V)\psi \tag{4.6}$$

Some trivial rearrangements and insertion of $\hbar = h/2\pi$ result in Schrödinger's equation:

$$\nabla^2\psi + \frac{2m}{\hbar^2}(E-V)\psi = 0 \tag{4.7}$$

or

$$\hat{H}\cdot\psi = E\cdot\psi \tag{4.8}$$

using the Hamiltonian operator $\hat{H} = H\left(x,\ldots,\frac{\hbar}{i}\frac{\partial}{\partial x},\ldots\right)$; ψ is called eigenfunction of the operator and E is the eigenvalue, which represents the energy.

The first step to calculate electronic properties of matter is to apply Schrödinger's equation to the hydrogen atom. Therefore, we look at the wavefunction of the electron in the potential field of the nucleus, which is basically the coulomb attraction between electron and nucleus:

$$V = -\frac{e^2}{4\pi\varepsilon_0 r} \tag{4.9}$$

Furthermore, the wavefunction has to be separated into equations in terms of spherical coordinates:

$$\psi(r,\vartheta,\varphi) = R(r)\cdot\Theta(\vartheta)\cdot\Phi(\varphi) \tag{4.10}$$

The solution of Schrödinger's equation with the separated variables leads to three quantum numbers associated with the three functions and the hydrogen atom's

3) $E = p^2/2m + V$ (total energy E = kinetic energy $p^2/2m$ + potential energy V)
$\Rightarrow p^2 = 2m(E-V) \Rightarrow (3)\lambda = h[2m(E-V)]^{-1/2}$

energy levels. Furthermore, each one-electron wavefunction can exist in two forms called "spin states".[4] The function ψ^2 ($\psi\psi^*$, respectively) gives the probability density of finding an electron at a given point. The different eigenfunctions ψ for different sets of quantum numbers are called "atomic orbitals" (AOs) with corresponding energies (eigenvalues) E [10, 11].

Schrödinger's equation for multi-electron atoms and molecules becomes increasingly complicated, since all of the interactions between different electrons and nuclei contribute to the potential energy. In the Born–Oppenheimer approximation, the terms that describe movement of the nuclei are decoupled from those that describe the movement of the electrons, which move much faster compared with the nuclei. Thus the Hamiltonian for a molecule with N atoms and K electrons reads

$$\hat{H} = \left(-\frac{\hbar^2}{2m} \sum_{i=1,K} \nabla^2 - \sum_{j=1,N} \sum_{i=1,K} Z_j / r_{ij} \right) + \sum_{i=1,K-1} \sum_{l=i+1,K} 1/r_{il} + \sum_{j=1,N-1} \sum_{m=j+1,N} Z_j Z_m / r_{jm} \tag{4.11}$$

where Z_i are charges of the nuclei and r_{ij} distances between two charges. Schrödinger equations for molecules cannot be solved analytically. Therefore, approximation methods have to be used. These methods can be categorized as either "*ab initio*" or "semiempirical". *Ab initio* methods use only natural constants to solve Schrödinger's equation. The most prominent *ab initio* method is the Hartree–Fock method. Semiempirical methods use measured values (usually spectroscopic data) for the same purpose.

The numerical solution of the Hartree–Fock equation is only valid for atoms and linear molecules. Therefore, another approximation is needed: orbitals are written as the product of so-called basis functions (usually exponential functions) and the wavefunctions of hydrogen. These functions are called basis sets and vary for different approaches. *Ab initio* methods have high computing requirements, so that they are restricted to the calculation of some dozens of atoms.

In order to calculate electronic states of larger molecules, one has to introduce further simplifications. Semiempirical methods make use of experimental or fitted data to replace variables or functions in the Hartree–Fock matrices. The general procedure for calculating solutions for Schrödinger's equation remains the same as described above, but due to the simplifications, semiempirical methods allow for the calculation of molecules with several hundred atoms, including nanocrystals.

Figure 4.1 displays schematically the "transition" from atomic orbitals (s, p or sp^3) over molecular orbitals (σ, σ^*) to quantum dots and semiconductor energy bands. Since electrons populate the orbitals with lowest energy first, there is a highest occupied (molecular) or "binding" orbital (the "valence band" in semiconductors) and a lowest unoccupied (molecular) or "antibinding" orbital (the "conduction band"). The energy gap between the highest occupied (molecular) orbital (HOMO)

4) Readers who are interested in the details of this solution can find it in every common physical phemistry textbook [10, 11].

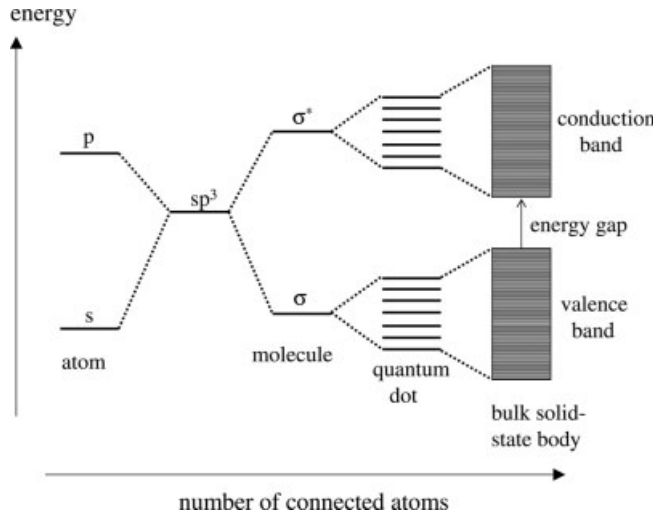


Figure 4.1 Electronic energy levels depending on the number of bound atoms. By binding more and more atoms together, the discrete energy levels of the atomic orbitals merge into energy bands (here shown for a semiconducting material) [12]. Therefore, semiconducting nanocrystals (quantum dots) can be regarded as a hybrid between small molecules and bulk material.

and lowest unoccupied (molecular) orbital (LUMO) is characteristic for luminophores, quantum dots and semiconductors. The energy gap decreases with increasing number of atoms in the transition region between molecules and bulk solids, as indicated in Figure 4.1. The exact calculation of this (most interesting) transition region follows the mathematical scheme described above. Alternatively, one can view the problem similar to a "particle-in-a-box" approach, which is outlined in Section 4.4.

First calculations of the energy of the first excited state in semiconductor quantum dots were carried out in the early 1980s by Brus [13, 14]. Brus did not solve Schrödinger's equation for the quantum dot, but for the exciton within a semiconductor nanocrystal by means of a variational method [effective-mass approximation (EMA)]. This approach thus resembles the particle-in-a-box method (see Section 4.4). The first semiempirical calculation was published in 1989 by Lippens and Lannoo [15]. They used the tight-binding approach to model CdS and ZnS quantum dots. As depicted in Figure 4.1 and calculated firstly by Brus, they found an increasing energy gap between HOMO and LUMO with decreasing nanocrystal size. Moreover, their results fit much better with experimental data than those obtained with the effective mass approximation (EMA).

Further refinements include the linear combination of atomic orbitals (LCAO) [16], the semiempirical pseudopotential calculation [17] and the *kp* method [18]. All of these methods provide estimates for the size-dependent bandgap of quantum dots. Even though the agreement of the calculations with the experimental data differs slightly, the general result is clear: the bandgap of the quantum dots increases with decreasing size of the nanocrystals. These results were as expected and thus not very

useful for the experimental scientist so far. However, they paved the way for more sophisticated calculations including the inclusion of defects in nanocrystals and the presence of surface ligands. Very few examples of such calculations have been published so far. One example is the calculation of the size-dependent behavior of the quantum dot–ligand bond [19].

4.4
Shrinking Bulk Material to a Quantum Dot

In this section, we will go back to the concept of quantum confinement of carriers in a solid from a physicist's point of view and will use it to derive a more detailed description of the electronic band structure in a low-dimensional solid. This description will catch the general physics of a solid when its dimensions shrink one by one down to few nanometers. We will start first with an elementary model of the behavior of electrons in a bulk solid. This model will then be adapted to the case of confined carriers.

4.4.1
Three-Dimensional Systems (Bulk Material)

We now consider the case of a three-dimensional solid with size d_x, d_y, d_z containing N free electrons. "Free" means that those electrons are delocalized and thus not bound to individual atoms. Furthermore, we will make the assumption that the interactions between the electrons, and also between the electrons and the crystal potential, can be neglected as a first approximation. Such a model system is called "free electron gas" [20, 21]. Astonishingly, this oversimplified model still captures many of the physical aspects of real systems. From more complicated theories, it has been learnt that many of the expressions and conclusions from the free electron model remain valid as a first approximation even when one takes electron–crystal and electron–electron interactions into account. In many cases it is sufficient to replace the free electron mass m by an "effective" mass m^* which implicitly contains the corrections for the interactions. To keep the story simple, we proceed with the free electron picture. In the free electron model, each electron in the solid moves with a velocity $\vec{v} = (v_x, v_y, v_z)$. The energy of an individual electron is then just its kinetic energy.[5)]

$$E = \frac{1}{2} m \vec{v}^2 = \frac{1}{2} m (v_x^2 + v_y^2 + v_z^2) \tag{4.12}$$

According to Pauli's exclusion principle, each electron must be in a unique quantum state. Since electrons can have two spin orientations ($m_s = +{}^1/_2$ and

5) The total energy E is the sum of the kinetic energy $p^2/2m$ and the potential energy V. For free particles (free electron gas) there is no potential energy and therefore their total energy is equal to their kinetic energy.

$m_s = -1/2$), only two electrons with opposite spins can have the same velocity $\vec{v}$. This case is analogous to the Bohr model of atoms, in which each orbital can be occupied by two electrons at maximum. In solid-state physics, the wavevector $\vec{k} = (k_x, k_y, k_z)$ of a particle is more frequently used instead of its velocity to describe the particle's state. Its absolute value $k = |\vec{k}|$ is the wavenumber. The wavevector $\vec{k}$ is directly proportional to the linear momentum $\vec{p}$ and thus also to the velocity $\vec{v}$ of the electron:

$$\vec{p} = m\vec{v} = \frac{h}{2\pi}\vec{k} \tag{4.13}$$

The scaling constant is Planck's constant h and the wavenumber is related to the wavelength λ associated with the electron through the de Broglie relation [20, 21] (Figure 4.2):

$$\pm k = |\vec{k}| = \pm\frac{2\pi}{\lambda} \tag{4.14}$$

The calculation of the energy states for a bulk crystal is based on the assumption of periodic boundary conditions (Figure 4.2). Periodic boundary conditions are a mathematical trick to "simulate" an infinite ($d \to \infty$) solid. This assumption

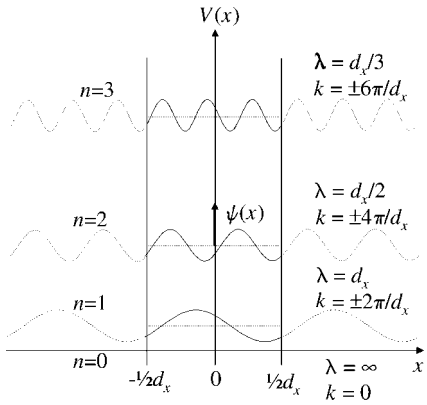


Figure 4.2 Periodic boundary conditions (only drawn for the x-dimension) for a free electron gas in a solid with thickness d. The idea of periodic boundary conditions is to "simulate" mathematically an infinite solid. Infinite extension is similar to an object without any borders. This means that a particle close to the "border" must not be affected by the border, but "behaves" exactly as in the bulk. This can be realized by using a wavefunction $\psi(x)$ that is periodic within the thickness d of the solid. Any electron that leaves the solid from its right boundary would re-enter under exactly the same conditions on its left side. For the electron the borders are quasi-nonexistent. The probability density $|\psi(x)|^2$ is the probability that an electron is at the position x in the solid. Different states for the electrons ($n = 0, 1, 2, \ldots$) have different wavefunctions. λ is the de Broglie wavelength of the electrons and k is their corresponding wavenumber. A "real" bulk solid can be approximated by an infinite solid ($d \to \infty$) and its electronic states in k-space are quasi-continuously distributed: $\Delta k = 2\pi/d_x \to 0$.

implies that the conditions at opposite borders of the solid are identical. In this way, an electron that is close to the border does not really "feel" the border. In other words, the electrons at the borders "behave" exactly as if they were in the bulk. This condition can be realized mathematically by imposing the following condition on the electron wavefunctions: $\psi(x, y, z) = \psi(x + d_x, y, z)$, $\psi(x, y, z) = \psi(x, y + d_y, z)$ and $\psi(x, y, z) = \psi(x, y, z + d_z)$. In other words, the wavefunctions must be periodic with a period equal to the whole extension of the solid number [21, 22]. The solution of the stationary Schrödinger equation under such boundary conditions can be factorized into the product of three independent functions $\psi(x, y, z) = \psi(x)\psi(y)\psi(z) = A\exp(ik_x x)\exp(ik_y y)\exp(ik_z z)$. Each function describes a free electron moving along one Cartesian coordinate. In the argument of the functions $k_{x,y,z}$ is equal to $\pm n\Delta k = \pm n2\pi/d_{x,y,z}$ and n is an integer [20–22]. These solutions are waves that propagate along the negative and positive directions for $k_{x,y,z} > 0$ and $k_{x,y,z} < 0$, respectively. An important consequence of the periodic boundary conditions is that all the possible electronic states in the $\vec{k}$ space are equally distributed. There is an easy way of visualizing this distribution in the ideal case of a one-dimensional free electron gas: there are two electrons ($m_s = \pm^1/_2$) in the state $k_x = 0$ ($v_x = 0$), two electrons in the state $k_x = +\Delta k$ ($v_x = +\Delta v$), two electrons in the state $k_x = -\Delta k$ ($v_x = -\Delta v$), two electrons in the state $k_x = +2\Delta k$ ($v_x = +2\Delta v$) and so on.

For a three-dimensional bulk material we can follow an analogous scheme. Two electrons ($m_s = \pm^1/_2$) can occupy each of the states $(k_x, k_y, k_z) = (\pm n_x\Delta k, \pm n_y\Delta k, \pm n_z\Delta k)$, again with $n_{x,y,z}$ being an integer. A sketch of this distribution is shown in Figure 4.3. We can easily visualize the occupied states in $\vec{k}$-space because all these states are included in a sphere whose radius is the wavenumber associated with the highest energy electrons. At the ground state, at 0 K, the radius of the sphere is the Fermi wavenumber k_F (Fermi velocity v_F). The Fermi energy $E_F \propto k_F^2$ is the energy of the last occupied electronic state. All electronic states with an energy $E \leq E_F$ are occupied, whereas all electronic states with higher energy $E > E_F$ are empty. In a solid, the allowed wavenumbers are separated by $\Delta k = \pm n2\pi/d_{x,y,z}$. In a bulk material $d_{x,y,z}$ is large and so Δk is very small. Then the sphere of states is filled quasi-continuously [21].

We need now to introduce the useful concept of the density of states $D_{3d}(k)$, which is the number of states per unit interval of wavenumbers. From this definition $D_{3d}(k)\Delta k$ is the number of electrons in the solid with a wavenumber between k and $k + \Delta k$. If we know the density of states in a solid we can calculate, for instance, the total number of electrons having wavenumbers less than a given k_{max}, which we will call $N(k_{max})$. Obviously, $N(k_{max})$ is equal to $\int_0^{k_{max}} D_{3d}(k)\mathrm{d}k$. In the ground state of the solid all electrons have wavenumbers $k \leq k_F$, where k_F is the Fermi wavenumber. Since in a bulk solid the states are homogeneously distributed in $\vec{k}$-space, we know that the number of states between k and $k + \Delta k$ is proportional to $k^2\Delta k$ (Figure 4.3). This can be visualized in the following way. The volume in three-dimensional $\vec{k}$-space scales with k^3. If we only want to count the number of states with a wavenumber between k and $k + \Delta k$, we need to determine the volume of a spherical shell with radius k and thickness Δk. This volume is proportional to product of the surface of the sphere (which scales as k^2)

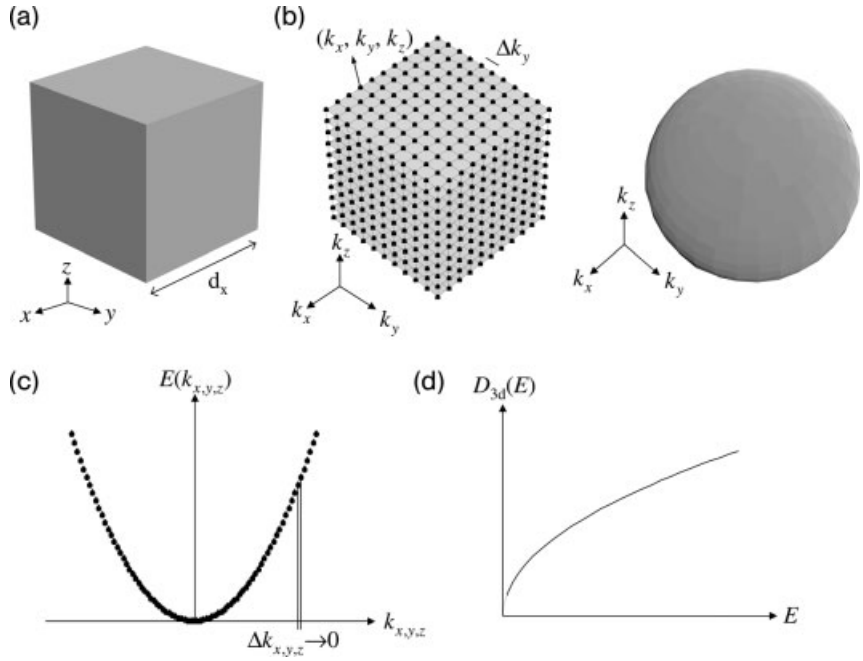


Figure 4.3 Electrons in a three-dimensional bulk solid [21]. (a) Such solid can be modeled as an infinite crystal along all three dimensions *x*, *y*, *z*. (b) The assumption of periodic boundary conditions yields standing waves as solutions for the Schrödinger equation for free electrons. The associated wavenumbers (k_x, k_y, k_z) are periodically distributed in the reciprocal *k*-space number [22]. Each of the dots shown in the figure represents a possible electronic state (k_x, k_y, k_z). Each state in *k*-space can be only occupied by two electrons. In a large solid the spacing $\Delta k_{x,y,z}$ between individual electron states is very small and therefore the *k*-space is quasi-continuously filled with states. A sphere with radius k_F includes all states with $k = (k_x^2 + k_y^2 + k_z^2)^{1/2} < k_F$. In the ground state, at 0 K, all states with $k < k_F$ are occupied with two electrons and the other states are empty. Since the *k*-space is homogeneously filled with states, the number of states within a certain volume scales with k^3. (c) Dispersion relation for free electrons in a three-dimensional solid. The energy of free electrons scales with the square of the wavenumber and its dependence on *k* is described by a parabola. For a bulk solid the allowed states are quasi-continuously distributed and the distance between two adjacent states (here shown as points) in *k*-space is very small. (d) Density of states D_{3d} for free electrons in a three-dimensional system. The allowed energies are quasi-continuous and their density scales with the square root of the energy $E^{1/2}$.

with the thickness of the shell (which is Δk). $D_{3d}(k)\Delta k$ is thus proportional to $k^2\Delta k$ and, in the limit when Δk approaches zero, we can write

$$D_{3d}(k) = \frac{dN(k)}{dk} \propto k^2 \tag{4.15}$$

Instead of knowing the density of states in a given interval of wavenumbers, it is more useful to know the number of electrons that have energies between E and $E + \Delta E$. From Equations 4.12 and 4.13 we know that $E(k)$ is proportional to k^2 and thus $k \propto \sqrt{E}$. Consequently, $dk/dE \propto 1/\sqrt{E}$. By using Equation 4.15, we obtain for

the density of states for a three-dimensional electron gas [22]

$$D_{3d}(E) = \frac{dN(E)}{dE} = \frac{dN(k)}{dk}\frac{dk}{dE} \propto E \times 1/\sqrt{E} \propto \sqrt{E} \tag{4.16}$$

This can be seen schematically in Figure 4.3. With Equation 4.16 we conclude our simple description of a bulk material. The possible states in which an electron can be found are quasi-continuous. The density of states scales with the square root of the energy. More details about the free electron gas model and more refined descriptions of electrons in solids can be found in any solid-state physics textbook [20].

4.4.2
Two-Dimensional Systems

We now consider a solid that is fully extended along the *x*-and *y*-directions, but whose thickness along the *z*-direction (d_z) is only a few nanometers (Figure 4.5). Free electrons can still move freely in the *x*–*y* plane. However, movement in the *z*-direction is now restricted. Such a system is called a two-dimensional electron gas (2DEG) [23]. As mentioned in Section 4.2, when one or more dimensions of a solid become smaller than the de Broglie wavelength associated with the free charge carriers, an additional contribution of energy is required to confine the component of the motion of the carriers along this dimension. In addition, the movement of electrons along such a direction becomes quantized. This situation is shown in Figure 4.4. No electron can leave the solid and electrons that move in the *z*-direction are trapped in a "box". Mathematically this is described by infinitely high potential wells at the border $z = \pm^1/2d_z$.

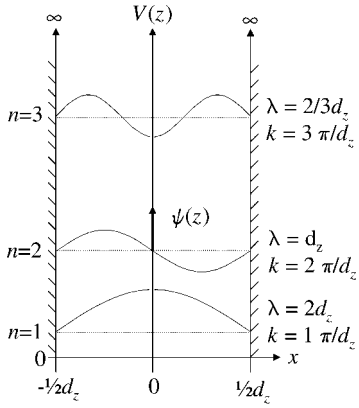


Figure 4.4 Particle in a box model for a free electron moving along in the *z*-axis. The movement of electrons in the *z*-direction is limited to a "box" with thickness *d*. Since electrons cannot "leave" the solid (the box), their potential energy *V*(*x*) is zero within the solid, but is infinite at its borders. The probability density $|\psi(z)|^2$ is the probability that an electron is located at position *x* in the solid. Different states for the electrons ($n = 1, 2, \ldots$) differ in their wavefunction.

The solutions for the particle in a box situation can be obtained by solving the one-dimensional Schrödinger equation for an electron in a potential $V(z)$, which is zero within the box but infinite at the borders. As can be seen in Figure 4.4, the solutions are stationarywaves with energies[6] $E_{nz} = \nabla^2 k_z^2/2m = h^2 k_z^2/8\pi^2 m = h^2 n_z^2/8md_z^2$, $n_z = 1, 2, \ldots$ [10, 22]. This is similar to states $k_z = n_z \Delta k_z$ with $\Delta k_z = \pi/d_z$. Again, each of these states can be occupied at the maximum by two electrons.

Let us compare the states in the k-space for three- and two-dimensional materials (Figures 4.3 and 4.5). For a two-dimensional solid that is extended in the x–y plane there are only discrete values allowed for k_z. The thinner the solid in the z-directions, the larger is the spacing Δk_z between those allowed states. On the other hand, the distribution of states in the $k_x - k_y$ plane remains quasi-continuous. Therefore, one can describe the possible states in the k-space as planes parallel to the k_x- and k_y-axes, with a separation Δk_z between the planes in the k_z-direction. We can number the individual planes with n_z. Since within one plane the number of states is quasi-continuous, the number of states is proportional to the area of the plane. This means that the number of states is proportional to $k^2 = k_x^2 + k_y^2$. The number of states in a ring with radius k and thickness Δk is therefore proportional to $k\Delta k$. Integration over all rings yields the total area of the plane in k-space. Here,

6) The particle-in-a-box approach (Figure 4.4) looks similar to the case of the periodic boundary conditions (Figure 4.2). There are indeed important differences between the two cases. Periodic boundary conditions "emulate" an infinite solid. A quantum mechanical treatment of this problem yields propagating waves that are periodic within the solid. Such waves can be seen as superposition of plane waves. For an idealized one-dimensional solid, with boundaries fixed at $x = \pm d/2$, a combination of plane waves can be for instance $\psi(x) = A\exp(ikx) + B\exp(-ikx)$ with $k = n \times 2\pi/d$. Written in another way, the solutions are of the type $\exp(ikx)$, with $k = \pm n \times 2\pi/d$. The solutions for $k = +n \times 2\pi/d$ and $k = -n \times 2\pi/d$ are linearly independent. The waves $\exp(+in \times 2\pi x/d)$ propagate to the right, the waves $\exp(-in \times 2\pi x/d)$ to the left side of the solid. Neither wave feels the boundaries. Since $\exp(ikx) = \cos(kx) + i\sin(kx)$ and $\exp(-ikx) = \cos(kx) - i\sin(kx)$, we also can write $\psi(x) = C\sin(kx) + D\cos(kx)$ with $k = n \times 2\pi/d$ as solutions. The only constraint here is that the wavefunction must be periodic throughout the solid. The state with wavenumber $k = 0$ is a solution, since $C\sin(0) + D\cos(0) = D \neq 0$. Therefore, the state with the lowest kinetic energy is $E \propto k^2 = 0$ for $k = 0$. The individual states in k-space are very close to each other because $\Delta k = 2\pi/d$ tends to 0 when d increases. On the other hand, the particle in a box model describes the case in which the motion of the electrons is confined along one or more directions. Outside the box the probability of finding an electron is zero. For a one-dimensional problem the solutions are standing waves of the type $\psi(x) = A\sin(kx)$ with $k = n\pi/d$. There is only one solution of this type. The function $\psi(x) = B\sin(-kx)$ can be written as $\psi(x) = -B\sin(kx)$ and therefore is still of the type $\psi(x) = A\sin(kx)$. Because of the boundary conditions $\psi(x = \pm d/2) = 0$ there is no solution of the type $\psi(x) = B\cos(kx)$. Since the standing wave is confined into the box, there is only the solution $k = +n\pi/d > 0$. For a small box the energy states are far apart from each other in k-space and the distribution of states and energies is discrete. An important difference with respect to the extended solid is the occurrence of a finite zero-point energy . There is no solution for $k = 0$, since $\psi(0) = A\sin(0) = 0$. Therefore, the energy of the lowest possible state ($n = 1$) is equal to $E = h^2/8md^2$, that is $k = \pi/d$. This energy is called zero-point energy and is a purely quantum mechanical effect. It can be understood as the energy that is required to "confine" the electron inside the box. For a large box the zero-point energy tends to zero. However, for small boxes this energy becomes significant as it scales with the square of the reciprocal of the box size d^2.

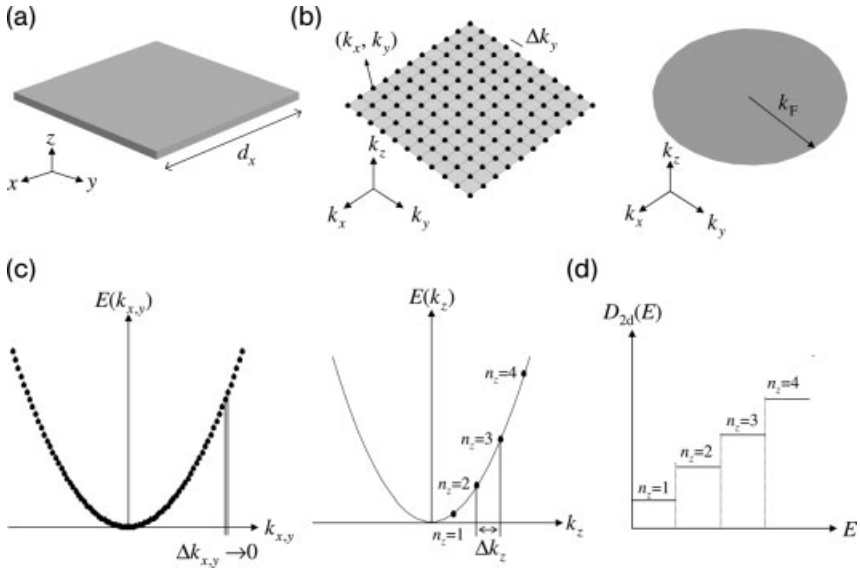


Figure 4.5 Electrons in a two-dimensional system. (a) A two-dimensional solid is (almost) infinitely extended in two dimensions (here x, y), but is very thin along the third dimension (here denoted z), which is comparable to the de Broglie wavelength of a free electron ($d_z \to \lambda$). (b) Electrons can still move freely along the x- and y-directions. The wavefunctions along such directions can be found again by assuming periodic boundary conditions. k_x and k_y states are quasi-continuously distributed in k-space. The movement of electrons in the z direction is restricted and electrons are confined to a "box". Only certain quantized states are allowed along this direction. For a discrete k_z state, the distribution of states in three-dimensional k-space can be described as a series of planes parallel to the k_x- and k_y-axes. For each discrete k_z state, there is a separate plane parallel to the k_x- and k_y-axes. Here only one of those planes is shown. The k_x and k_y states within one plane are quasi-continuous, since $\Delta k_{x,y} = 2\pi/d_{x,y} \to 0$. The distance between two planes for two separate k_z states is large, since $\Delta k_z = \pi/d_z \gg 0$. For each k_z value the k_x and k_y states are homogeneously distributed on the $k_x - k_y$ plane [22]. The number of states within this plane is therefore proportional to the area of a disk around $k_x = k_y = 0$. This means that the number of states for a certain wavenumber scales with k^2. In the ground state all states with $k \leq k_F$ are occupied with two electrons, while the remaining states are empty. (c) Free electrons have a parabolic dispersion relation [$E(k) \; (k^2]$. The energy levels $E(k_x)$ and $E(k_y)$ for the electron motion along the x- and y-directions are quasi-continuous (they are shown here as circles). The wavefunction $\psi(z)$ at the border of a small "box" must be zero, leading to standing waves inside the box. This constraint causes discrete energy levels $E(k_z)$ for the motion along the z-direction. Electrons can only occupy such discrete states (n_{z1}, n_{z2}, ..., shown here as circles). The position of the energy levels now changes with the thickness of the solid in the z-direction or in other words with the size of the "box". (d) Density of states for a two-dimensional electron gas. If electrons are confined in one direction (z) but can move freely in the other two directions (x, y), the density of states for a given k_z state ($n_z = 1, 2, \ldots$) does not depend on the energy E.

in contrast to the case of a three-dimensional solid, the density of states scales linearly with k:

$$D_{2d}(k) = \frac{dN(k)}{dk} \propto k \tag{4.17}$$

In the ground state, all states with $k \leq k_F$ are occupied with two electrons. We now want to know how many states exist for electrons that have energies between E and $E + \Delta E$. From Equations 4.12 and 4.13 we know the relation between k and E: $E(k) (k^2$ and thus $k \propto \sqrt{E}$ and $dk/dE \propto 1/\sqrt{E}$. By using Equation 4.17 we obtain the density of states for a two-dimensional electron gas; see also Figure 4.5 [22].

$$D_{2d}(E) = \frac{dN(E)}{dE} = \frac{dN(k)}{dk}\frac{dk}{dE} \propto \sqrt{E} \times 1/\sqrt{E} \propto 1 \tag{4.18}$$

The density of electronic states in a two-dimensional solid is therefore remarkably different from the three-dimensional case. The spacing between the allowed energy levels in the bands increases, because fewer levels are now present. As soon as one dimension is reduced to nanometer size, dramatic changes due to quantum confinement occur, as for example the non-negligible zero-point energy. In two-dimensional materials the energy spectrum is still quasi-continuous, but the density of states is now a step function [22, 24].

The quantum mechanical behavior of electrons in a two-dimensional solid is the origin of many important physical effects. With recent progress in nanoscience and -technology, the fabrication of two-dimensional structures has become routine. Two-dimensional systems are usually formed at interfaces between different materials or in layered systems in which some of the layers may be only a few nanometers thick. Structures such as this can be grown, for example, by successive deposition of the individual layers with molecular beam epitaxy. In such geometry, charge carriers (electrons and holes) can move freely parallel to the semiconductor layer, but their movement perpendicular to the interface is restricted. The study of these nanostructures led to the discovery of remarkable two-dimensional quantized effects, such as the integer and the fractional quantum Hall effect [25–28].

4.4.3
One-Dimensional Systems (Quantum Wires)

Let us now consider the case in which the solid also shrinks along a second (y) dimension. Now electrons can only move freely in the x-direction and their motion along the y- and z-axes is restricted by the borders of the solid (Figure 4.6). Such a system is called "quantum wire" and – when electrons are the charge carriers – a one-dimensional electron system (1DES). The charge carriers and excitations now can move only in one dimension and occupy quantized states in the other two dimensions.

The states of a one-dimensional solid can now be obtained by methods that are analogous to those described for the three- and two-dimensional materials. In the x-direction electrons can move freely and again we can apply the concept of periodic boundary conditions. This gives a quasi-continuous distribution of states parallel to the k_x-axis and for the corresponding energy levels. Electrons are confined along the remaining directions and their states can be derived from the Schrödinger equation for a "particle in a box" potential. Again, this yields discrete k_y and k_z states. We can now visualize all possible states as lines parallel to the k_x-axis. The lines are separated by discrete intervals along k_y and k_z, but within one line the distribution of k_x states is

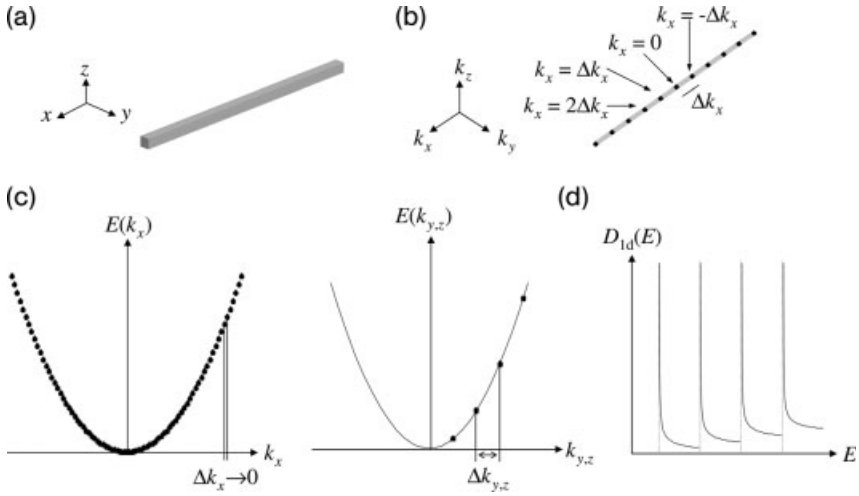


Figure 4.6 (a) One-dimensional solid. (b) The allowed (k_x, k_y, k_z) states can be visualized as lines parallel to the k_x-axes in the three-dimensional k-space. In this figure only one line is shows as an example. Within each line, the distribution of states is quasi-continuous, since $\Delta k_x \to 0$. The arrangement of the individual lines is discrete, since only certain discrete k_y and k_z states are allowed. (c) This can also be seen in the dispersion relations. Along the k_x-axes the energy band $E(k_x, k_y, k_z)$ is quasi-continuous, but along the k_y- and k_z-axes only certain energies exist. (d) The density of states within one line along the k_x-axes is proportional to $E^{-1/2}$. Each of the hyperbolas shown in the D_{1d} diagram corresponds to an individual (k_y, k_z) state.

quasi-continuous (Figure 4.6). We can count the number of states along one line by measuring the length of the line. The number of states is therefore proportional to $k = k_x$. Hence the number of states with wavenumbers in the interval between k and $k + \Delta k$ is proportional to Δk:

$$D_{1d}(k) = \frac{dN(k)}{dk} \propto 1 \tag{4.19}$$

In the ground state, all states with $k \leq k_F$ are occupied with two electrons. From Equations 4.12 and 4.13, we know the relation between k and E for free electrons: $E(k) \propto k^2$, and thus $k \propto \sqrt{E}$ and $dk/dE \propto 1/\sqrt{E}$. By using Equation 4.19, we obtain the density of states for a one-dimensional electron gas:

$$D_{1d}(E) = \frac{dN(E)}{dE} = \frac{dN(k)}{dk}\frac{dk}{dE} \propto 1 \times 1/\sqrt{E} \propto 1/\sqrt{E} \tag{4.20}$$

The density of states is depicted in Figure 4.6. In one-dimensional systems the density of states has an $E^{-1/2}$ dependence and thus exhibits singularities near the band edges [22]. Each of the hyperbolas contains a continuous distribution of k_x states, but only one discrete k_y and k_z state.

The quantization of states in two dimensions has important consequences for the transport of charges. Electrons can only flow freely along the x-axes but are limited to discrete states in the y- and z-directions. Therefore, they are only transported in

discrete "conductivity channels". This may be of considerable importance for the microelectronics industry. If the size of electronic circuits is reduced more and more, at one point the diameter of wires will become comparable to the de Broglie wavelength of the electrons. The wire will then exhibit the behavior of a quantum wire. Quantum aspects of 1D transport were first observed in so-called quantum point contacts which were lithographically defined in semiconductor heterostructures [29, 30]. More recent examples for such 1D wires include short organic semiconducting molecules [31–36], inorganic semiconductor and metallic nanowires [37–42] and break junctions [43–45]. A particular role is played by carbon nanotubes [32, 46–52]. Carbon nanotubes have been extensively studied both as model systems for one-dimensional confinement and for potential applications, such as electron emitters [53].

4.4.4
Zero-Dimensional Systems (Quantum Dots)

When charge carriers and excitations are confined in all three dimensions, the system is called "quantum dot". The division is somewhat arbitrary since, for instance, clusters made of very few atoms are not necessarily considered as quantum dots. Although clusters are smaller than the de Broglie wavelength, their properties depend critically on their exact number of atoms. Larger clusters have a well-defined lattice and their properties no longer depend critically on their exact number of atoms. We shall then refer to such systems with the term "quantum dots" [54–64] (Figure 4.7).

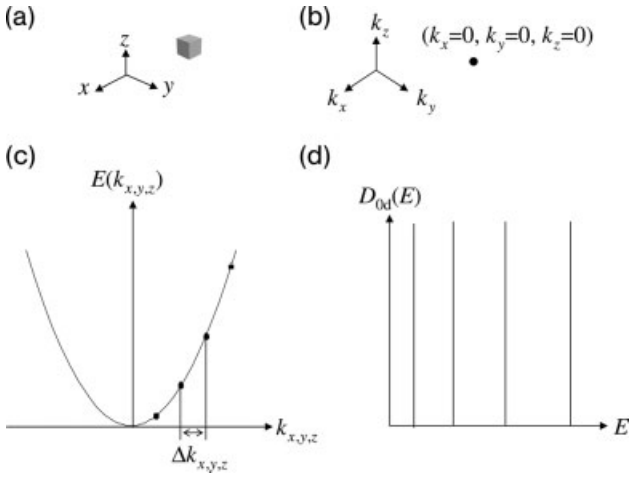


Figure 4.7 A zero-dimensional solid. (a) The solid is shrunk in all three dimensions to a thickness that is comparable to the de Broglie wavelength of its charge carriers. (b) Because of such confinement, all states (k_x, k_y,k_z) are discrete points in the three-dimensional *k*-space. (c) Only discrete energy levels are allowed. (d) The one-dimensional density of states $D_{0d}(E)$ contains delta peaks, that correspond to the individual states. Electrons can occupy only states with these discrete energies.

In a quantum dot, the movement of electrons is confined in all three dimensions and there are only discrete (k_x, k_y, k_z) states in the k-space. Each individual state in k-space can be represented by a point. The final consequence is that only discrete energy levels are allowed, which can be seen as delta peaks in the distribution $D_{0d}(E)$. As we can see, the energy bands converge to atom-like energy states, with the oscillator strength compressed into few transitions. This change is most dramatic at the edges of the bands and influences semiconductors more than metals. In semiconductors, the electronic properties are in fact strongly related to the transitions between the edges of the valence band and the conduction band. In addition to the discreteness of the energy levels, we want to stress again the occurrence of a finite zero-point energy. In a dot, even in the ground state electrons have energies higher than bulk electrons at the conduction band edge. These points will be discussed in more detail in the next section.

4.5
Energy Levels of a (Semiconductor) Quantum Dot

In this section, we will describe in more detail a zero-dimensional solid. Since many quantum effects are more pronounced in semiconductors than metals, we will focus on the case of a semiconducting material. In Section 4.4 we described how the properties of a free electron gas change when the dimensions of the solid are reduced. The model of the free electron gas does not include the "nature" of the solid. However, from a macroscopic point of view we distinguish between metals, semiconductors and insulators [20]. The model of a free electron gas describes relatively well the case of electrons in the conduction band of metals. On the other hand, electrons in an insulating material are only poorly described by the free electron model. In order to extend the model of free electrons for semiconducting materials, the concept of a new charge carrier, the hole, was introduced [21]. If one electron from the valence band is excited to the conduction band, the "empty" electronic state in the valence band is called a hole. Some basic properties of semiconducting materials can be described by the model of free electrons and free holes. The energy bands for electrons and holes are separated by a bandgap [20, 21]. The dispersion relations for the energy of electrons and holes in a semiconductor are parabolic to a first approximation. This approximation holds true only for electrons (holes) occupying the levels that lie at the bottom (top) of the conduction (valence) band. Each parabola represents a quasi-continuous set of electron (hole) states along a given direction in k-space. The lowest unoccupied energy band and the highest occupied energy band are separated by an energy gap E_g(bulk), as shown in Figure 4.8. The bandgap for a bulk semiconductor can range from a fraction of an electronvolt up to a few electronvolts.

We could expect that the energy dispersion relations would still be parabolic in a quantum dot. However, since only discrete energy levels can exist in a dot, each of the original parabolic bands of the bulk case is now fragmented into an ensemble of points. The energy levels of a quantum dot can be estimated with the particle-in-a-box

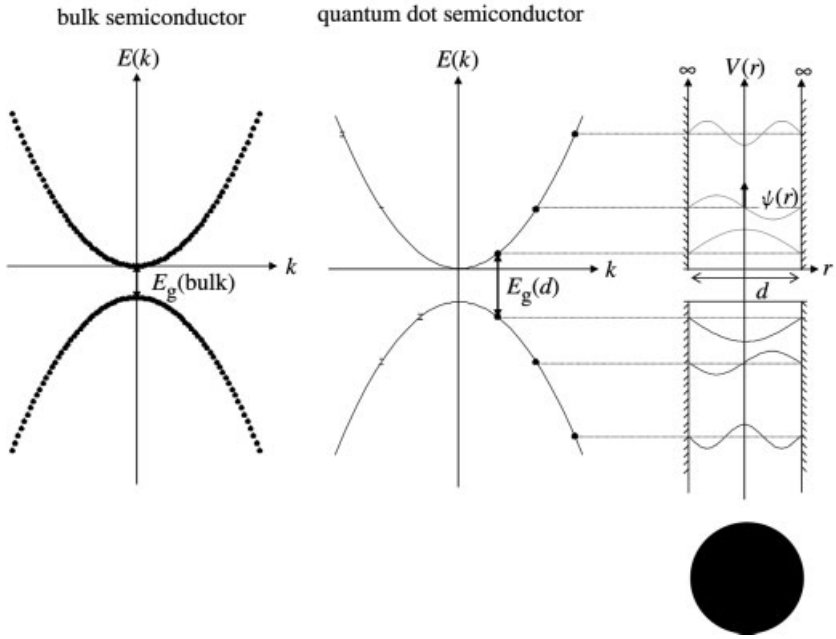


Figure 4.8 Free charge carriers in a solid have a parabolic dispersion relation [$E(k) \propto k^2$]. In a semiconductor the energy bands for free electrons and holes are separated by an energy gap E_g. In a bulk semiconductor, the states are quasi-continuous and each point in the energy bands represents an individual state. In a quantum dot the charges are confined to a small volume. This situation can be described as a charge carrier confined in an infinite potential well of width d. Here, the width d of the potential well corresponds to the diameter of the quantum dot. The only allowed states are those whose wavefunctions vanish at the borders of the well 8. This leads to discrete energy levels [8, 11]. The energy gap between the lowest possible energy level for electrons and holes $E_g(d)$ is larger than that of a bulk material E_g(bulk).

model. As described in the previous section (Figure 4.4), the lowest energy for an electron in a one-dimensional potential well is

$$E_{\text{well,1d}} = (1/8)h^2/md^2 \tag{4.21}$$

where d is the width of the well. In a quantum dot, the charge carriers are confined in all three dimensions and this system can be described as an infinite three-dimensional potential well. The potential energy is zero everywhere inside the well but is infinite on its walls. We can also call this well a "box". The simplest shapes for a three-dimensional box can be, for instance, a sphere or a cube. If the shape is cubic, the Schrödinger equation can be solved independently for each of the three translational degrees of freedom and the overall zero-point energy is simply the sum of the individual zero point energies for each degree of freedom [10, 65]:

$$E_{\text{well,3d(cube)}} = 3E_{\text{well,1d}} = (3/8)h^2/md^2 \tag{4.22}$$

If the box is a sphere of diameter d, the Schrödinger equation can be solved by introducing spherical coordinates and by separating the equation in a radial part and in a part that contains the angular momentum [66, 67]. The lowest energy level (with angular momentum = 0) is then

$$E_{\text{well,3d(sphere)}} = (1/2)h^2/md^2 \tag{4.23}$$

The effect of quantum confinement is again remarkable. More confined charge carriers lead to a larger separation between the individual energy levels, and also to a greater zero-point energy. If carriers are confined into a sphere of diameter d, the zero-point energy is higher than that for charges that are confined to a cube whose edge length is equal to d [$E_{\text{well,3d(sphere)}} > E_{\text{well,3d(cube)}}$]. This is because such a sphere simply has a smaller volume [$(\pi/6)d^3$] than the cube (d^3).

An electron–hole pair can be generated in the quantum dot, for instance by a photoinduced process or by charge injection. The minimum energy E_g required for creating an electron–hole pair in a quantum dot is made up of several contributions. One contribution is the bulk band gap energy, E_g(bulk). Another important contribution is the confinement energy for the carriers, which we call $E_{\text{well}} = E_{\text{well}}(e^-) + E_{\text{well}}(h^+)$. For large particles (bulk: $d \to \infty$), E_{well} tends to zero. We can estimate the overall confinement energy for an electron–hole pair in a spherical quantum dot. It is the zero point energy of the potential well or in other words the energy of the state of a potential box with the lowest energy. This can be written as

$$E_{\text{well}} = h^2/2m^*d^2 \tag{4.24}$$

where m^* is the reduced mass of the exciton and is given by [68]

$$1/m^* = 1/m_e + 1/m_h \tag{4.25}$$

were m_e and m_h are the effective masses for electrons and holes, respectively. In order to calculate the energy required to create an electron–hole pair, another term (E_{Coul}) has to be considered. The Coulomb interaction E_{Coul} takes into account the mutual attraction between the electron and the hole, multiplied by a coefficient that describes the screening by the crystal. In contrast to E_{well}, the physical content of this term can be understood within the framework of classical electrodynamics. However, an estimate of such a term is only possible if the wavefunctions for the electron and the hole are known. The strength of the screening coefficient depends on the dielectric constant ε of the semiconductor. An estimate of the coulomb term yields

$$E_{\text{Coul}} = -1.8e^2/2\pi\varepsilon\varepsilon_0 d \tag{4.26}$$

This term can be fairly significant because the average distance between an electron and a hole in a quantum dot can be small [13, 14, 55, 69, 70]. We can now estimate the size-dependent energy gap of a spherical semiconductor quantum dot, which is given by the following expression [13, 14, 55, 68–70]:

$$E_g(\text{dot}) = E_g(\text{bulk}) + E_{\text{well}} + E_{\text{Coul}} \tag{4.27}$$

Then, by inserting Equations 4.24 and 4.26 into Equation 4.27, we obtain

$$E_g(d) = E_g(\text{bulk}) + h^2/2m^*d^2 - 1.8e^2/2\pi\varepsilon\varepsilon_0 d \tag{4.28}$$

Here we have emphasized the size dependence in each term. Equation 4.28 is only a first approximation. Many effects, such as crystal anisotropy and spin–orbit coupling, have to be considered in a more sophisticated calculation. The basic approximation for the bandgap of a quantum dot comprises two size-dependent terms: the confinement energy, which scales as $1/d^2$, and the Coulomb attraction, which scales as $1/d$. The confinement energy is always a positive term and thus the energy of the lowest possible state is always raised with respect to the bulk situation. On the other hand, the Coulomb interaction is always attractive for an electron–hole pair system and therefore lowers the energy. Because of the $1/d^2$ dependence, the quantum confinement effect becomes the predominant term for very small quantum dot sizes (Figure 4.9).

The size-dependent energy gap can be a useful tool for designing materials with well-controlled optical properties. A much more detailed analysis on this topic can be found in, for example, a paper by Efros and Rosen [61].

In this chapter, we have shown how the dependence of the energy gap of semiconductors on the size of the material can be explained by either shrinking down the material from bulk to nanometer dimensions or by assembling the material atom by atom. Both views, one in a top-down and the other in a bottom-up approach, ultimately lead to the same physics. In Chapter 3, a more general description of these two types of approaches is given.

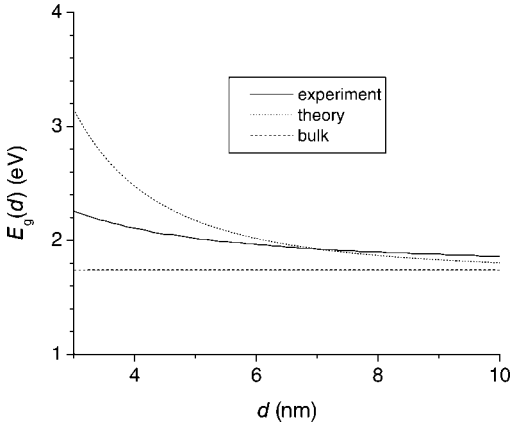


Figure 4.9 Size dependence of the energy gap $E_g(d)$ for colloidal CdSe quantum dots with diameter d. The bulk value for the energy gap is $E_g(\text{bulk}) = 1.74$ eV [68]. The theoretical curve was obtained using Equation 4.28 with the following parameters: effective mass of electrons/holes $m_e = 0.13m_0$, $m_h = 0.4m_0$, m_0 = mass of free electrons ($m = 9.1095 \times 10^{-31}$ kg) $\Rightarrow m^* = 0.098m$ [68]; dielectric constant $\varepsilon_{CdSe} = 5.8$ [71], permittivity constant $\varepsilon_0 = 8.854 \times 10^{-12}\,C^2\,N^{-1}\,m^{-2}$, Planck's constant $h = 6.63 \times 10^{-34}$ J s, $1\,eV = 1.602 \times 10^{-19}$ J. The experimental values were obtained by recording the absorption spectra of CdSe quantum dots of different sizes and determining the size of the quantum dots by transmission electron microscopy (TEM) [72].

References

1 Parak, W.J., Manna, L., Simmel, F.C., Gerion, D. and Alivisatos, P. (2004) *Nanoparticles – from Theory to Application* (ed. G. Schmid), 1st edn., Wiley-VCH, Weinheim, 4.

2 Lane, N. (2001) *Journal of Nanoparticle Research*, **3**, 95.

3 Service, R.F. (2000) *Science*, **290**, 1526.

4 Kingon, A.I., Maria, J.-P. and Streiffer, S.K. (2000) *Nature*, **406**, 1032.

5 Lloyd, S. (2000) *Nature*, **406**, 1047.

6 Ito, T. and Okazaki, S. (2000) *Nature*, **406**, 1027.

7 Peercy, P.S. (2000) *Nature*, **406**, 1023.

8 Cohen-Tannoudji, C., Diu, B. and Laloe, F. (1997) *Quantum Mechanics*, 1st edn., Wiley, New York.

9 Yoffe, A.D. (2001) *Advances in Physics*, **50**, 1.

10 Atkins, P.W. (1986) *Physical Chemistry*, 4th edn., W. H. Freeman, New York.

11 Karplus, M. and Porter, R.N. (1970) *Atoms and Molecules*, 1st edn., W. A. Benjamin, New York.

12 Alivisatos, A.P. (1997) *Endeavour*, **21**, 56.

13 Brus, L.E. (1983) *Journal of Chemical Physics*, **79**, 5566.

14 Brus, L.E. (1984) *Journal of Chemical Physics*, **80**, 4403.

15 Lippens, P.E. and Lannoo, M. (1989) *Physical Review B-Condensed Matter*, **39**, 10935.

16 Delerue, C., Allan, G. and Lannoo, M. (1993) *Physical Review B-Condensed Matter*, **48**, 11024.

17 Wang, L.-W. and Zunger, A. (1996) *Physical Review B-Condensed Matter*, **53**, 9579.

18 Fu, H., Wang, L.-W. and Zunger, A. (1998) *Physical Review B-Condensed Matter*, **57**, 9971.

19 Schrier, J. and Wang, L.-W. (2006) *Journal of Physical Chemistry B*, **110**, 11982.

20 Kittel, C. (1989) *Einführung in die Festkörperphysik*, 8th edn., R. Oldenbourg Verlag, Munich.

21 Ashcroft, N.W. and Mermin, N.D. (1976) *Solid State Physics*, Saunders College, Phiadelphia, PA.

22 Davies, J.H. (1998) *The Physics of Low-dimensional Semiconductors*, Cambridge University Press, Cambridge.

23 Ando, T., Fowler, A.B. and Stern, F. (1982) *Reviews of Modern Physics*, **54**, 437.

24 Moriarty, P. (2001) *Reports on Progress in Physics*, **64**, 297.

25 Zhitenev, N.B., Fulton, T.A., Yacoby, A., Hess, H.F., Pfeiffer, L.N. and West, K.W. (2000) *Nature*, **404**, 473.

26 Suen, Y.W., Engel, L.W., Santos, M.B., Shayegan, M. and Tsui, D.C. (1992) *Physical Review Letters*, **68**, 1379.

27 Stormer, H.L. (1998) *Solid State Communications*, **107**, 617.

28 Stormer, H.L., Du, R.R., Kang, W., Tsui, D.C., Pfeiffer, L.N., Baldwin, K.W. and West, K.W. (1994) *Semiconductor Science and Technology*, **9**, 1853.

29 Wharam, D.A., Thornton, T.J., Newbury, R., Pepper, M., Ahmed, H., Frost, J.E.F., Hasko, D.G., Peacock, D.C., Ritchie, D.A. and Jones, G.A.C. (1988) *Journal of Physics C: Solid State Physics*, **21**, L209.

30 van Wees, B.J., van Houten, H., Beenakker, C.W.J., Williams, J.G., Kouwenhoven, L.P., van der Marel, D. and Foxon, C.T. (1988) *Physical Review Letters*, **60**, 848.

31 Bumm, L.A., Arnold, J.J., Cygan, M.T., Dunbar, T.D., Burgin, T.P., L. Jones, II Allara, D.L., Tour, J.M. and Weiss, P.S. (1996) *Science*, **271**, 1705.

32 Anantram, M.P., Datta, S. and Xue, Y.Q. (2000) *Physical Review B-Condensed Matter*, **61**, 14219.

33 Cobden, D.H. (2001) *Nature*, **409**, 32.

34 Freemantle, M. (2001) *Chemical & Engineering News*, 5 March, 38.

35 Cui, X.D., Primak, A., Zarate, X., Tomfohr, J., Sankey, O.F., Moore, A.L., Moore, T.A., Gust, D., Harris, G. and Lindsay, S.M. (2001) *Science*, **294**, 571.

36 Reed, M.A. (2001) *MRS Bulletin*, 113.

37 Hu, J.T., Odom, T.W. and Lieber, C.M. (1999) *Accounts of Chemical Research*, **32**, 435.

38 Cui, Y., Duan, X., Hu, J. and Lieber, C.M. (2000) *Journal of Physical Chemistry B*, **104**, 5213.

39 Rodrigues, V., Fuhrer, T. and Ugarte, D. (2000) *Physical Review Letters*, **85**, 4124.

40 Rao, C.N.R., Kulkarni, G.U., Govindaraj, A., Satishkumar, B.C. and Thomas, P.J. (2000) *Pure and Applied Chemistry*, **72**, 21.

41 Häkkinen, H., Barnett, R.N., Scherbakov, A.G. and Landman, U. (2000) *Journal of Physical Chemistry B*, **104**, 9063.

42 Cui, Y. and Lieber, C.M. (2001) *Science*, **291**, 851.

43 Reed, M.A., Zhou, C., Muller, C.J., Burgin, T.P. and Tour, J.M. (1997) *Science*, **278**, 252.

44 van den Brom, H.E., Yanson, A.I. and Ruitenbeek, J.M. (1998) *Physica B*, **252**, 69.

45 Xe, H.X., Li, C.Z. and Tao, N.J. (2001) *Applied Physics Letters*, **78**, 811.

46 Tans, S.J., Devoret, M.H., Dai, H., Thess, A., Smalley, R.E., Geerligs, L.J. and Dekker, C. (1997) *Nature*, **386**, 474.

47 Saito, S. (1997) *Science*, **278**, 77.

48 McEuen, P.L., Bockrath, M., Cobden, D.H. and Lu, J.G. (1999) *Microelectronic Engineering*, **47**, 417.

49 Yao, Z., Postma, H.W.C., Balents, L. and Dekker, C. (1999) *Nature*, **402**, 273.

50 Odom, T.W., Huang, J.-L., Kim, P. and Lieber, C.M. (2000) *Journal of Physical Chemistry B*, **104**. 2794.

51 McEuen, P.L. (2000) *Physics World*, June 2000, 31.

52 Jacoby, M. (2001) *Chemical & Engineering News*, 30 April 13.

53 de Heer, W.A., Chatelain, A. and Ugarte, D. (1995) *Science*, **270**, 1179.

54 Bastard, G. and Brum, J.A. (1986) *IEEE Journal of Quantum Electronics*, **QE22**, 1625.

55 Bawendi, M.G., Steigerwald, M.L. and Brus, L.E. (1990) *Annual Review of Physical Chemistry*, **41**, 477.

56 Alivisatos, A.P. (1996) *Science*, **271**, 933.

57 Alivisatos, A.P. (1998) *MRS Bulletin*, **23**, 18.

58 Kouwenhoven, L.P., Marcus, C.M., McEuen, P.L., Tarucha, S., Westervelt, R.M. and Wingreen, N.S. (1997) *Mesoscopic Electron Transport, NATO ASI Series E*, (ed. L.P.K.L.L. Sohn), Kluwer, Dordrecht.

59 Warburton, R.J., Miller, B.T., Dürr, C.S., Bördefeld, C., Kotthaus, J.P., Medeiros-Riberio, G., Petroff, P.M. and Huant, S. (1998) *Physical Review B-Condensed Matter*, **58**, 16221.

60 Alivisatos, P. (2000) *Pure and Applied Chemistry*, **72**, 3.

61 Efros, A.L. and Rosen, M. (2000) *Annual Review of Materials Science*, **30**, 475.

62 Soloviev, V.N., Eichhofer, A., Fenske, D. and Banin, U. (2000) *Journal of the American Chemical Society*, **122**, 2673.

63 Zrenner, A. (2000) *Journal of Chemical Physics*, **112**, 7790.

64 Petroff, P.M., Lorke, A. and Imamoglu, A. (2001) *Physics Today*, May, 46.

65 Landau, L.D. and Lifschitz, E.M. (1979) *Quantenmechnik*, 9th edn., Vol. 3, Akademie-Verlag, Berlin.

66 Schwabl, F. (1990) *Quantenmechanik*, 2nd ed. Springer, Berlin.

67 Messiah, A. (1976) *Quantenmechanik*, Band 1 Walter de Gruyter, Berlin.

68 Trindade, T., O'Brien, P. and Pickett, N.L. (2001) *Chemistry of Materials*, **13**, 3843.

69 Brus, L. (1986) *Journal of Physical Chemistry*, **90**, 2555.

70 Steigerwald, M.L. and Brus, L.E. (1990) *Accounts of Chemical Research*, **23**, 183.

71 Gorska, M. and Nazarewicz, W. (1974) *Physica Status Solidi B-Basic Research*, **65**, 193.

72 Yu, W.W., Qu, L., Guo, W. and Peng, X. (2003) *Chemistry of Materials*, **15**, 2854.

5
Fundamentals and Functionality of Inorganic Wires, Rods and Tubes

Jörg J. Schneider, Alexander Popp, and Jörg Engstler

5.1
Introduction

Nanostructured one-dimensional inorganic tubes, wires and rods are known for a variety of single elements and combinations thereof. The number of studies towards their synthesis and properties is now vast. For nanowires – anisotropic nanocrystals with large aspect ratio (length to diameter) – around 5000 papers have been published during the last 2 years. An excellent comprehensive monograph and timely reviews presenting the current state of the art up to 2005 exist [1].

In this chapter, physical properties of 1D inorganic structures will be discussed first, followed by a section devoted to general techniques for the synthesis of inorganic wires, rods and tubes. The chapter then highlights some of the material developments made over the last few years in the very active area of nanostructured inorganic rods, wires and tubes with respect to their specific materials functionality. [The field of carbon nanotubes (CNTs) is probably still the fastest growing of all 1D materials. CNTs will only be touched upon in this chapter as far as sensing and nano–micro integration in functional devices are concerned. For further reading on this topic, the reader is referred to numerous excellent monographs in the field.] Due to the breadth of the field, the selection of materials and applications is somewhat subjective and reflects what the authors personally feel are "hot" topics. Where could the materials discussed impact on future technological developments? Fields of sensing and micro/nanoelectronics integration will be selectively addressed here. It is the intention of this chapter to introduce the reader to these fields and the currently ongoing rapid developments in these promising future fields of functional 1D nanomaterials.

A drastic change in materials properties is often connected with the nanoscale range (1–100 nm). In addition to an understanding of fundamental size-related electronic effects [quantum size effects (QSE)], which are connected with this "miniaturization" of matter (for an intriguing description of how quantum phenomena arise in 0D and 1D nanostructured matter, see Chapter 4), interest in nanostructured materials often arises from the fact that the small size connected with nanoscaled matter creates new chemistry. For example, the extremely high number of interfaces connected with

Nanotechnology. Volume 1: Principles and Fundamentals. Edited by Günter Schmid

ISBN: 978-3-527-31732-5

small-scale matter, be it in 0D (particles), 1D (wires, rods, tubes) or 2D (films) dimensions create high chemical reactivity. Interfaces control important material properties such as catalytic activity or analytical sensing behavior in addition to electronic properties of nanomaterials, which are highly dependent on such interfacial contacts of individual nano-building blocks and also on the individual QSE of the nano-building blocks. Besides such effects connected with the nanoscale regime, morphological properties of assembled nanomaterials such as habit (size, shape) and surface structure are also important for new and desired materials properties arising from the sequential build-up of larger structures from nano-building blocks.

The morphology of mesostructures can be tailored by synthetic techniques, for example, self-assembly, which arranges individual small building blocks such as molecules or even nanoparticles into larger mesostructured objects by employing secondary interactions such as hydrogen bonding, van der Waals, capillary or hydrophobic forces. Therefore, the shaping of materials, be it on the molecular or the nanoscale, presents a major task to the experimentalist since it is the key to organizing matter on the mesoscopic scale (Greek *mesos*, in between); the dimension between the pure nanostructured regime in which QSE reign nearly every type of material property and the macroscopic world in which solid-state physics is the key to describing and understanding material properties.

A very successful and intriguing example of controlled 1D assembly from the chemist's workbench can be found in the technique of supramolecular assembly of compact or porous 1D structures. Such structures can be built up sequentially starting first from defined molecular precursors, followed by self-assembly of 0D (nano)particle aggregates via controlled condensation. In this approach, the individual building blocks arrange in hierarchical order (from the molecular to the nanoscale and finally to the mesoscale) by supramolecular organization (Figure 5.1) [2].

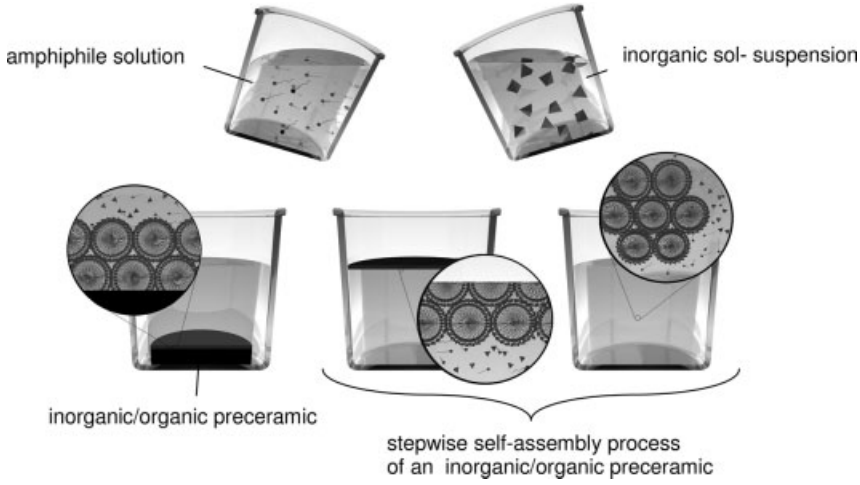


Figure 5.1 Schematic of the process of self assembly of inorganic sol particles on a polymeric micellar amphiphile assembly leading to an organic/inorganic composite. The composite structure is formed via self assembly. Adapted from Ref. [2f].

A polymeric macromolecule (called a template) steers, in addition to temperature, solvent and concentration of all reactants, the complex arrangement of individual building blocks into the final macro-sized structure.

5.2
Physical Properties of 1D Structures

Starting from a three-dimensional solid, the confinement of electrons into a one-dimensional structure leads to the quantization of electronic states in two directions (lets us say x and y). In the z-direction electrons can move freely and give rise to a quasi-continuous distribution of energy states along this dimension. Along x and y they are confined and only one discrete state is possible. Once the diameter of the 1D system is comparable to the de Broglie wavelength of the electrons, the 1D structure will become a quantum wire. Probably the most important application-related aspect of this dimensionality reduction in 1D materials is the restricted flow of charge carriers in only a single direction, the "conductivity channel" (see Section 5.4.3.1). An intriguing example can be found in single crystalline silicon nanowires smaller than

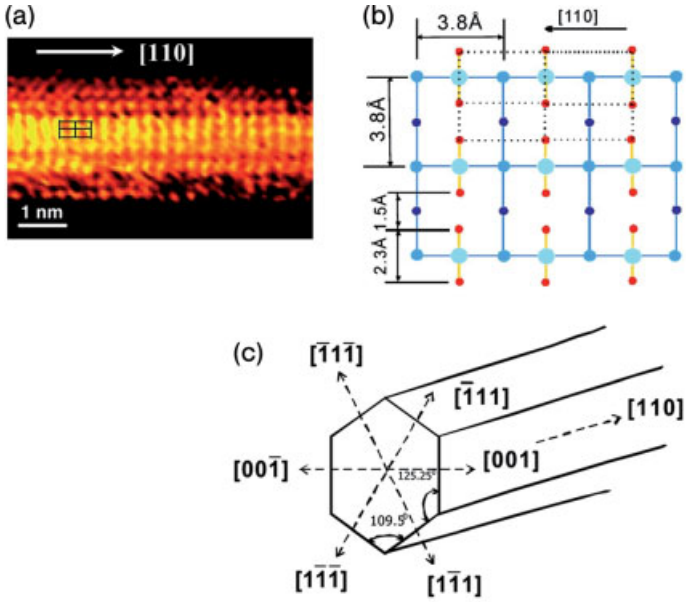


Figure 5.2 Scanning tunneling microscopic (STM) image and schematic view of an Si nanowire with an Si (0 0 1) facet. (a) Constant-current STM image of an Si nanowire on a HOPG substrate. The wire's axis is along the [110] direction. (c) Schematic view of an Si nanowire bounded by four (11)-type facets and two (0 0 1)-type facets. The wire's axis is along the [110] direction. Reprinted with permission from Ref. [3].

5 nm in diameter (5 nm is the exciton size for Si). Via scanning tunneling microscopy (STM), the electronic states and the bandgaps of such Si wires have been probed for wires with different diameters <10 nm. Gaps ranging from 1.1 eV (7 nm Si wire diameter) up to 3.5 eV (1.3 nm Si wire diameter) have been determined, demonstrating the extreme quantum confinement effect in such structures (Figure 5.2) [3].

A variety of direction-dependent electronic effects, for example, polarizability of light, are also different for anisotropic 1D materials compared with 0D materials [4].

Although electron conductivity in 1D wires and rods is favored along the preferred direction, phonon transport is greatly impeded, in thin 1D nanostructures, due to boundary scattering in the confined directions. Electrons may suffer elastic scattering events during their journey along the wire. This has important implications on the heat conductivity of 1D nanowire structures and for application of such structures in wiring or circuiting next-generation semiconductor devices. On the other hand, poor heat transport in confined nanowires may be used for development of thermoelectric materials. The thermoelectric effect (Seebeck effect), which is responsible for this property, describes the enhancement of the thermal electronic conductivity through a material as phonon transport in the structure worsens (due to the 1D confinement effect). Theory has predicted a significant increase over bulk values of this thermoelectric effect depending on the diameter, composition and charge carrier concentration of the 1D material of choice [5]. Nevertheless, research in that technologically important area is still in its infancy. An example is discussed in Section 5.3.1. For a detailed discussion of mesoscopic transport phenomena, for example, boundary scattering in 1D confined structures, the reader is referred to Ref. [6].

5.3
Synthetic Methods for 1D Structures

Synthetic methods for obtaining 1D nanostructures are numerous and have been reviewed by various authors recently [7]. They can be divided into top-down and bottom-up techniques; in the former a desired nanostructure is formed by sophisticated physical techniques, for example, electron beam structuring or laser structuring of a bulk material. Such techniques, however, are always restricted to the wavelength of the structuring beam, thus nanomaterials with dimensions well below 10 nm are still beyond the scope of these methods.

Bottom-up techniques, however, show a rich diversity towards desirable chemical and materials compositions and also structure (amorphous, crystalline), morphology (0D, 1D and 2D) and size. Several bottom-up synthetic techniques leading to the formation of 1D structures which seem to have a more general impact can be identified: the template method, self-assembly techniques, vapor to liquid–solid synthesis of 1D nanostructures and electrochemical techniques. These techniques show an enormous breadth since a variety of elemental compositions for different structures and morphologies and also sizes are accessible. Often combinations of individual techniques are employed, broadening further the scope for the experimentalist.

5.3.1
The Template Approach

This is probably the most versatile method for the synthesis of 1D structures. A host structure with a pore morphology – the template – is filled with a compact, more or less dense material or as film. The former produces compact wires or rods and the latter results in the formation of tubes.

Filling of the pores is possible either via solution (simply by capillary filling), by electrochemical deposition techniques or via the gas phase.

Oxidation of valve metals such as aluminum, titanium, tantalum and hafnium leads to porous metal oxide films on metal surfaces [8]. For aluminum this is a useful and outstanding technique to prepare both surface-attached and free-standing porous 2D alumina films (after detachment from the metal surface) with varying pore diameters. The pore size of these films is strongly dependent on the experimental conditions and can be varied between 10 and several hundred nm [9]. Especially in the case of aluminum, these films can be fairly thick (up to several tens of μm) or thin (down to several hundred nm), but still self-supporting, free-standing and handable. The diameter of the pores (D_p) and the cells (D_c) of a porous alumina membrane are dependent on the anodization potential applied in the electrolytic process [9]. Additional experimental parameters governing the pore size are temperature and current density. The latter is influenced by the concentration and type of electrolyte used in the electrolysis process. Nearly perfectly ordered pores are accessible by prestructuring the metallic surface [9a] (Figure 5.3).

Consequently, porous alumina membranes have been widely used to prepare various types of mesoscale materials within their pores. Their enormous synthetic impact in the area of mesostructured 1D materials comes from the ability to combine well-established wet chemical synthesis techniques (e.g., sol–gel chemistry) with the straightforward synthesis and subsequent filling of porous alumina templates.

Preparation of active sol precursors, filling and aging within the mesopores of alumina, followed by final calcination steps, have led to a huge variety of different mesostructured 1D materials, for example, in the ceramics field [1e,11]. The pores can even be used to arrange silica in a columnar or circular arrangement with a defined internal mesostructure. Further entrapment of a variety of 1D nanostructures into the mesopores of alumina such as metallic rods (Pt, Au, Pd), semiconductor rods and carbon nanostructures have been reported recently [1e,11].

Using a combination of sol–gel processing on the surface of (1 0 0)-oriented silicon templates, a porous structure with an ordered arrangement of pores is formed. These studies show that a topographic structure (here Si) can be used to engineer pores on this solid surface, allowing a high degree of freedom. The filling of such pores with cobalt has been demonstrated [12].

Another template technique giving porous, free-standing membranes, but of polymeric materials, uses heavy ion track etching of polymer foils [13] to fill the statistically formed ion tracks within the polymer with nanoscale materials in order to form rod-like structures [14, 15] (Figure 5.4). This led to bunches of randomly

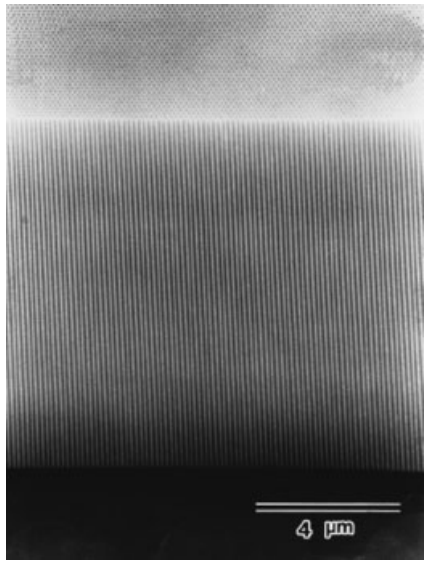


Figure 5.3 Cross-sectional view of a channel array of a porous alumina membrane which can be used for arranging 1D nanomaterials. The scanning microscopic (SEM) observation was carried out without removal of Al and barrier layer of the oxide film. Reprinted with permission from Ref. [9i].

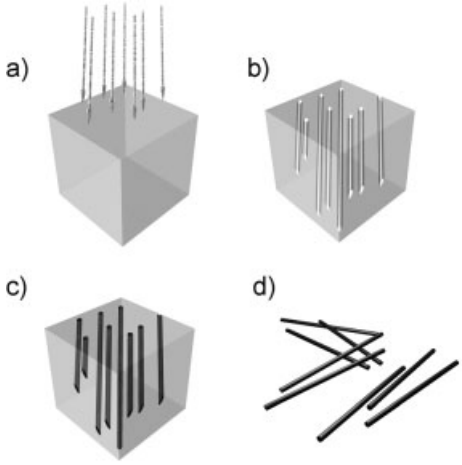


Figure 5.4 Ion track etching process of a polymer film. Heavy ions hit the polymer foil, and generate tracks in the membrane (a). The tracks are stochastically arranged (b); solid rods are synthesized in the tracks, for example, via electrochemical deposition (c); free rods after complete polymer membrane dissolution (d).

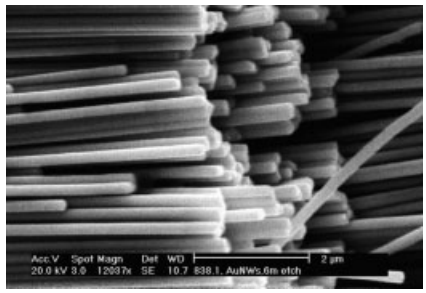


Figure 5.5 Bunch of aligned gold nanowires, synthesized in a polymer membrane, after dissolution of the membrane. Reproduced with permission from Ref. [14].

oriented composite 1D structures. Careful dissolution of the template gives free-standing aligned metallic rods with well-defined crystal lattices (Figures 5.5 and 5.6).

Using the template approach in connection with electrodeposition allows the pores of such templates to be filled with single- or multicomponent metallic rods. After deposition of a metal electrode on the back side of such a porous template, one or even more metals can be densely deposited inside the pores on purpose. This has been used, for example, for deposition of Au/Pt segments. The deposition of gradient materials via electrodeposition is also possible. This technique can also be has used for the deposition of magnetic structures. Their magnetic behavior depends on the aspect (length-to-diameter) ratio of the individual rods. The easy axis of magnetization is parallel to the nanorod or nanowire axis if the electrodeposited structure is longer than its width [16]. Otherwise, it is perpendicular to the deposited structure. This difference depending on particle shape (morphology, platelet vs. rod structure) can be used to align bimetallic nanorods side-by-side (Figure 5.7) [17a] or one after the other when the magnetic field is applied parallel to the substrate and to the 1D magnetic materials easy axis [17b].

Hybrid Co/Au nanorod structures are accessible by using organometallic chemistry. The appropriate choice of the molecular Co and Au precursors and the

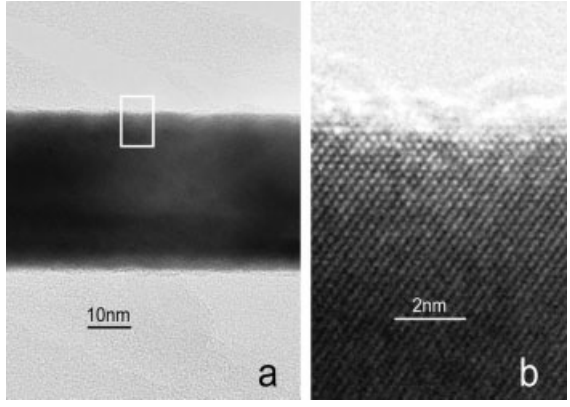


Figure 5.6 High-resolution TEM micrograph of a 70-nm single-crystalline Au nanowire. Low magnification (a); enlargement area (b). Reproduced with permission from Ref. [14].

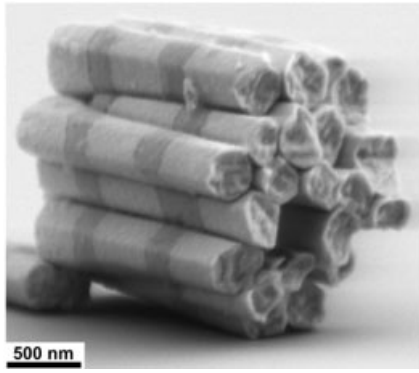


Figure 5.7 SEM image of a nanorod bundle stapled via attractive forces between disk-shaped magnetic inner-rod sections. Reproduced with permission from Ref. [17].

stabilizing ligand allows control of the growth process and of the overall morphology of the rod (tip or whole body growth) [18]. Combining this technique with the template method should lead to 2D-arranged hybrid metallic rods.

Devising methods for aligning 1D materials with a uniform growth front is of general importance for optimizing device performance, for example, in thermoelectric materials such as bismuth telluride. For example, overgrowth of rods when the template pores are already filled has to be avoided in order not to cover the pores of the template with an active material, which may lead to short-circuiting. This has been achieved by applying a pulsed potential deposition technique and has yielded a uniform growth front of Bi_2Te_3 nanowire arrays in porous alumina [19]. Using nanorods as sacrificial templates to generate polymeric or ceramic nanorod structures is another method that uses porous alumina as the initial template and shows the high versatility of this porous template structure. Electrodeposition of, for example, nickel rods into porous alumina, followed by dissolution of the oxide template and coating of the metallic wires with either organic polymers or inorganic polyelectrolyte ceramics, finally gives the corresponding polymeric or ceramic (after subsequent calcinations) 1D structures. The complete method uses a hybrid technique of first templating followed by a layer-by-layer deposition technique [20].

Although the nano-templating approach using porous alumina (Figure 5.8) works well for template diameters down to about 20 nm, reports on 1D nanomaterials with smaller diameters prepared with this so-called "hard template technique" [1e] are still scarce so far [21].

5.3.2 Electrochemical Techniques

5.3.2.1 Electrospinning

This technique can be used to create nano- to microscale fibers of mainly polymeric materials [22]. Recent advances in the technique of electrospinning have brought this

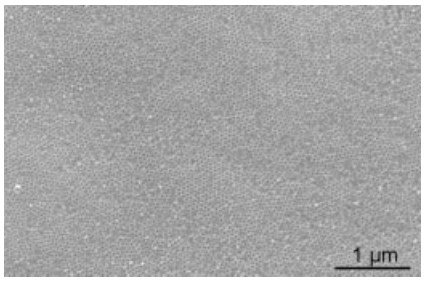


Figure 5.8 Scanning electron microscopic (SEM) view of a porous alumina template membrane with pores in the range 15–17 nm (J.J. Schneider, J. Engstler, M. Naumann, TU Darmstadt).

already well-established method again in the focus of the experimentalists as a valuable method for synthesizing large quantities of, for example, ceramic nanofibers of various compositions with dense or hollow morphology [22–24].

In the technique, a precursor solution is held in the hollow tip of a needle-like capillary by its own surface tension. This capillary is subjected to an electric field and induces an electric field on the surface of the liquid precursor. On reaching a critical value, the repulsive electric field overcomes the surface tension and a charged jet of the solution is ejected from the capillary. Once the jet has started, its trajectory can be controlled by the electric field. These charged fibers are then deposited on a grounded collector. The process depends on a number of parameters such as solution properties (viscosity, conductivity and surface tension) and also process parameters such as hydrostatic pressure and electric potential at the capillary tip, distance between the tip and the collecting metal plate and general parameters such as temperature and humidity [23]. This technique gives access to isolated fibers [24–26], fiber agglomerates (mats) and dense or hollow fibers of different composition (e.g., single phase or composite) [27]. However, experimental electrospinning conditions are critical for the morphology of the resulting material; nevertheless, the morphology of the material can be tuned from 0D particles up to 1D filaments (hollow or dense morphology) (Figure 5.9).

5.3.2.2 **Electrophoretic Deposition**

This technique is widely used in the deposition of both mesoporous and dense thin films from colloidal suspensions. The electrophoretic technique can be ideally combined with the sol–gel synthetic approach since in the latter charged species are the active components (colloids or charged stabilized sols). These can be moved in an external electric field [28–30]. In general, reduction or oxidation of the particles occurs at the electrodes, the initial deposition surface for the preceramic material. Combining sol–gel and electrophoretic synthesis techniques basically uses the (field) oriented motion of the charged sol colloids (electrophoretic motion), (Figure 5.10) [31]. To produce, for example, a 1D ceramic material, a structure-directing template is necessary. Porous alumina is an ideal host for this purpose and a number of 1D ceramics obtained by this technique have been reported (see Chapter 4).

As the deposition of the particles is from suspension, an additional post-deposition compaction or annealing step has to be performed in order to obtain, for example, a

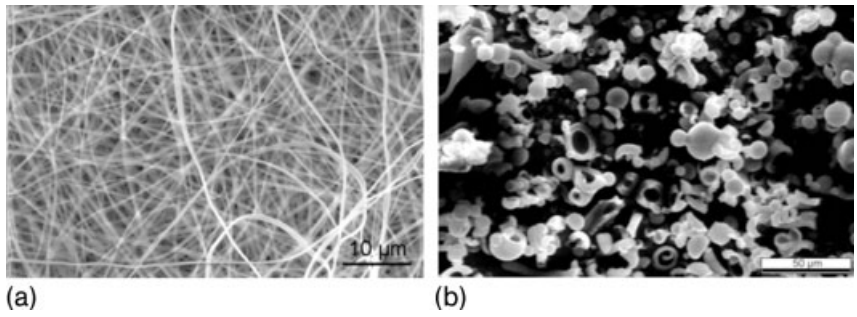


Figure 5.9 (a) Electrospun polymer fibers derived from Eudragit L100-55 (methacrylic acid–ethyl acrylate (1 : 1) copolymer. (b) Hollow spheres obtained via elctrospinning of aluminum *sec*-butylate in light petroleum (J.J. Schneider, J. Engstler, TU Darmstadt).

dense mesostructured ceramic. Since the as-prepared structure is already built from nanosized particles, this step can often be performed under smoother conditions as typically used in conventional ceramic processing [32].

5.3.3
Vapor–Liquid–Solid (VLS) and Related Synthesis Techniques

For the synthesis of 1D structures from the gas phase, vapor–liquid–solid (VLS) and vapor–solid (VS) processes are the typical growth mechanisms which are accepted to explain the 1D growth of mesoscopic structures [33]. In this growth process, a catalyst particle first melts, becomes saturated with a gaseous precursor and, when oversaturated, either an elemental wire or a compound wire depending on the precursor extrudes from this catalyst droplet to form a single-crystal nanowire. This is essentially the method proposed for whisker growth from the vapor [33]. Nanowires

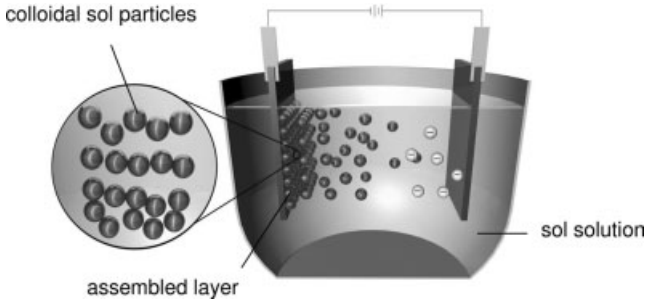


Figure 5.10 Schematic drawing of an experimental setup used for synthesis of mesostructured films by electrophoretic deposition. Particles are deposited from a colloid suspension to form a film on the surface of the electrodes. If the deposition is done in a porous template nanostructured rods are accessible. The compaction of the as prepared rods or films into a ceramic material follows in a post process.

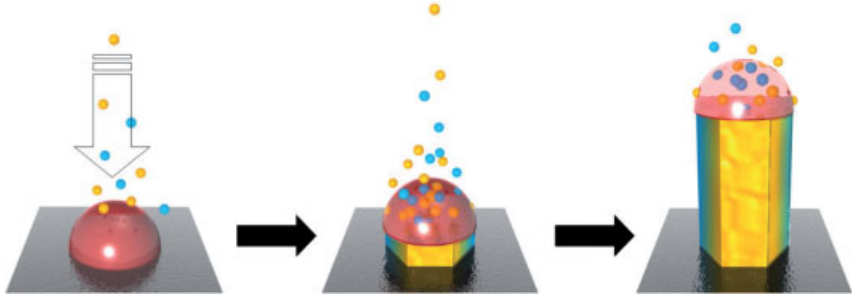


Figure 5.11 Growth mechanism for a pseudo-1D crystalline morphology. $V_{apor}L_{iquid}S_{olid}$ mechanism proposed by Wagner and Ellis for growth under CVD conditions.

grow as long as active catalyst is supplied and the growth temperature is maintained (Figure 5.11).

Recent model studies on the influence of various metal catalysts on the growth, structure morphology and size of inorganic nanowires have shown that different catalyst metals with different crystal morphologies are able to generate different wire morphologies based on the VLS formation process [34, 35]. For the growth of Si nanowires with Au nanocluster catalysts, it was found that the lowest energy surface, which is a {1 1 1} plane for Si, controls nucleation and growth (Figure 5.12) [34]. The results point towards the importance of an additional effect of oxygen in the gold-catalyzed growth kinetics of Si wires. The presence of oxygen can suppress Au catalyst

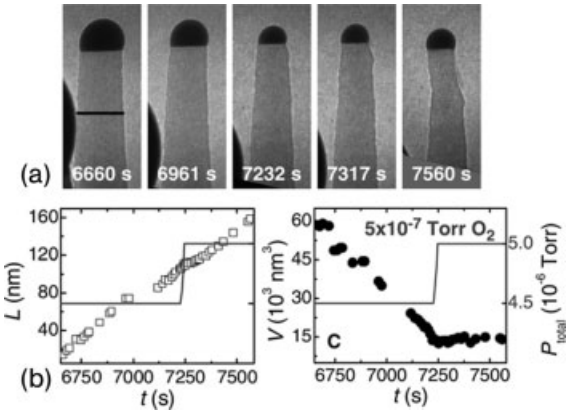


Figure 5.12 Effect of increasing oxygen pressure on Si nanowire growth kinetics. (a) Series of images extracted from a video sequence showing the effect of introducing oxygen to an Si wire that was previously growing in disilane (Si_2H_6). The growth was carried out in 4.5×10^{-6} torr disilane at 610 °C for 111 min; 5×10^{-7} torr oxygen was introduced while maintaining the disilane pressure constant. Scale bar is 50 nm. (b) Length *L* of the Si wire as a function of time *t*. (c) Volume *V* of the droplet as a function of time *t*. Reprinted with permission from Ref. [36].

migration and influences the wire diameter and the overall wire morphology [36]. The presence of oxygen, be it in the form of gas-phase oxygen or surface-bound oxygen, could be an important experimental parameter to modulate nanowire morphology further in a more general way during VLS growth [36].

For ZnO nanowires, it has been shown recently that control of the partial O_2 pressure under growth conditions is crucial for reproducible large-scale wire formation (Figure 5.13). It is very likely that the gas-phase conditions play a major role in the VLS growth of this and speculatively also other 1D materials made by this technique. Studies towards understanding the influence of reactive gas-phase species are central to deducing how individual morphologies of 1D materials depend on catalyst composition, shape and crystallinity in addition to the overall reaction conditions [37].

The current understanding of growth control of nanowires via the VLS technique is that (a) interplay given by the phase diagram of temperature–pressure and (b) the composition ofthe precursor elements and the catalyst particle under consideration are crucial. Careful control of these conditions can give rise to a variety of wire morphologies which are accessible at will once the conditions are thoroughly adjusted.

As shown, indicative for the VLS mechanism are catalyst tips at the faceted ends of the nanowires. Silicides are interesting semiconductors, with promising thermoelectric properties. Even though silicide wires (MSi_2, M = Fe, Co, Cr) were grown in the presence of nickel or iron [38], no catalyst metal particles are detected on the faceted nanowire ends as usually observed for the VLS growth mode. Obviously, in the chemical vapor transport technique (CVT) which is used for the synthesis of silicide wires a different gas-phase mechanism than in the most widely found catalyst-driven VLS growth may operate.

A promising technique related to the VLS Au seeded nanowire growth method has been reported. It allows oriented low-temperature gas-phase growth of pure Si nanowires [39]. Plasma growth at temperatures as low as 300 °C enables an additional degree of synthetic control over nanowire orientation (Figure 5.14). The crucial step in the formation process seems to be Si incorporation at the vapor–liquid interface.

Additional techniques also related to the traditional VLS-type growth mechanism are catalyst-driven growth from the solution/liquid phase into the solid state (SLS) and also from a supercritical liquid (fluid) solution (SLFS) into the solid state. The SLFS technique is synthetically complementary to the long-established technique of seed-mediated growth from solution, which yields one-dimensional colloidal metallic nanostructures [40]. In the former, a molecular precursor and in the latter a nanoparticle is decomposed under controlled conditions in solution or under supercritical conditions and then serves as a source for the metal catalyst particles from which the crystalline 1D structure grows (Figure 5.15) [41]. For this growth process from solution, low-melting metals as catalyst particles are essential (e.g., In, Bi, Sn). For metals with higher melting points (e.g., Au, Ge), supercritical solvent conditions have been successfully employed [41–43]. This synthetic technique allowed for the first time access to colloidal Au quantum wires which show a spectroscopic QSE.

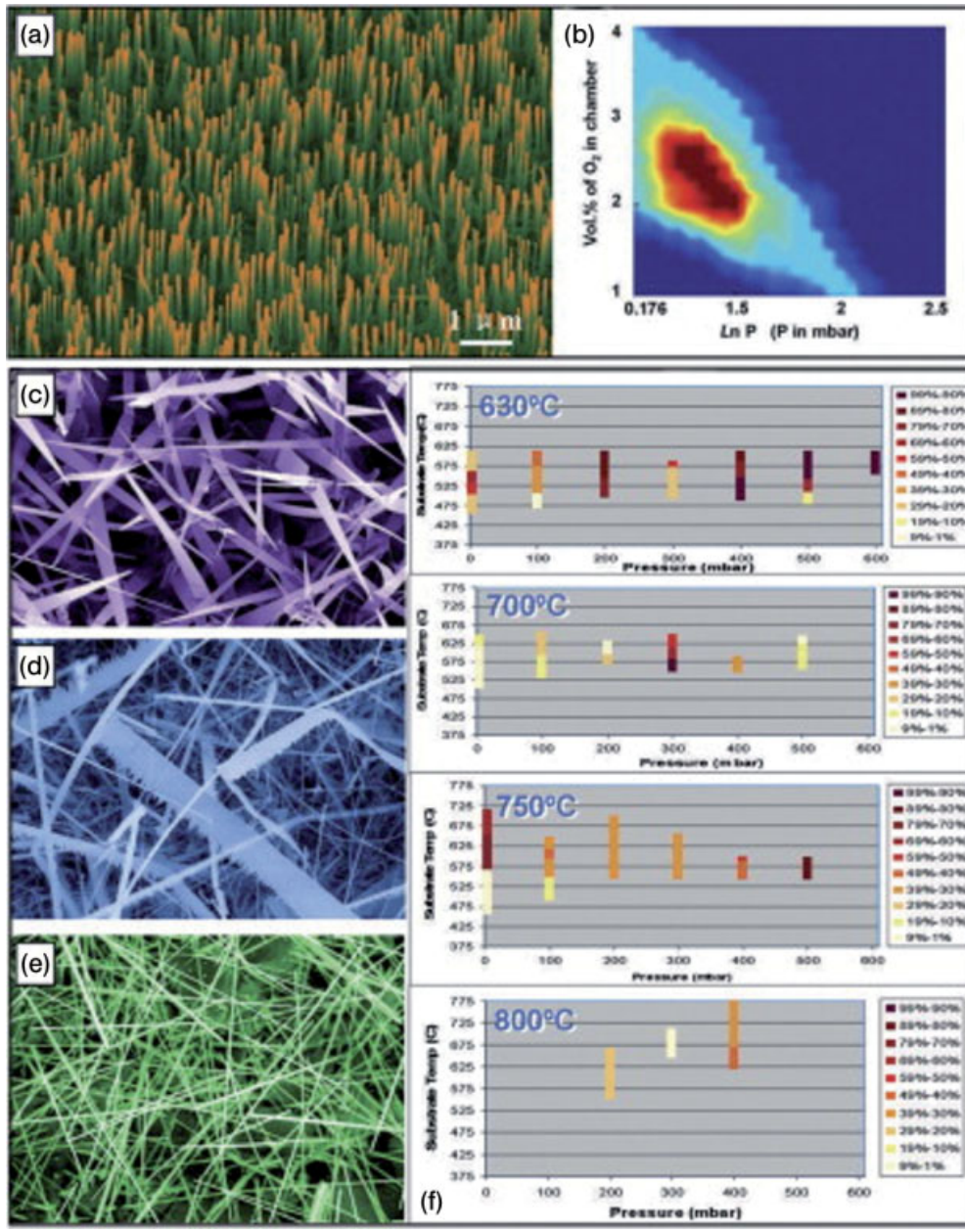


Figure 5.13 Aligned ZnO nanowires grown on a single-crystal alumina substrate with a honeycomb pattern which defined by the catalyst mask (a); (b) diagram showing the effect of oxygen partial pressure and total pressure in the growth chamber on the growth of ZnO NWs. Diversity of the morphology of possible ZnO nanomaterials: (c) ZnO nanobelts; (d) ZnO nanosaws; (e) nanowires of ZnO. (f) Yield of ZnO nanosaws as a function of reaction chamber pressure at four different furnace temperatures. The strength of the color indicates the percentage yield of the products. Reprinted with permission from Ref. [37b].

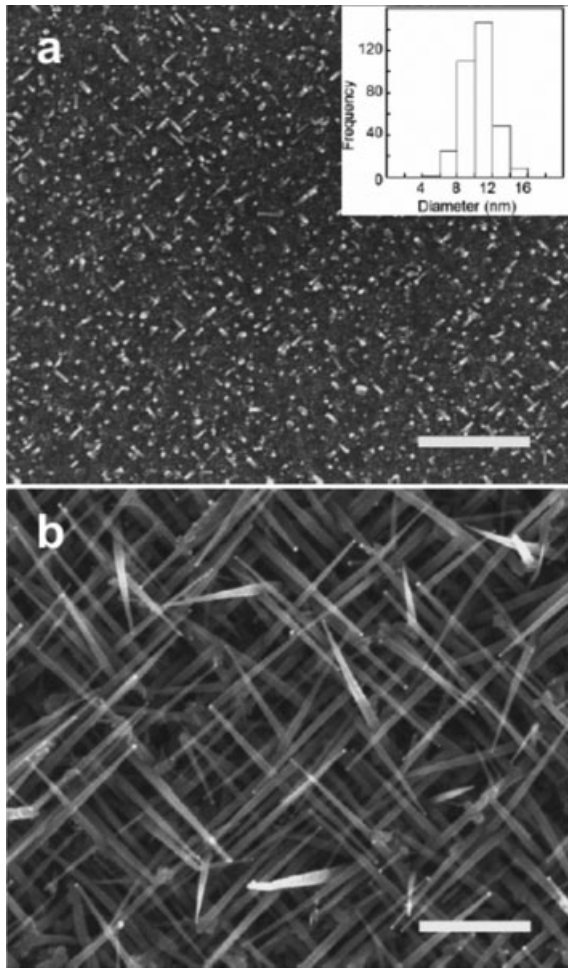


Figure 5.14 SEM images of Si nanowires grown on Si (100) substrates at 350 °C, 0.5 torr growth pressure for 15 min. (a) Thermal; (b) r.f. plasma enhanced conditions. All scale bars are 500 nm. Inset shows Au seed nanodot diameter distribution. Reprinted with permission from Ref. [39].

5.4
Contacting the Outer World: Nanowires and Nanotubes as Building Blocks in Nano/Micro/Macro-Integration

The key for applying 1D structures in future nanotechnological applications will be first assembly of nanostructures and then integration of such assembled nanostructures into existing micro-building blocks and their subsequent packaging into larger structures to allow for micro and macro manipulation of these integrated

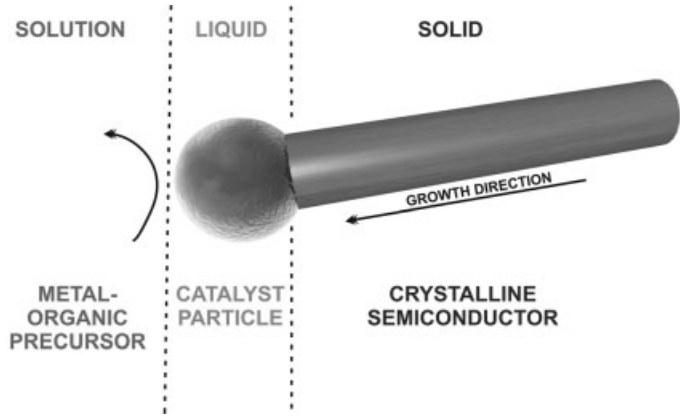


Figure 5.15 $S_{olution}L_{iquid}S_{olid}$ mechanism for nanowire growth from solution. Adapted from Ref. [41].

assemblies. This is especially challenging since a priori pertinent and incompatible length scales, nano vs. micro vs. macro, have to be bridged. Here the so called pitch problem is an important issue. It is related to the controlled formation of any nanowire structure into an array of individual wires packed into a defined order with definite distances between the individual objects.

For this purpose, techniques such as field-assisted assembly, fluid flow alignment and Langmuir–Blodgett techniques for placement and assembly of nanowires have been successfully studied and progress in these areas has been impressive in recent years. Precise positioning which allows the use of a huge number of 1D structures is no longer fiction [44–46]. Furthermore, the ability to develop even more powerful methods which have the potential to bridge that gap is still a major goal, especially for large-scale integration into electronic and sensing devices.

In 2003 it was still pointed out that "precise positioning of single-walled CNTs is beyond the capability of current growth and assembly technology and presents a major hurdle for CNT based electronic applications" [47]. In the following section selected examples of assembly and integration of 1D structures from different material classes will be highlighted to show current progress in that field.

5.4.1
Nanowire and Nanotube Sensors

Sensors based on a one-dimensional morphology (wire, tube) offer selective recognition for biological and chemical species of interest [48, 49]. This is based on their unique electronic and optical properties, which have been studied in detail either for isolated objects or in an unordered fashion for bundles of them. However, the control and use of well-arranged, aligned 1D structures have also made enormous progress within the last few years and have an impact on the applicability of such structures in sensor devices. The basis for this is the field effect transistor (FET) geometry, which

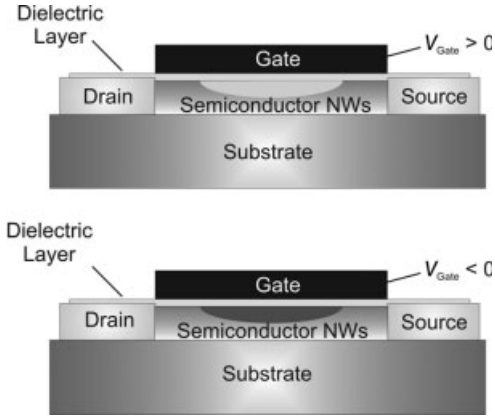


Figure 5.16 Schematics of field effect transistor (FET) geometry with nanowires as active channel material and different channel sensitivity. With a depletion of charge carriers conductance increases (upper FET schematic) When charge carriers are accumulated the conductance increases. For a p-type nanowire FET device, generation of positive charges on the nanowire's surface leads to a decrease in conductance. This can be induced, for example, by binding a protein to the wire's surface, which has a net positive charge in aqueous solution.

has been extensively explored as a device bridging the micro- and the nanoworld. In a FET device the contact between the source and drain electrode consists of a semiconductor. Its conductivity can be modulated by a third electrode, the gate coupled to the semiconductor by a dielectric layer through which charge is injected into the semiconductor. Since the binding of charged or polar molecules to the 1D semiconductor structure alters the channel electrode characteristics, this may lead to an accumulation or depletion of charge carriers and thus an increase or decrease in device conductance (Figure 5.16).

The idea was already put forth in the 1980s, but only for planar devices, which however, show only limited applicability for this sensor principle [50–52]. In a planar 2D film, only the near-surface region is altered by this effect, whereas in a single-crystalline rod surface, binding of an analyte has a strong impact on the depletion or accumulation of charge carriers in the overall structure of the nanoscale 1D object. The size of this effect depends, of course, on the size of the wire (which needs to be in the region of 2–5 nm) and its hybridization with a biomolecule (e.g., DNA). This can create a high charge density on the nanowire surface, which produces an electrostatic gating effect. This so-called field effect reduces the charge carrier concentration and results in an increase in resistance ($V_{gate} > 0$) (Figure 5.16).

With respect to CNT structures, recently the separation of metallic and semiconducting single-walled CNTs, albeit still in low quantities, has been achieved. Dielectrophoresis is the key to the nanoscale manipulation and separation of CNTs. An AC field induces polarization in a nanoscale object and this results in a force which can be used to manipulate and assemble nano-objects [53].

Dielectrophoresis in physiologically relevant saline solution has been performed and it has been shown that this technique is useful in manipulating nanowires across

electrode gaps in saline solutions. Individual carbon nanowires can be switched between bridging and unbridging events in microelectrode gaps. This may be important for, for example, contacting biomolecules, DNA recognition or protein binding, all of which typically operate in highly conducting aqueous saline solutions [54]. Also intriguing is the arrangement of semiconducting or metallic CNTs using electrophoresis techniques into highly integrated structures [55]. Altogether these findings these findings may pave the way to integrated sensor devices since they allow to use the whole individual tube length of the CNT as active element.

When dispersing commercial single-walled CNTs into an appropriate solvent with the aid of a dispensing polymer, printable and conductive inks have been obtained which allow the fabrication of robust, flexible, transparent (85%) and conductive (100 kΩ) single-walled CNT–polymer composite films on plastic substrates via inkjet printing. Their use as sensors for alcohol vapors under static and dynamic flow conditions has been reported [56]. Obviously no discrimination between metallic and semiconducting CNT can be made here. However, the approach's beauty lies in its simplicity of handling, storage and fabrication of mechanically stable, flexible and sensitive CNT sensor devices.

The precise alignment of single Si wires (diameter 20–30 nm) has been achieved by a combination of electrostatic positioning and conventional lithography (Figure 5.17) [57]. With this technique, the individual Si nanowire resistance and their contact resistance has been determined. It can be envisaged that this technique is capable of arranging 1D structures up to a high precision and opens up new avenues into sensing of single particles with individually arranged ensembles of nanowires.

Directed assembly of Si nanowires between two electrodes has been achieved by electrical field assembly out of a nanowire suspension. The aligned nanowires remained bound to the electrodes by van der Waals forces. So far arrays of 18 planar electrode pairs with micrometer spacings have been studied for nanowire alignment [58].

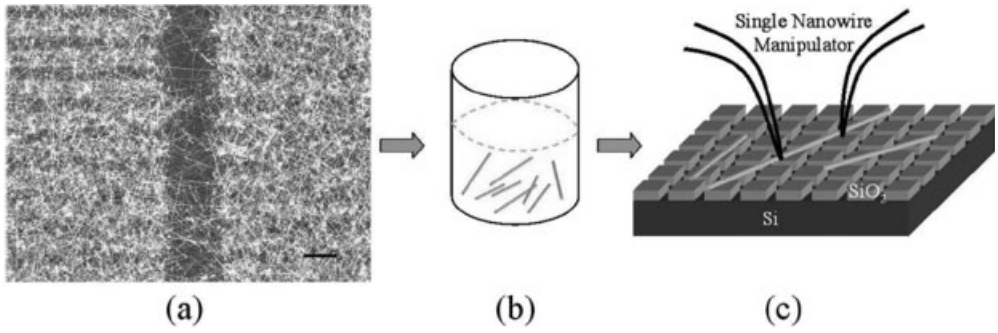


Figure 5.17 (a) SEM image of Si nanowires grown from patterned Au catalyst on a silicon wafer (VLS growth). Nanowires were removed into a suspension of deionized water. (b) A drop of the nanowire solution was dispersed on a template substrate and evaporated under vacuum. (c) Schematic of manipulator tips picking up a nanowire from a template substrate. The patterned structure of the template minimizes the adhesion forces of the wire.

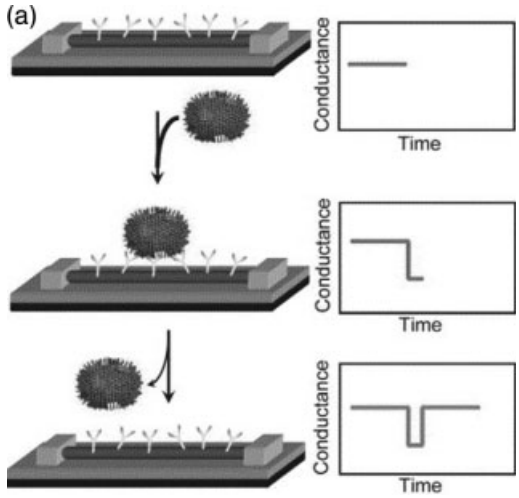


Figure 5.18 Schematic of a single virus binding and unbinding event at the surface of an Si nanowire device modified with antibody receptors and the corresponding time-dependent change in conductance. Reprinted with permission from Ref. [62].

The sensing mechanism in nanowires is due to a change in charge density on the wire's surface and probably deep into the "bulk" of the nanowire structure. This has been proposed to lead to ultrasensitive biosensors [59] and indeed was realized a few years later [60]. The upper sensing limitation of viruses as disease carriers and potential warfare agents with single-wire devices has been the subject of recent studies [61]. The working principle relies on the modification of a single wire, modified with antibody receptors (Figure 5.18) [62]. The corresponding time-dependent change in conductance is monitored. A real binding event monitored with an influenza A virus has been unraveled by measuring the change in the time-dependent conductance (Figure 5.19) [62].

With the use of the recently developed electrostatic positioning technique [57] for individual nanowires (see Figure 5.17), an integrated device configuration seems conceivable. Other successful approaches for assembling arrays of wires for multi-detection of analytes have been described [63]. Intriguing with respect to multiplexing methods for the detection of analytes is a transfer technique for arranging arrays of doped silicon wires on flexible plastic substrates. First, on to a prepatterned p-type doped Si substrate, highly aligned Si nanowires were prepared via anisotropic etching. Transfer on to plastic substrates was made possible by a polydimethylsiloxane (PDMS) stamping technique. Via selective chemical functionalization techniques of the Si nanowires using silanes, a sensor library with different end-group functionalization of the wires could be formed (Figure 5.20) [64]. This array finally worked as a highly sensitive integrated electronic nose for different vapors (Figure 5.21) [64].

The arrangement and contacting of up to 1 million cm^{-2} CNTs into individually contacted nanotube devices was achieved by dielectrophoretic deposition of metallic

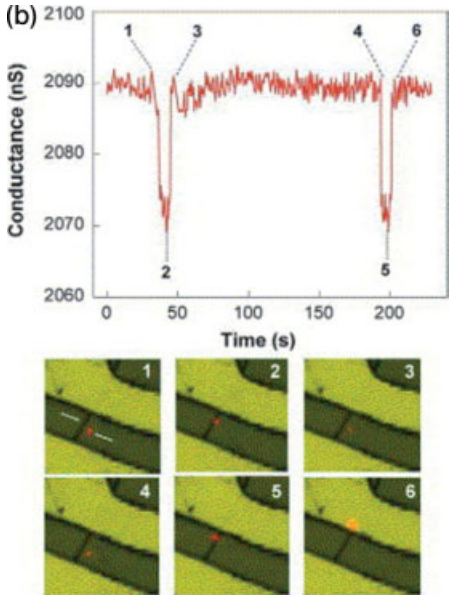


Figure 5.19 Simultaneous conductance and optical data recorded for an Si nanowire device after introduction of a influenza A virus solution. The images correspond to the two binding/unbinding events highlighted by time points 1–3 and 4–6 in the conductance data, with the virus appearing as a red dot in the images. Reprinted with permission from Ref. [62].

and semiconducting CNTs. The technique uses a field change occurring during nanotube deposition between individual electrodes of a preformed microstructured device (Figure 5.22) [55, 65]. Although not fully understood yet, redistribution of the electric field around the CNT in the gap seems to be important for the organization, rather than short-circuiting of the gap electrodes by the entrapped object [65].

In this approach, the overall dimensions of an individually contacted CNT device are only limited by the dimensions of the contacting electrodes rather than the

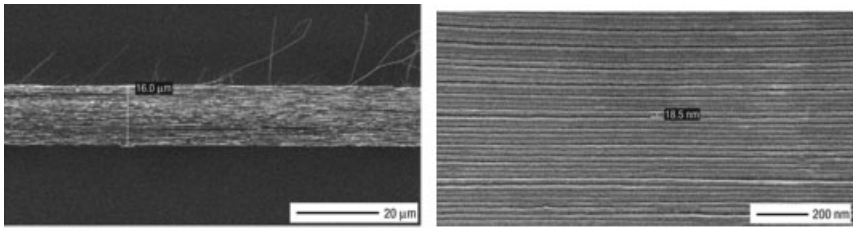


Figure 5.20 Superlattice nanowire pattern transfer film consisting of about 400 nanowires. High-magnification SEM image. The diameter of a typical nanowire is about 18.5 nm. Reprinted with permission from Ref. [64].

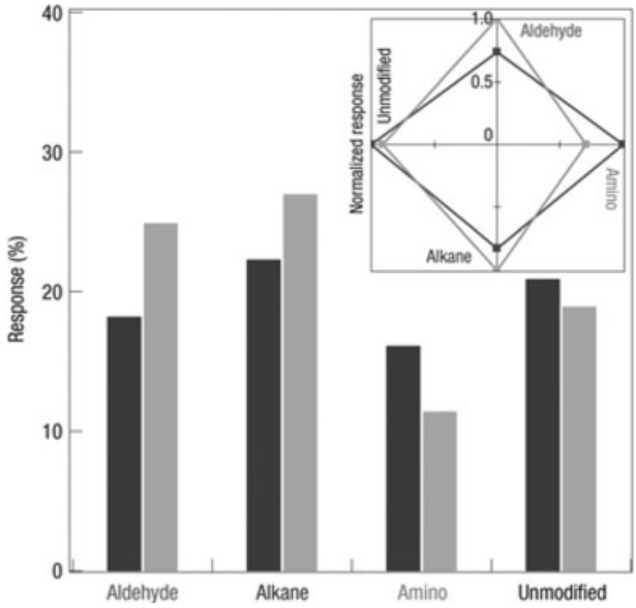


Figure 5.21 Sensor characteristics of a "nano-electronic nose". The bar graph summarizes the percentage change in response of the array to acetone (dark gray) and hexane (light gray) vapors. The inset shows the normalized response of the sensor library to acetone and hexane vapors. Each of the four axes represents the four unique surface functionalities. Reproduced with permission from Ref. [64].

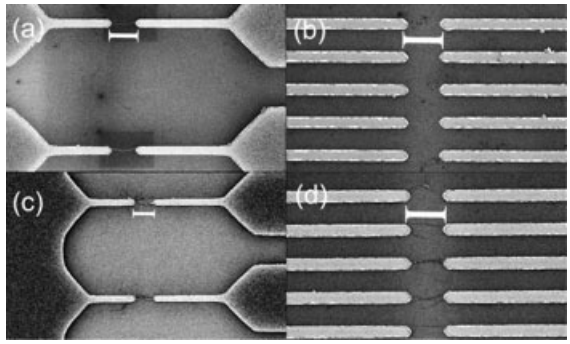


Figure 5.22 SEM images after nanotube deposition on chips with a designed electrode configuration. Effect of oxide thickness and counter electrode area of the chip on the directed assembly of nanotube devices, in terms of number of bridging nanotubes per electrode pair. Single nanotubes were assembled for 800-nm thick oxide layer and 10-μm^2 counter electrode area (a). No nanotubes were assembled for 800-nm oxide and 1-μm^2 electrode area (b). Only a few nanotubes were assembled for 50-nm oxide and 10-μm^2 electrode area (c). One or two nanotubes were assembled for 50-nm oxide and 1-μm^2 electrode area (d). Scale bars equal 1 μm in all images. Reprinted with permission from Ref. [55].

dimensions of the CNT itself. However, this seems to be the case for other approaches also [66]. So far the electrical device characteristics of over 100 CNTs have thus been measured individually. Only up to 10% of the electrodes were bridged by multiple CNTs [55].

Taking further into account the established routes for functionalizing CNTs [67], this approach seems viable for applying CNTs in a comparable way to that described above for Si wire arrays, for example, as sensor devices for the multiple detection of analytes. The field of immobilization of biomolecules is a sector of intense research activity [68]. Due to the electronic properties of individual CNTs (e.g., as semiconducting or metallic tubes), their sensing towards biomolecules could be very selective. Therefore, combining nanotubes with biosystems may provide access to nanosized biosensors. The first step for biomolecular interaction is the attachment of the biomolecule to the tube. This could be achieved either via a covalent bonding interaction or via a wrapping mechanism which uses non-covalent interactions, for example, on the basis of van der Waals interactions. The former needs functionalization of the CNTs followed by covalent bonding of the target biomolecule and is already well established. The latter rely on physisorption mechanisms and work typically for proteins.

For the development of covalent CNT functionalization techniques, a tool box of methods which have shown success in fullerene functionalization have been advantageously employed so far [67]. For example, for the case of functionalization of CNTs with bromomalonates containing thiol groups, attachment to gold surfaces is possible, allowing the synthesis of electrode arrays for sensor applications [69].

Chemical functionalization of CNTs has also paved the way to DNA–CNT adducts and subsequently to studies of their properties as biosensors. However, it seems that DNA attachment occurs mainly towards the end of the nanotubes [70]. This site-specific interaction points towards a sequence-specific poly(nucleic acid)–DNA base pairing rather than an unspecific interaction. This might be helpful for differentiation between two DNA sequences [68]. Pyrene functionalization of DNA itself can lead to a very specific DNA–CNT adduct. Its stability is mediated via hydrophobic interactions of the graphite-type sidewalls of the nanotubes and the pyrene anchors of the DNA [71].

DNA–CNT adducts are highly water soluble and therefore DNA functionalization of nanotubes avoids the use of surfactants, which are often used to solubilize CNTs alone. As with native DNA, the adducts with CNTs are still charged species. Specific DNA sequence detection with electrochemical CNT–DNA biosensors has been reported [72]. So far still a theoretical promise, the specific size-selective major groove binding mechanism of B-DNA towards single-walled CNTs could lead to a composite structure with unique sensor properties for ultrafast DNA sequencing or electronic switching based on the combination of the individual electronic properties of the single components CNT and DNA joined together in this composite structure (Figure 5.23). [73].

In this respect, it is intriguing to see the results that a strong binding capability of the major groove of DNA towards 0D nanoparticles has already been proven for gold nanoclusters, both theoretically and experimentally [74].

Figure 5.23 Theoretically proposed DNA–CNT composite structure. The single-wall CNTs are intercalated within the major groove of DNA. Adapted from Ref. [73].

5.4.2
Piezoelectrics Based on Nanowire Arrays

Converting mechanical energy into electric energy or signals is the area of piezoelectronics and relies on specific structures and morphologies of materials. Recently, nanomaterials have come into the focus of this application-driven area. ZnO is a material which probably exhibits the most diverse morphological configurations of any nanomaterial so far studied in detail. In addition to particles and wires there are nanobelts, nanosprings, nanorings, nanobows and nanohelices known and

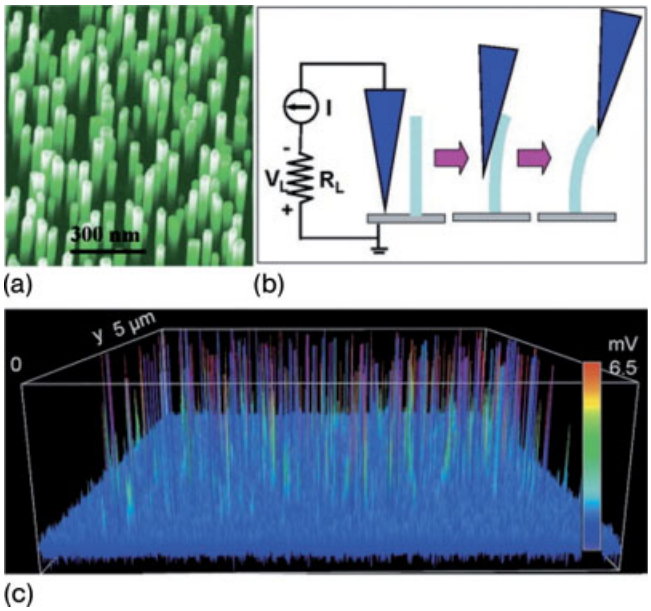


Figure 5.24 Scanning electron microscopy (SEM) images of aligned ZnO nanowires grown on an α-Al_2O_3 substrate (a). Experimental setup for generating electricity through the deformation of a semiconducting and piezoelectric nanowire using a conductive AFM tip. The root of the nanowire is grounded and an external load of $R_{load} = 500$ MW is applied, which is much larger than the inner resistance R_{inner} of the nanowire (b). The AFM tip is scanned across the nanowire array in contact mode. Output voltage image obtained when the AFM tip scans across the nanowire array (c). Reprinted with permission from Ref. [78].

characterized for ZnO [75]. Apart from use as a catalyst and sensor material which can be considered as more traditional areas for 1D nanoscale ZnO [76], the semiconducting and piezoelectric properties of ZnO nanowire arrays have recently been probed as piezoelectrics with unique properties. Such a ZnO-based nanogenerator converts mechanical energy into electric power and vice versa using massively aligned ZnO nanowires (Figure 5.24) [77, 78]. An electric field is created by deformation of a ZnO nanowire within the array by the piezoelectric effect. Across the top of the nanowire the potential distribution varies from negative at the strained side of the surface (V_s^-) to positive at the stretched surface (V_s^+). This potential difference is measured in millivolts and is due to the piezoelectric effect. On an atomic scale, the displacement of Zn^{2+} ions in the wurtzite lattice with respect to the O^{2-} counterions is responsible for that. The charges cannot move or combine without releasing strain via mechanical movement. As long as the potential difference is maintained (via constant deformation), the system is stable. When external free charges (e.g., via a metal tip) are induced, the wire is discharged. As an effect, the current which then flows is the result of the ΔV (V_s^+/V_s^-)-driven flow of electrons from the semiconducting ZnO wire to the metal tip. This flow of electrons neutralizes the ionic charges in the volume of the ZnO wire and reduces the potentials V_s^+ and V_s^- [78, 77].

The potential technological impact of such a nano-piezoelectronic effect might be in converting various type of energies (mechanical, vibration, hydraulic) into electrical energy. Thus extremely large deformations are possible, which are interesting for flexible electronics as a power source and realizing a much larger power/volume density output. The effect has already been shown to work for field effect transistor devices [79], piezoelectric gated diodes [80] and piezoelectric resonators [81], all on a ZnO material basis.

5.4.3 With Nanowires and Nanotubes to Macroelectronics

Using top-down techniques for scaling conventional microelectronic devices has by now reached the size regime of about 50 nm with a gate length of about 30 nm for a FET. This demonstrates impressively how far the conventional top-down technique using lithography, deposition techniques and etching methods have developed. Nevertheless, scaling down characteristic feature sizes and the channel gate length even further seems limited using these techniques.

On the other hand, despite tremendous success in recent years, incorporating nanosized materials into typical FET device architectures is still in its infancy. Exciting new developments have been reported over the last few years which might have the potential to bridge the nano–micro–macro gap for certain device architectures. Due to their aspect ratio, 1D materials can be produced in micrometer-long structures, but displaying nanometer size diameters. Hence they are *a priori* ideally suited to be incorporated in current microdevice technology.

Due to the possibility that a variety of semiconductor nanowires, including the workhorse element Si, show great promise to be processable from solution, a couple of areas of high technological input might appear on the horizon for such materials (e.g., flat panel displays or flexible solar cells). Especially when these nanomaterials are available as single-crystalline matter in 1D morphology, they show electronic performances even exceeding those of high-quality macro (bulk)-sized single crystals or thin films. Combining their high electronic performance with the possibility of arranging 1D inorganic nanowires over larger dimensions offers the ability finally to bridge the gap from nano- over micro- to macroelectronics and thus puts these materials one step closer to the realm of application.

5.4.3.1 Inorganic Nanowire and Nanotube Transistors

In general, inorganic semiconductors have the intrinsic advantage over organic-based semiconductors that the electronic mobility μ (electron–hole) usually drastically exceeds those of organic materials. The mobility μ is the proportionality constant between an applied electric field and the corresponding average charge carrier drift velocity. It is therefore a direct measure of the switching speed of an electronic device. The carrier mobility thus depends critically on the crystallinity of the material. In inorganic semiconductor wires of high crystallinity, a high order of atomic or molecular building blocks over long distances is maintained due to strong ionic or covalent bonds, whereas in organic semiconductors the mobility μ seems fundamentally

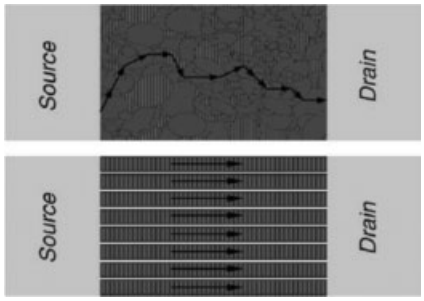


Figure 5.25 In polycrystalline silicon thin field effect transistors (TFTs), electrical carriers have to travel across multiple grain boundaries, resulting in low carrier mobility. Nanowire TFTs have conducting channels consisting of multiple single-crystal nanowires in parallel. Therein charges travel from source to drain within single crystals, ensuring high carrier mobility. Reprinted with permission from Ref. [82].

limited due to weak van der Waals bonds between electronically active molecular building blocks. This situation is a general one for an inorganic crystalline semiconductor compared with crystalline organic ones.

It is important for the exploitation of inorganic semiconductor nanowires to integrate them in massive arrangements in semiconductor devices, for example, thin-film transistors (TFTs), in which they might offer an increased electronic performance and future device integration. The former is due to the fact that drastically higher charge carrier mobilities μ can be obtained when using multiple single-crystal nanowires in FET devices instead of polycrystalline thin films (Figure 5.25) [82].

In a poly-Si TFT channel material, the electrical carriers have to travel across multiple grain boundaries (curved pathway), whereas in a massive parallel-arranged single-crystal nanowire array, charge carriers travel from source to drain within a single-crystalline structure, which ensures a high carrier mobility μ. The same effect has been observed, for example, for compound semiconductors such as ZnO. First sintering of isolated ZnO 0D nanoparticles has to be employed to obtain coarse grain polycrystalline thin films (Figure 5.26) [83]. When comparing the electronic performance of polycrystalline ZnO thin films with that of single-crystalline ZnO nanowires deposited between the source and drain of a FET device, the performance of ZnO nanowires is intriguing (Table 5.1).

Calculations have shown that for a channel length (geometric source to drain distance) of 20 μm in a FET device, the number of grain boundary contacts is diminished by a factor of 6 compared with a polycrystalline ZnO semiconductor electrode against one composed of a 1D wire morphology. In addition, the reduced hoping frequency for the charge carriers across the grain boundaries in the 1D ZnO structures compared with the polycrystalline ZnO thin-film morphology adds towards their increased mobility (Figure 5.27) [85]. One may imagine how this can

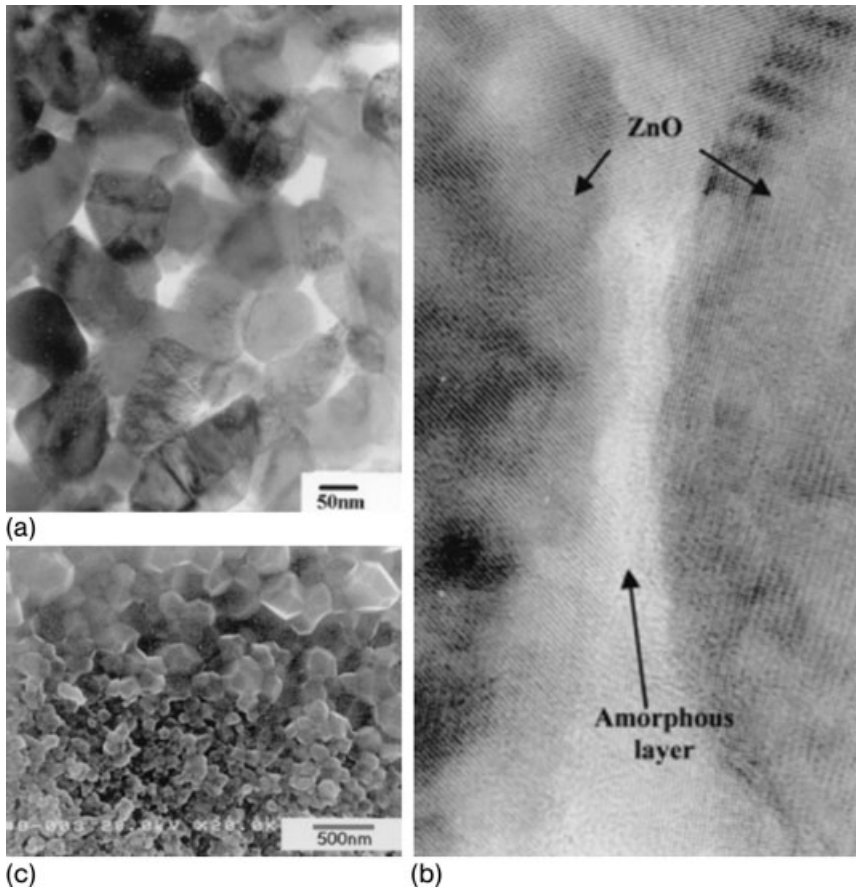


Figure 5.26 Transmission electron microscope (TEM) image of a nanocrystalline ZnO film. Processing temperature 550 °C (a); TEM of the grain boundary of two ZnO nano particles (b), Processing temperature 600 °C, sintered ZnO nanostructures can be observed (c). Reprinted with permission from Ref. [84].

Table 5.1 Charge carrier mobilities μ and on/off ratios of FET architectures with ZnO as active material (2D thin films vs. 1D rods).

Material	Morphology (diameter) (nm)	Mobility μ ($cm^2 V^{-1} s^{-1}$)	On/off ratio	Ref.
Si	Rods 20 or 40	119	10^8	92
Si	Particles 300	50–100; 6.5 (printed)	100	84a
Si	Rods 2000	180	1000	89
ZnO	Particles	0.23	1.6×10^5	84b
ZnO	Particles 6	0.00023	5×10^3	85
	Rods 10	0.023	1×10^5	85

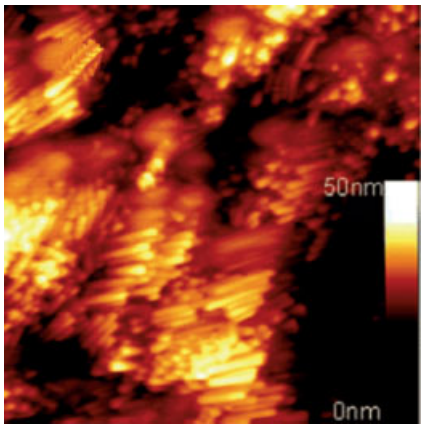


Figure 5.27 Atomic force microscope (AFM) picture of ZnO nanorods, length 65 nm, diameter 10 nm, deposited between source and drain of a FET device structure. Reprinted with permission from Ref. [85].

be further improved if it were possible to arrange such wires in a massive parallel fashion.

Dramatically higher carrier mobilities for well-aligned single-walled CNTs compared with randomly oriented networks of CNTs [86] have been measured [87]. The CNT alignment was achieved along the $[21\bar{0}]$ plane of a right-handed α-quartz substrate. The process is attractive since Y-cut α-quartz is a commercial substrate. The tube orientation is so far not fully understood, but is directed along step edges and/or micro/nanofacets on the surface of the quartz (Figure 5.28). Device mobilities of up to 125 $cm^2 V s^{-1}$ have been measured, which correspond to individual tube

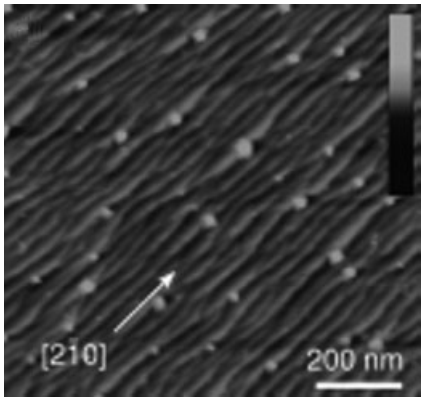


Figure 5.28 AFM image of terraced quartz surface structures after thermal annealing. The steps are 0.7–1 nm in height with spacing 30–35 nm. The dots seen are ferritin-derived Fe catalyst particles. Scale bar, 5 nm height. Reprinted with permission from Ref. [87].

mobilities within the array comparable to those already measured in pristine single tubes. These results clearly indicate that alignment of 1D nanostructures is crucial for the most effective device integration of CNTs [87].

The channel conductivity for nanowire and nanotube transistor devices has been treated theoretically and a universal analytical description has been developed. It has been found that the transconductance of the channel differs from classical device theory because of the specific nanowire charge distribution. Mainly different electrostatics for the 1D channel structure are responsible for the different device characteristics [88].

Due to the microscale length dimensions in which many semiconductor nanowire materials are obtainable, large-area substrates such as standard FET device structures can already be applied in device fabrication using 1D nanomaterials. An intriguing example is the arrangement of crystalline, several tens of micrometer long Si nanowires, which have been arranged between the source and drain of a FET device from a solution-derived process (Figure 5.29) [84]. Therein growth and integration of the particular nanowire material are separated from the device fabrication. Wire synthesis is done via a CVD approach, which of course is not compatible with the device fabrication process. However, the wires can be solution processed and in a second follow-up step deposited in a highly aligned fashion. The two-step route is applicable to sensitive FET substrates such as plastics and points towards a general route for future integration of nanowires into micro–macro integrated device architectures. Always a single-crystalline Si wire connects the source and drain; the distance between such individual wires is 500–1000 nm. This assures that

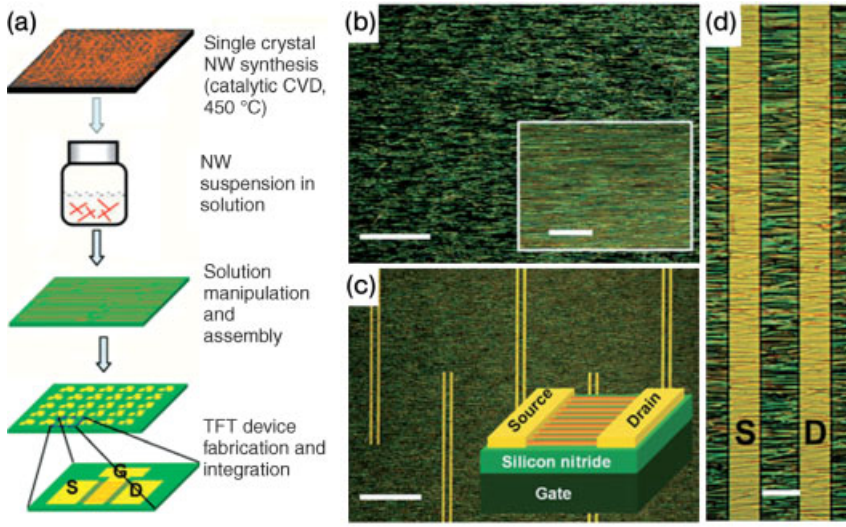


Figure 5.29 Scheme of parallel aligned crystalline Si nanowires between source and drain structure of a FET (b–d). Distance between individual wires is 500–1000 nm and can be adjusted via the solution deposition process (a). Reprinted with permission from Ref. [89c].

the wires have no contact. Extraordinary high charge carrier mobilities have been realized [82, 89].

The idea of separating the two individual processes, (i) material synthesis and (ii) materials processing, into device architectures has been used to arrange semiconductor nanowires on flexible polymer-based substrates. Seminal to the development of this field of printed inorganic devices was probably the report of the first printed all-inorganic thin-film field effect transistor based on 0D CdSe nanoparticles [90].

Synthesis of semiconductor rods or wires was performed, for example, via a CVD process on a silicon wafer. First the wafer-based source material was structured via a lithographic etching procedure. This technique works for a variety of 1D materials such as single-walled CNTs, GaN, GaAs and Si wires [84a]. An independent dry transfer process of the previously synthesized nanowires using an elastomeric stamp-based printing technique involves transfer of the 1D structures to a flexible substrate, for example, polyimide. It was thus possible to fabricate transistor devices with the highest electronic performance and also very high mechanical flexibility (Figure 5.30).

A general applicable condensed-phase approach for the alignment of nanowires uses the Langmuir–Blodgett technique [91]. It allows a massive parallel arrangement on planar substrates. The technique can be used in a layer-by-layer process allowing crossed nanowire structures with defined pitch to be formed (Figures 5.31 and 5.32). Moreover, this approach may allow the use of nanowires of different composition within the layering process. This gives rise to organized nanowire heterostructures.

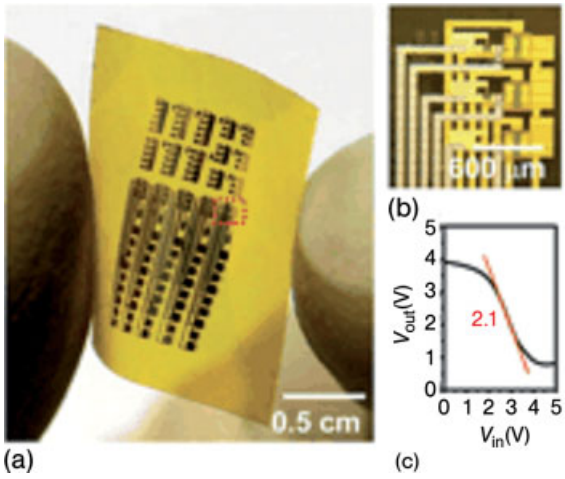


Figure 5.30 (a) Image of a printed array of 3D silicon n-channel metal oxide semiconductor inverters on a polyimide substrate. The inverters consist of MOSFETs (channel lengths of 4 μm, load-to-driver width ratio of 6.7 and a driver width of 200 μm) on two different levels. (b) View of the region indicated by the red box in (a). (c) Transfer characteristics of a typical inverter. Reprinted with permission from Ref. [89b].

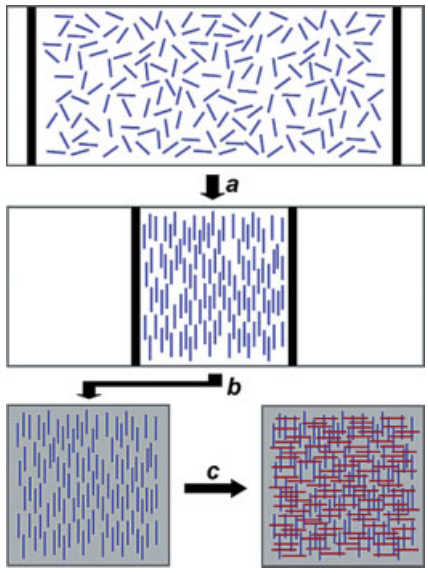


Figure 5.31 Nanowires (blue lines) in a monolayer of surfactant at the air–water interface are (a) compressed on a Langmuir–Blodgett trough to a specified pitch. (b) The aligned nanowires are transferred to the surface of a substrate to make a uniform parallel array. (c) Crossed NW structures are formed by uniform transfer of a second layer of aligned parallel nanowires (red lines) perpendicular to the first layer (blue lines). Reprinted with permission from Ref. [91].

Due to the hierarchical scaling of the structures from the nanometer up to the micrometer regime, reliable electrical contacting is possible.

Molecular precursors for the synthesis of polycrystalline silicon wires represent a promising alternative route to the widely employed synthetic CVD technique to inorganic semiconductor materials. Known since the early days of organosilicon chemistry, hydrogenated silicon compounds are known as chain (Si_nH_{2n+2}) or ring molecules (Si_nH_{2n}). For $n \geq 3$, these molecules are liquids and decompose around

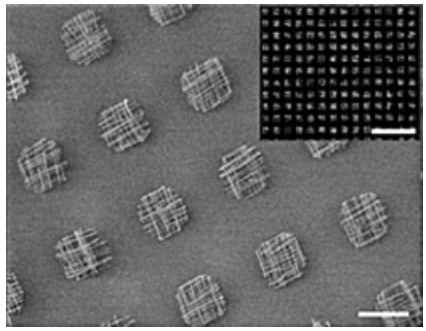


Figure 5.32 SEM image of patterned crossed nanowire arrays; scale bar 10 μm. Inset: Large area dark-field optical micrograph of the patterned crossed nanowire arrays; scale bar 100 μm. Reprinted with permission from Ref. [91].

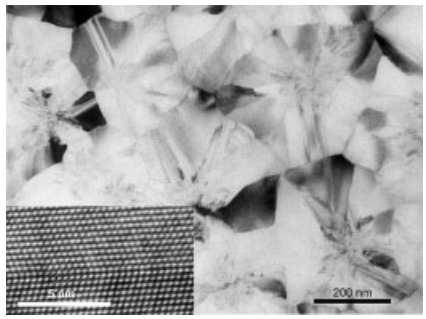


Figure 5.33 The polycrystalline Si thin film was formed by spin-coating and baking of the liquid single source Si_5H_{10} precursor followed by laser crystallization. The TEM inset picture highlights the atomic image of the silicon crystal. Grain size in the film is about 300 nm, which is comparable to that of conventional CVD-formed poly-Si film. Reprinted with permission from Ref. [92].

300 °C to give Si. Using cyclopentasilane Si_5H_{10}, thin films of this precursor have been solution processed and converted to polycrystalline Si (Figure 5.33) [92].

This route is compatible with inkjet printing and has been used to set up a complete TFT device architecture via printing. Charge carrier mobilities of 6.5 $cm^2 V^{-1} s^{-1}$ have been reported [92]. Other FET device architectures have also been studied using nanowires as active materials. FETs with surrounding gates have been synthesized and their electrical performance studied for Si [93]. Similar structures have been fabricated for InAs as active semiconductor. InAs performs with a high electron mobility and tends to form ideal ohmic contacts to many metals (Figure 5.34) [94]. Such structures are expected to show enhanced transconductance along the wire [95].

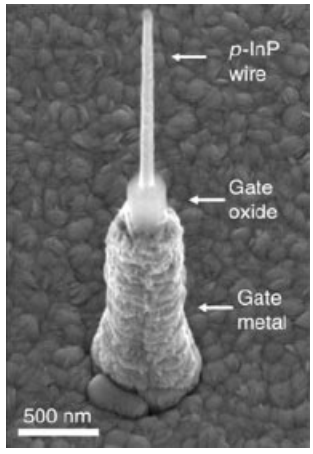


Figure 5.34 SEM micrograph of a vertical device with surrounding gate structure consisting of a p-type InP wire covered with gate oxide and gate metal. Reprinted with permission from Ref. [94].

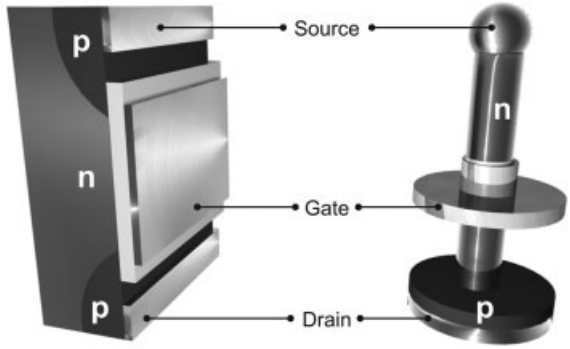


Figure 5.35 Schematics of a conventional p-channel MOSFET and a silicon nanowire surround gate FET. Adapted from Ref. [96].

A complete flow process for the fabrication of a silicon nanowire array with vertical surround gate FETs has been devised (Figure 5.35) [96]. Such an architecture was proposed earlier for CNTs but inorganic nanowires have the advantage that they can be grown as mechanically stiff vertical objects, whereas this is more complicated with single-walled CNTs [97].

Surround gate-type FETs have also been reported for Ge nanowire arrays. First, the nanowire array with the surround gate shell structure is synthesized via a multistep CVD process, followed by patterning of the nanosized core shell $Ge/Al_2O_3/Al$ from solution on to an Si substrate. Finally, the structure is electrically contacted. This leads to a macroscopic device with a multiple number of surround gate nanowires arranged in a quasi-parallel fashion (Figure 5.36) [98].

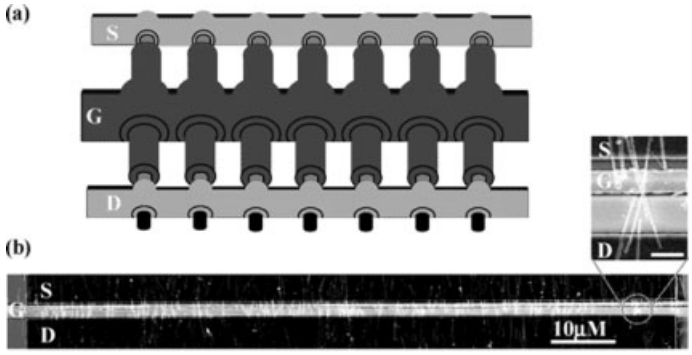


Figure 5.36 Transistor comprised of multiple surround-gate nanowires in parallel. (a) An idealized schematic presentation of a device. (b) SEM image of a device with ~35 surround gate nanowires in parallel. Crossing wires (each with its own gate shell) are seen in the zoomed-in image (scale bar = 1 μm). Reprinted with permission from Ref. [98].

5.4.3.2 Branched Nanowire Structures

Growth of 1D nanowires into certain defined directions leads to branched nanowire structures. In these, a higher degree of complexity is reached than in isolated wires. For example, the number of connection points and interconnecting possibilities to other nanostructures are enhanced compared with classical 1D structures. Their potential as interconnects for nanoelectronics, for example, circuiting, is obvious and certainly drives research into that area to some extent. One prerequisite for forming branched nanowire structures is the wide ability of semiconductors to undergo polymorphism, its characteristic feature to exist in more than one stable crystal structure.

Due to their nanosized dimensions, efficient strain release due to lattice mismatch of element combinations is much more efficient in nanowire architectures than in bulk film heterostructures of semiconductors. Therefore, lattice mismatches of $>3\%$ are not detrimental for a sharp interface growth process of wires [99]. In recent years, the scope of materials within the area of branched nanowire structures has broadened significantly. Condensed-phase (solution) and gas-phase synthetic techniques, the latter based on both physical vapor and chemical vapor deposition, have been used successfully to generate branched structures for many elemental combinations [100]. Compositions such as CdSe, PdSe, CdS, MnS and CdTe have been made available through solution-based techniques [101]. A successful technique to initiate branching of individual nanorods is to add nanoparticles as new growth sites. This has been demonstrated for GaP, InP, Ga As, AlN and GaN [102]. Successful studies even towards the hierarchical growth of interconnected nanobranched structures have been undertaken and reviewed recently [103]. Again, ZnO is a material which has shown an enormous breadth of structures with branched characteristics [104]. The growth of ZnO tetrapods is well understood and may serve as a model system to illustrate some general points of the growth of branched nanostructures. Although several models have been discussed and differ in some detail, it is accepted that growth of the cylindrical arms of a tetrapod proceeds via nucleation of a core structure. As determined by detailed high-resolution transmission electron microscopy studies, a zinc blende structure represents the core structure from which wurtzite arms grow. The former thus serve as seed templates for wurtzite arm growth. Growth of the $[00\bar{1}]$ wurtzite facets is outwards from the four [111] crystal faces of the zinc blende seeds. The same growth mode has been discussed for CdSe and CdTe tetrapod architectures [105]. Although the synthesis and growth of ZnO tetrapods is well studied, the electronic properties of this architecture have only recently been addressed.

Making electrical contacts to the edges of the tetrapod nanocrystal is, of course, difficult. A diode configuration has been realized with rectification characteristics (W forms an ohmic contact, Pt forms a Schottky contact to the tetrapod) (Figure 5.37) [106].

Branching out of nanowires has been achieved by using Au particles as catalyst seeds, resulting in the formation of Group III and V compound semiconductor rods (Figure 5.38).

Au catalyst particles are prepared via a gas-phase aerosol process and are size selected using a differential mobility analyzer [107]. The wire growth process is

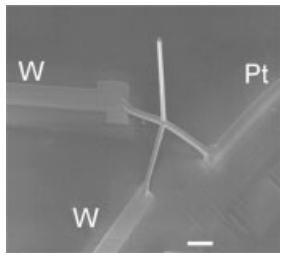


Figure 5.37 SEM image of a ZnO tetrapod with two W contacts and one Pt contact made to each arm of the structure. Scale bar = 1 μm. Reprinted with permission from Ref. [106].

repeated on the initially formed nanowire trunks to yield finally the branch structure in a follow-up process step. In this second process step, individual synthesis conditions, for example, the molecular wire precursor, can be varied, giving rise to branched nanowire heterostructures. The individual nanowire diameter of the branches is determined by the Au catalyst particle size; however, additional deposition of material on the side facets of the wires leads to thickening of the wires, and a tapered structure is therefore often observed [108].

The hierarchical growth of already branched nanowire structures to higher organized ensembles can also be realized in a solution-based growth process [105]. The precursor (e.g., an elemental chalcogenide) and a capping agent [e.g., an alkyl

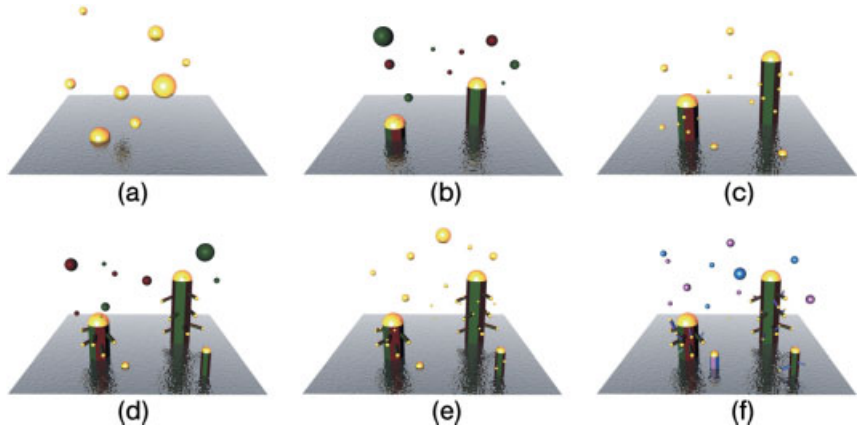


Figure 5.38 Branched nanowire and hetero nanowire formation process. Au particles are deposited on a substrate (a); precursor molecules (arrows) are introduced at elevated temperatures and combine beneath seed particles to form compound nanowire trunks (b); a second set of Au particles is deposited onto these trunks (c); branched nanowires grow in the same way as for (b) in (d); a third set of Au particles is introduced (e), followed by a new set of precursor molecules, which results in the growth of branched hetero nanowires (f). Adapted from Ref. [108].

(aryl)phosphine] first initiates the nanorod growth, followed by branching due to the facetted growth mechanism into a nanotetrapod structure. An end selective extension and further branching of such nanotetrapods are then possible by adding a new precursor to the same mixture. However, removal of the nano-objects from the initial synthesis mixture of the grown nanostructures, redispersion and regrowth, for example, with a different precursor, are also possible and allow further hierarchical growth at the ends of the original tetrapod structures. This growth gives rise to the formation of branched heterostructures. Thus, interfaces with different compositions can be designed in these structures, allowing control over charge carriers (electrons or holes) due to tailoring of the interface.

A detailed catalyst-driven (via Au aerosol particles) growth study of branched heterostructures by combining Group III–V materials, GaAs, GaP, InP and AlAs, has led to the conclusion that the growth mechanisms of such heterostructures depend on the relationship of the interface energies between the growing materials and the catalyst particle. It turned out that the growth of straight heterostructures in a particular direction seems favored over growth in another direction.

Finally, comparing gas-phase techniques and solution-based routes to branched nanowire structures, both methods show great potential towards the synthesis of higher organized morphologies based on 1D wires. It is remarkable that in both synthetic approaches the growth can be initiated on already formed branched structures just by adding fresh catalyst particles. However, the solution-based methods surely are favored due to the relative ease of formation of such structures and their further assembly potential towards higher organized multibranched wire structures. One further step towards this end has been demonstrated by tipping the ends of individual CdSe rods or tetrapods on both sides with Au nanoparticles, in solution (Figure 5.39) [109].

Although already a truly composite structure on its own (e.g., the excitonic spectra of the CdSe nanorod structure and the plasmonic Au spectra are not a superposition of the individual features of the individual semiconductor material and the metal nanoparticle), functionalization of the nanoparticle ends with alkanedithiols has led to assembled lines of individual composite nanowires (Figure 5.39) [109].

Figure 5.39 Schematics of dumbbell-like structures of a CdSe nanorod with Au gold tips at both ends. Single CdSe dumbbell doubly tipped with Au nanoparticles (a); typical unordered arrangement of Au tipped CdSe nanorods, where the Au tip size can be varied by increasing the $AuCl_3$ concentration during tip growth (b); self-assembled chain of nano CdSe dumbbells (real size 29 × 4 nm) formed by adding hexanedithiol bifunctional linker molecules which connect Au tipped ends of CdSe dumbbells. Adapted from Ref. [110].

5.5 Outlook

Studies towards functionality in inorganic 1D materials are an interesting and interdisciplinary field, bringing together fundamental science and opening up opportunities towards possible applications of these materials in various areas. Directed, organized growth of such 1D objects is surely a key in most of the envisioned areas of application of nanoscience.

In recent years, it has become clear that in addition to a bottom-up synthetic approach to 1D nanomaterials, a top-down approach which allows structuring, assembly and integration of 1D nanomaterials on the next higher length scale is necessary to organize and use 1D materials as building blocks in functional devices. The question no longer seems to be which of the two techniques, bottom-up or top-down, is the more powerful approach to new functional devices, but rather how we can make use in the most efficient way of both technologies to bridge the dimension gap: nano–micro–macro. In areas where current microtechnologies and materials are best compatible with bottom-up techniques for the synthesis and handling of 1D nanomaterials, they already work hand in hand and the unique properties of, for example, 1D materials have to be exploited successfully. This has already been demonstrated in current electronic and optoelectronic devices such as field effect transistors, light-emitting diodes, gas sensors and nanoresonators [110]. A key to this development of the field is certainly strong interdisciplinary research efforts between chemists, physicists and engineers.

References

1 (a) Rao, C.N.R. and Govindaraj, A. (2005) *Nanotubes and Nanowires*, Royal Society of Chemistry, Cambridge; (b) Yang, P. and Poeppelmeier, K.R. (2006) Inorganic Chemistry Forum: Special Issue on Nanowires. *Inorganic Chemistry*, **45**, 7509–7510; (c) Tenne, R. (2006) *Nature Nanotechnology*, **1**, 103–111; (d) Murphy, C.J., Gole, A.M., Hunyadi, S.E. and Orendorff, C.J. (2006) *Inorganic Chemistry*, **45**, 7544–7554; (e) Kline, T.R., Tan, M., Wang, J., Sen, A., Chan, M.W.H. and Mallouk, T.E. (2006) *Inorganic Chemistry*, **45**, 7555–7565; (f) Xiang, X., Yang, P., Sun, Y., Wu, Y., Mayers, B., Gates, E., Yin, Y., Kim, F. and Yan, H. (2003) *Advanced Materials*, **15**, 353–389; (g) Goldberger, J., Fan, R. and Yang, P. (2006) *Accounts of Chemical Research*, **39**, 239–248; (h) Pokropivnyi, V.V. (2001) *Powder Metallurgy and Metal Ceramics*, **40**, 485–496; (i) Pokropivnyi, V.V. (2001) *Powder Metallurgy and Metal Ceramics*, **40**, 582–594; (j) Remškar, M. and Mrzel, A. (2003) *Science Direct*, **71**, 177–183; (k) Tenne, R. (2006) *Journal of Materials Research*, **21**, 2726–2743; (l) Rao, C.N.R. and Nath, M. (2003) *Dalton Transactions*, 1–24; (m) Remškar, M. (2004) *Advanced Materials*, **16**, 1497–1504; (n) Monthioux, M., Flahaut, E. and Cleuziou, J.-P. (2006) *Journal of Materials Research*, **21**, 2774–2793; (o) Yan, Y., Chan-Park, M.B. and Zhang, Q. (2007) *Small*, **3**, 24–42; (p) Tenne, R., *Angewandte Chemie*, **2003**, **115**, 5280–5289; (2003) *Angewandte Chemie-International Edition*, **42**, 5124–5132; (q) Patzke, G.R., Krumeich, F. and Nesper, R.

(2002) *Angewandte Chemie*, **114**, 2554–2571; (2002) *Angewandte Chemie-International Edition*, **42**, 2446–2461.

2 (a) Beck, J.S., Vartulli, J.C., Kennedy, G.J., Kresge, C.T., Roth, W.J. and Schramm, S.E. (1994) *Chemistry of Materials*, **6**, 1816–1821; (b) Hamley, I.W. (2000) *Introduction to Soft Matter*, Wiley, Chichester; (c) Texter, J. (ed.) (2001) *Reactions and Synthesis in Surfactant Systems*, Vol. 100, Marcel Dekker, New York; (d) Förster, S. and Antonietti, M. (1998) *Advanced Materials*, **10**, 195–217; (e) John, V.T., Simmons, B., McPherson, G.L. and Bose, A. (2002) *Current Opinion in Colloid and Interface Science*, **7**, 288–295; (f) Dabbs, D.M. and Aksay, I.A. (2000) *Annual Review of Physical Chemistry*, **51**, 601–622; (f) Harrison, W.T.A. (2002) *Current Opinion in Solid State & Materials Science*, **6**, 407–413; (g) Liang, Z. and Susha, A.S. (2004) *Chemistry – A European Journal*, **10**, 4910–4914.

3 Ma, D.D.D., Lee, C.S., Au, F.C.K., Tong, S.Y. and Lee, S.T. (2003) *Science*, **299**, 1874–1877.

4 Pokropivnyi, V.V. (2002) *Powder Metallurgy and Metal Ceramics*, **41**, 123–135.

5 (a) Hicks, L.D. and Dresselhaus, M.S. (1993) *Physical Review B-Condensed Matter*, **47**, 16631–16634; (b) Hicks, L.D. and Dresselhaus, M.S. (1993) *Physical Review B-Condensed Matter*, **47**, 12727–12731.

6 Heinzel, T. (2007) *Mesoscopic Electronics in Solid State Nanostructures*, 2nd edn, Wiley-VCH, Weinheim.

7 (a) Xiong, Y., Mayers, B.T. and Xia, Y. (2005) *Chemical Communications)*, 5013–5022; (b) Pokropivnyi, V.V. (2001) *Powder Metallurgy and Metal Ceramics*, **40**, 485–496; (c) Ivanaovskii, A.L. (2002) *Russian Chemical Review*, **71**, 175–194; (d) Tenne, R. (2006) *Journal of Materials Research*, **21**, 2726–2743; (e) Remskar, M. and Mrzel, A. (2003) *Vacuum*, **71**, 177–183; (e) Roveri, N., Falini, G., Foresti, E., Fracasso, G., Lesci, I.G. and Sabatino, P. (2006) *Journal of Materials Research*, **21**, 2711–2725.

8 Lohrengel, M.M. (1993) *Materials Science and Engineering*, **R11** (6), 243–294.

9 (a) Keller, F., Hunter, M.S. and Robinson, D.L. (1953) *Journal of the Electrochemical Society*, **100**, 411–419; (b) Wood, G.C. and O'Sullivan, J.P. (1970) *Proceedings of the Royal Society of London. Series A*, **317**, 511–543; (c) Ono, S. and Masuko, N. (1992) *Corrosion Science*, **33**, 503–505; (d) Diggle, J.W., Downie, T.C. and Goulding, C.W. (1969) *Chemical Reviews*, **69**, 365–405; (e) Kniep, R., Lamparter, P. and Steeb, S. (1989) *Advanced Materials*, **1**, 229–231; (f) Hoar, T.P. and Yahalom, J. (1963) *Journal of the Electrochemical Society*, **110**, 614–621; (g) McDonald, D.D. (1993) *Journal of the Electrochemical Society*, **140**, L27–L30; (h) Thompson, G.E., Furneuax, R.C., Wood, G.C., Richardson, J.A. and Goode, J.S. (1978) *Nature*, **271**, 433; (i) Jessensky, O., Müller, F. and Gösele, U. (1998) *Applied Physics Letters*, **72**, 1173–1175; (j) Ono, S., Ichinose, H. and Masuko, N. (1992) *Corrosion Science*, **33**, 841–850; (k) Uchi, H., Kanno, T. and Alwitt, R.S. (2001) *Journal of the Electrochemical Society*, **148**, B17–B23; (l) Masuda, H., Yamada, H., Satoh, M., Asoh, H., Nakao, M. and Tamamura, T. (1997) *Applied Physics Letters*, **71**, 2770–2772.

10 (a) Schneider, J.J. and Engstler, J. (2008) Mesoscopic Ceramic Structures in 1D, 2D and 3D. (eds Riedel, R. and Chen, J.) *Ceramics Science and Technology*, Vol. 1, Structures, Wiley-VCH, Weinheim.

11 Petkov, N., Platschek, B., Morris, M.A., Holmes, J.D. and Bein, T. (2007) *Chemistry of Materials*, **19**, 1376–1381.

12 Rice, R.L., Arnold, D.C., Shaw, M.T., Iacopina, D., Quinn, A.J., Amenitsch, H., Holmes, J.D. and Morris, M.A. (2007) *Advanced Functional Materials*, **17**, 133–141.

13 (a) Possin, G.E. (1970) *Review of Scientific Instruments*, **41**, 772–774; (b) Price, P.B. and Walter, R.M. (1962) *Journal of Applied*

Physics, **33**, 3407–3412; (c) Kawai, S. and Ueda, R. (1975) *Journal of the Electrochemical Society*, **22**, 32–36.

14 Karim, S., Toimil-Molares, M.E., Maurer, F., Miehe, G., Ensinger, W., Liu, J., Cornelius, T.W. and Neumann, R. (2006) *Applied Physics A*, **84**, 403–407.

15 Ensinger, W. and Vater, P. (2005) *Materials Science and Engineering C*, **25**, 609–613.

16 Ferre, R., Ounadjela, K., George, J.M., Pireuax, L. and Dubois, S. (1997) *Physical Review B-Condensed Matter*, **56**, 14066–14075.

17 (a) Parallel rod (bunch) arrangement, Love, J.C., Urbach, A.R., Prentiss, M.G. and Whitesides, G.M. (2003) *Journal of the American Chemical Society*, **125**, 12696; (b) one after each other arrangement, Volkov, V.V., Schofield, M.A. and Zhu, Y. (2003) *Modern Physics Letters B*, **17**, 791–801.

18 Wetz, F., Soulantica, K., Falqui, A., Respaud, M., Snoeck, E. and Chaudret, B. (2007) *Angewandte Chemie*, **119**, 7209–7211. (2007) *Angewandte Chemie-International Edition*, **46**, 7079–7081.

19 Trahey, L., Becker, C.R. and Stacy, A. (2007) *Nano Letters*, **7**, 2535–2539.

20 Decher, G. (1997) *Science*, **277**, 1232–1237.

21 (a) Lu, Q., Feng, G., Kormaneni, S. and Mallouk, T.E. (2004) *Journal of the American Chemical Society*, **126**, 8650–8651; (b) Wu, Y., Cheng, G., Katsov, K., Sides, S.W., Wang, J., Tiang, J., Frederickson, G.H., Moskovits, M. and Stucky, G.D. (2004) *Nature Materials*, **3**, 816; (c) Wu, Y., Livneh, T., Zhang, Y.X., Cheng, G., Wang, J., Tang, J., Moskovits, M. and Stucky, G. (2004) *Nano Letters*, **4**, 2337–2342.

22 (a) Reneker, D.H. and Chun, I. (1996) *Nanotechnology*, **7**, 216–223; (b) Greiner, A. and Wendorff, H. (2007) *Angewandte Chemie*, **119**, 5770–5805; (2007) *Angewandte Chemie-International Edition*, **46**, 5670–5703.

23 (a) Taylor, G. (1969) *Proceedings of the Royal Society of London. Series A*, **313**, 453–475; (b) Doshi, J. and Reneker, D.H. (1995) *Journal of Electrostatics*, **35**, 151–160; (c) Reneker, D.H., Yarin, A.L., Fong, H. and Koanbhongse, S. (2000) *Journal of Applied Physics*, **87**, 4531–4547; (d) Larsen, G., Velarde-Ortiz, R., Minchow, K., Barrero, A. and Loscertales, I.G. (2003) *Journal of the American Chemical Society*, **125**, 1154–1155.

24 Dai, H., Gong, J., Kim, H. and Lee, D. (2002) *Nanotechnology*, **13**, 674–677.

25 Li, D. and Xia, Y. (2003) *Nano Letters*, **3**, 555–560.

26 (a) Li, D., Wang, Y. and Xia, Y. (2003) *Nano Letters*, **3**, 1167–1171; (b) Li, D. and Xia, Y. (2004) *Advanced Materials*, **16**, 1151–1170.

27 McCann, J.T., Li, D. and Xia, Y. (2005) *Journal of Materials Chemistry*, **15**, 735–738.

28 Hunter, R.J. (1981) *Zeta Potential in Colloid Science: Principles and Applications*, Academic Press, London.

29 Everett, D.H. (1988) *Basic Principles of Colloid Science*, Royal Society of Chemistry London.

30 Schubert, U. and Hüsing, N. (2000) *Synthesis of Inorganic Materials*, Wiley-VCH, Weinheim.

31 Cao, G. (2004) *The Journal of Physical Chemistry B*, **108**, 19921–19931.

32 Antonietti, M. and Ozin, G.A. (2004) *Chemistry – A European Journal*, **10**, 28–41.

33 Wagner, R.S. and Ellis, W.C. (1964) *Applied Physics Letters*, **4**, 89–90.

34 Wu, Y., Cui, Y., Huynh, L., Barrelet, C.J., Bell, D.C. and Lieber, C.M. (2004) *Nano Letters*, **4**, 433–436.

35 Holmes, J.D., Johnston, K.P., Doty, R.C. and Korgel, B.A. (2000) *Science*, **287**, 1471–1473.

36 Kodambaka, S., Hannon, J.B., Tromp, R.M. and Ross, F.M. (2006) *Nano Letters*, **6**, 1292–1296.

37 (a) Kovalev, D. and Fujii, M. (2005) *Advanced Materials*, **17**, 2531–2544; (b) Song, J., Wang, X., Riedo, E. and Wang, Z.L. (2005) *The Journal of Physical Chemistry B*, **109**, 9869–9872; (c) Wang, X., Summers, C.J. and Wang, Z.L. (2004) *Nano Letters*, **4**, 423–426.

38 (a) Schmitt, A.L., Bierman, M.J., Schmeisser, D., Himpsel, F.J. and Jiu, S. (2006) *Nano Letters*, **6**, 1617–1621; (b) Szczech, J.P., Schmitt, A.L., Bierman, M.J. and Jin, S. (2007) *Chemistry of Materials*, **19**, 3238–3243.

39 Aella, P., Ingole, S., Petuskey, W.T. and Picraux, S.T. (2007) *Advanced Materials*, **19**, 2603–2607.

40 Murphy, C.J., Gole, A.M., Hunyadi, S.E. and Orendorff, C.J. (2006) *Inorganic Chemistry*, **45**, 7544–7554.

41 Wang, F., Dong, A., Sun, J., Tang, R., Yu, H. and Buhro, W.E. (2006) *Inorganic Chemistry*, **45**, 7511–7521.

42 Hanrath, T. and Korgel, B.A. (2002) *Journal of the American Chemical Society*, **124**, 1424–1429.

43 Hanrath, T. and Korgel, B.A. (2003) *Advanced Materials*, **15**, 437–440.

44 Smith, P.A., Nordquist, C.D., Jackson, T.N., Mayer, T.S., Martin, B.R., Mbindyo, J. and Mallouk, T.E. (2000) *Applied Physics Letters*, **77**, 1399–1401.

45 Whang, D., Jin, S., Wu, Y. and Lieber, C.M. (2003) *Nano Letters*, **3**, 1255–1259.

46 Huang, Y., Duan, X., Wei, Q. and Lieber, C.M. (2001) *Science*, **291**, 630–633.

47 Snow, E.S., Novak, J.P., Campbell, P.M. and Park, D. (2003) *Applied Physics Letters*, **82**, 2145–2147.

48 Patolsky, F. and Lieber, C.M. (2005) *Materials Today*, **8**, 20–28.

49 Patolsky, F., Timko, B.P., Zheng, G. and Lieber, C.M. (2007) *MRS Bulletin*, **32**, 142–149.

50 Bergveld, P. (1972) *IEEE Transactions on Bio-Medical Engineering*, **BME-19**, 342–351.

51 Blackburn, G.F. (1987) in *Biosensors: Fundamentals and Applications* (ed. A.P.F. Turner), Oxford University Press, Oxford, p. 481.

52 Hafeman, D.G., Parce, J.W. and McConnell, H.M. (1988) *Science*, **240**, 1182–1185.

53 (a) Smith, P.A., Nordquist, C.D., Jackson, T.N., Mayer, T.S., Martin, B.R., Mbindyo, J. and Mallouk, T.E. (2000) *Applied Physics Letters*, **77**, 1399–1401; (b) Barsotti, R.J., Vahey, M.D., Wartena, R., Chiang, Y.M., Voldman, J. and Stellacci, F. (2007) *Small*, **3**, 488–499; (c) Shang, L., Clare, T.L., Eriksson, M.A., Marcus, M.S., Metz, K.M. and Hamers, R.J. (2005) *Nanotechnology*, **16**, 2846–2851; (d) Dong, L., Bush, J., Chiarayos, V., Solanki, R., Jiao, J., Ono, Y., Conley, J.F. Jr. and Ulrich, B.D. (2005) *Nano Letters*, **5**, 2112–2115.

54 Marcus, M.S., Shang, L., Li, B., Streifer, J.A., Beck, J.D., Perkins, E., Eriksson, M.A. and Hamers, R.J. (2007) *Small*, **3**, 1610–1617.

55 Vijayaraghavan, A., Blatt, S., Weissenberger, D., Oron-Carl, M., Hennrich, F., Gerthsen, D., Hahn, H. and Krupke, R. (2007) *Nano Letters*, **7**, 1556–1560.

56 Small, W.R. and in het Panhuis, M. (2007) *Small*, **3**, 1500–1503.

57 Li, Q., Koo, S.-M., Richter, C.A., Edelstein, M.D., Bonevich, J.E., Kopanski, J.J., Suehle, J.S. and Vogel, E.M. (2007) *IEEE Transactions on Nanotechnology*, **6**, 256–262.

58 Ingole, S., Aella, P., Haerne, S.J. and Picraux, S.T. (2007) *Applied Physics Letters*, **91**, 033106–033106.

59 Cui, Y., Wei, Q., Park, H. and Lieber, C.M. (2001) *Science*, **293**, 1289–1292.

60 Hahm, J. and Lieber, C.M. (2004) *Nano Letters*, **4**, 51–54.

61 (a) Gao, Z., Agarwal, A., Trigg, A.D., Singh, N., Fang, C., Tung, C.-H., Fan, Y., Buddharaju, K.D. and Kong, J. (2007) *Analytical Chemistry*, **79**, 3291–3297; (b) Patolsky, F. (2005) *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 3208–3212; (c) Stadler, K. (2003) *Nature Reviews Microbiology*, **1**, 209–218; (d) Atlas, R.M. (2003) *Nature Reviews Microbiology*, **1**, 70–74; (e) Nijler, E. (2002) *Nature Biotechnology*, **20**, 21–25.

62 Patolsky, F., Zheng, G., Hayden, O., Lakadamyali, M., Zhuang, X. and Lieber, C.M. (2004) *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 14017–14022.

63 (a) Whang, D., Jin, S. and Lieber, C.M. (2004) *Japanese Journal of Applied Physics*, **43**, 4465–4470; (b) Jin, S., Whang, D., McAlpine, M.C., Friedman, R.S., Wu, Y. and Lieber, C.M. (2004) *Nano Letters*, **4**, 915–919.

64 (a) McAlpine, M.C., Ahmad, H., Wang, D. and Heath, J.R. (2007) *Nature Materials*, **6**, 379–384; (b) Boukai, A., Diana, F., Gerardot, B., Badolato, A., Petroff, P.M. and Heath, J.R. (2003) *Science*, **300**, 112–115.

65 (a) Krupke, R., Hennrich, F., Weber, H.B., Beckmann, D., Hamper, O., Malik, S., Kappes, M.M. and Löhneysen, H.v. (2003) *Applied Physics A*, **76**, 397–400; (b) Chung, J., Lee, K.-H. and Ruoff, R.S. (2004) *Langmuir*, **20**, 3011–3017.

66 see e.g. Ingole, S., Aella, P., Haerne, S.J. and Picraux, S.T. (2007) *Applied Physics Letters*, **91**, 033106–033106.

67 (a) Hirsch, A. (1994) *The Chemistry of Fullerenes*, Georg Thieme, Stuttgart; (b) Krüger, A. (2007) *Neue Kohlenstoffmaterialien – eine Einführung*, Teubner, Wiesbaden.

68 Daniel, S., Rao, T.P., Rao, K.S., Rani, S.U., Naidu, G.K.R., Lee, H.Y. and Kawai, T. (2007) *Sensors and Actuators B-Chemical*, **122**, 672–682.

69 Coleman, K.S., Bailey, S.R., Fogden, S. and Green, M.L.H. (2003) *Journal of the American Chemical Society*, **125**, 8722–8723.

70 Williams, K.A., Peter, T.M., Veenhuizen Torre, B.G., Eritja, R. and Dekker, C. (2002) *Nature*, **420**, 761–762.

71 Taft, B.J., Lazareck, A.D., Withey, G.D., Yin, A., Xu, J.M. and Kelley, S.O. (2004) *Journal of the American Chemical Society*, **126**, 12750–12751.

72 Cai, H., Cao, X., Jiang, Y., He, P. and Fang, Y. (2003) *Analytical and Bioanalytical Chemistry*, **375**, 287–293.

73 Lu, G., Maragakis, P. and Kaxiras, E. (2005) *Nano Letters*, **5**, 897–900.

74 Liu, Y., Meyer-Zaika, W., Franzka, S., Schmid, G., Leis, M. and Kuhn, H. (2003) *Angewandte Chemie*, **115**, 2959–2963; (2003) *Angewandte Chemie-International Edition*, **42**, 2853–2857.

75 (a) Huang, M.H., Wu, Y., Feick, H., Tran, N., Weber, E. and Yang, P. (2001) *Advanced Materials*, **13**, 113–116; (b) Pan, Z.W., Dai, Z.R. and Wang, Z.L. (2001) *Science*, **291**, 1947–1949; (c) Kong, X.Y. and Wang, Z.L. (2003) *Nano Letters*, **3**, 1625–1631; (d) Kong, X.Y., Ding, Y., Yang, R. and Wang, Z.L. (2004) *Science*, **303**, 1348–1351; (e) Hughes, W.L. and Wang, Z.L. (2004) *Journal of the American Chemical Society*, **126**, 6703–6709; (f) Gao, P.X., Ding, Y., Mai, W., Hughes, W.L., Lao, C. and Wang, Z.L. (2005) *Science*, **309**, 1700–1704.

76 (a) Glaspell, G.G., Hassan, H.M.A., Elzatahry, A., Fuoco, L., Radwan, N.R.E. and El-Shall, M.S. (2006) *The Journal of Physical Chemistry B*, **110**, 21387–21393; (b) Yang, H.M. and Liao, P.H. (2007) *Applied Catalysis A-General*, **317**, 226–233; (c) Dalal, S.H., Baptista, D.L., Teo, K.B.K., Lacerda, R.G., Jefferson, D.A. and Milne, W.I. (2006) *Nanotechnology*, **17**, 4811–4818; (d) Comini, E., Faglia, G., Ferroni, M. and Sberveglieri, G. (2007) *Applied Physics A*, **88**, 45–48; (e) Jeong, M.C., Oh, B.Y., Nam, O.H., Kim, T. and Myoung, J.M. (2006) *Nanotechnology*, **17**, 526–530.

77 Wang, Z.L. (2007) *Materials Today*, **10**, 20–28.

78 Wang, Z.L. and Song, J. (2006) *Science*, **312**, 242–246.

79 Gao, P.X., Song, J., Liu, J. and Wang, Z.L. (2007) *Advanced Materials*, **19**, 67–72.

80 He, H., Hsiu, C.L., Liu, J., Chen, L.J.V. and Wang, Z.C. (2007) *Advanced Materials*, **19**, 781–784.

81 Buchine, B.A., Hughes, W.L., Degertekin, F.L. and Wang, Z.L. (2006) *Nano Letters*, **6**, 1155–1159.

82 Duan, X. (2007) *MRS Bulletin*, **32**, 134–141.

83 Gao, L., Li, Q., Luan, W., Kawaoka, H., Sekino, T. and Niihara, K. (2002) *Journal of the American Ceramic Society*, **85**, 1016–1018.

84 (a) Duan, X., Niu, Ch., Sahi, V., Chen, J., Parce, J.W., Empedocles, St. and Goldman, J.L. (2003) *Nature*, **425**, 274–278; (b) Ong, B.S., Li, Ch., Li, Y., Wu, Y. and Loutfy, R. (2007) *Journal of the American Chemical Society*, **129**, 2750–2751.

85 Sun, B. and Sirringhaus, H. (2005) *Nano Letters*, **5**, 2408–2413.

86 (a) Durkop, T., Getty, S.A., Cobas, E. and Fuhrer, M.S. (2004) *Nano Letters*, **4**, 35–39; (b) Snow, E.S., P: Novak, J., M: Campbell, P. and Park, D. (2003) *Applied Physics Letters*, **82**, 2145–2147; (c) Xiao, K., Liu, Y., Hu, P., Yu, G., Wang, X. and Zhu, D. (2003) *Applied Physics Letters*, **82**, 2145–2147; (d) Bradley, K., Gabriel, J.C.P. and Grüner, G. (2003) *Nano Letters*, **3**, 1353–1355; (e) Seidel, R., Graham, A.P., Unger, E., Duesberg, G.S., Liebau, M., Steinhoegl, W., Kreupl, F. and Hoehnlein, W. (2004) *Nano Letters*, **4**, 831–834; (f) Zhou, Y., Gaur, A., Hur, S.-H., Kocabas, C., Meitl, M.A., Shim, M. and Rogers, J.A. (2004) *Nano Letters*, **4**, 2031–2035.

87 Kocabas, C., Hur, S.H., Gaur, A., Meitl, M.A., Shim, M. and Rogers, J.A. (2005) *Small*, **1**, 1110–1116.

88 Rotkin, S.V., Ruda, H.E. and Shik, A. (2003) *Applied Physics Letters*, **83**, 1623–1626.

89 (a) Ridley, B.A., Nivi, B. and Jacobson, J.M. (1999) *Science*, **286**, 746–748; (b) Ahn, J.-H., Kim, H.-S., Lee, K.J., Jeon, S., Kang, S.J., Sun, Y., Nuzzo, R.G. and Rogers, J.A. (2006) *Science*, **314**, 1754–1757; (c) Menard, E., Lee, K.J., Khang, D.-Y., Nuzzo, R.G. and Rogers, J.A. (2004) *Applied Physics Letters*, **84**, 5398–5400.

90 Ridley, B.A., Nivi, B. and Jacobson, J.M. (1999) *Science*, **286**, 746–749.

91 Whang, D., Jin, S., Wu, Y. and Lieber, C.M. (2003) *Nano Letters*, **3**, 1255–1259.

92 Shimoda, T., Matsuki, Y., Furusawa, M., Aoki, T., Yudasaka, I., Tanaka, H., Iwasawa, H., Wang, D., Miyasaka, M. and Taakeuchi, Y. (2006) *Nature*, **440**, 783–786.

93 Becker, J.S., Suh, S. and Gordon, R.G. (2003) *Chemistry of Materials*, **15**, 2969–2976.

94 (a) Doh, Y.-J., van Dam, J.A., Roest, A.L., Bakkers, E.P.A.M., Kouwenhoven, L.P. and De Franceschi, S. (2005) *Science*, **309**, 272–275; (b) van Dam, J.A., Nazarov, Y.V., Bakkers, E.P.A.M., De Franceschi, S. and Kouwenhouven, L.P. (2006) *Nature*, **442**, 667–670.

95 Wong, H.S.P. (2002) *IBM Journal of Research and Development*, **46**, 133–167.

96 Schmidt, V., Riel, H., Senz, S., Karg, S., Riess, W. and Gösele, U. (2006) *Small*, **2**, 85–88.

97 For a survey on growth of nanotubes for electronics: Robertson, J. (2007) *Materials Today*, **10**, 36–43.

98 Zhang, L., Tu, T. and Dai, H. (2006) *Nano Letters*, **6**, 2785–2789.

99 (a) Zakharov, N.D., Werner, P., Gerth, G., Schubert, L., Sokolov, L. and Gösele, U. (2006) *Journal of Crystal Growth*, **290**, 6–10; (b) Borgström, M.T., Verheijen, M.A., Immink, G., de Smet, T. and Bakkers, E.P.A.M. (2006) *Nanotechnology*, **17**, 4010–4013.

100 Mieszawska, A.J., Jalilian, R., Sumanasekera, G.U. and Zamborini, F.P. (2007) *Small*, **3**, 722–756.

101 (a) Manna, L., Scher, E.C. and Alivisatos, A.P. (2000) *Journal of the American Chemical Society*, **122**, 12700–12706; (b) Grebinski, J.W., Hull, K.L., Zhang, J., Kosel, T.H. and Kuno, M. (2004) *Chemistry of Materials*, **16**, 5260–5272; (c) L: Hull, K., Grebinski, J.W., Kosel, T.H. and Kuno, M. (2005) *Chemistry of Materials*, **17**, 4416–4425; (d) Yun, Y.W., Lee, S.M., Kang, N.J. and Cheon, J. (2001) *Journal of the American Chemical Society*, **123**, 5150–5151; (e) Jun, Y.W., Jun, Y.Y. and Cheon, J. (2002) *Journal of the American Chemical Society*, **124**, 615–619; (f) Manna, L., Milliron, D.J., Mesiel, A., Scher, E.C. and Alivisatos, A.P. (2003) *Nature Materials*, **2**, 382–385.

102 (a) Dick, K.A., Deppert, K., Larsson, M.W., Martenssson, T., Seifert, W., Wallenberg, L.R. and Samuelson, L. (2004) *Nature Materials*, **3**, 380–384; (b) Dick, K.A., Deppert, K., Karlsson, L.S., Wallenberg,

L.R., Samuelson, L. and Seifert, W. (2005) *Advanced Functional Materials*, **15**, 1603–1610; (c) Dick, K.A., Getertovszky, Zs., Mikkelsen, A., Karlsson, L.S., Lundgren, E., Malm, J.-O., Andersen, J.N., Samuelson, L., Seifert, W., Waacaser, B.A. and Deppert, K. (2006) *Nanotechnology*, **17**, 1344–1350; (d) Su, J., Cui, G., Gherasimova, M., Tsukamoto, H., Han, J., Ciuparu, D., Lim, S., Pfeferrle, L., He, Y., Nurmikko, A.V., Broadbridge, C. and Lehman, A. (2005) *Applied Physics Letters*, **86**, 013105.

103 Dick, K.A., Deppert, K., Karlsson, L.S., Larsson, M.W., Seifert, W., Wallenberg, L.R. and Samuelson, L. (2007) *MRS Bulletin*, **32**, 127–133.

104 (a) Bae, S.Y., Seo, H.W., Choi, H.C., Park, J. and Park, J. (2004) *The Journal of Physical Chemistry B*, **108**, 12318–12326; (b) Zhang, T., Dong, W., Keeter-Brewer, M., Konar, S., Njabon, R.N. and Tian, Z.R. (2006) *Journal of the American Chemical Society*, **128**, 10960–10968.

105 (a) Milliron, D.J., Hughes, S.M., Cui, Y., Manna, L., Li, J.B., Wang, L.W. and Alivisatos, A.P. (2004) *Nature*, **430**, 190–194; (b) for a discussion in more breadth, see Ozin, G.A. and Arsenault, A.C. (2005) *Nanochemistry, a Chemical Approach to Nanomaterials*, Royal Society of Chemistry, Cambridge, Chapter 6, pp. 297–303.

106 Newton, M.C., Firth, S. and Warburton, P.A. (2006) *Applied Physics Letters*, **89**, 072104.

107 (a) Magnusson, M.H., Deppert, K., Malm, J.-O., Bovin, J.-O. and Samuelson, L. (1999) *Nanostructured Materials*, **12**, 45–48; (b) Magnusson, M.H., Deppert, K., Malm, J.-O., Bovin, J.-O. and Samuelson, L. (1999) *Journal of Nanoparticle Research*, **1**, 243–251; (c) Karlsson, M.N.A., Deppert, K., Karlsson, L.S., Magnusson, M.H., Malm, J.-O. and Srinivasan, N.S. (2005) *Journal of Nanoparticle Research*, **7**, 43–49.

108 Dick, K.A., Deppert, K., Karlsson, L.S., Larsson, M.W., Seifert, W., Wallenberg, L.R. and Samuelson, L. (2007) *MRS Bulletin*, **32**, 127–133.

109 Mokari, T., Rothenberg, E., Popov, I., Costi, R. and Banin, U. (2004) *Science*, **304**, 1787–1790.

110 Lieber, C.M. and Wang, Z.L., guest ed., Special issue on Functional Nanowires. (2007) *MRS Bulletin*, **32**, 99–142.

6
Biomolecule–Nanoparticle Hybrid Systems

Maya Zayats and Itamar Willner

6.1
Introduction

Metal and semiconductor nanoparticles (NPs) or quantum dots (QDs) exhibit unique electronic, optical and catalytic properties. The comparable dimensions of NPs or QDs and biomolecules such as enzymes, antigens/antibodies and DNA suggest that by the integration of the biomolecules with NPs (or QDs) combined hybrid systems with the unique recognition and catalytic properties of biomolecules and with the electronic, optical and catalytic features of NPs might yield new materials with predesigned properties and functions. For example, metallic NPs, such as Au or Ag NPs, exhibit size-controlled plasmon excitons. These plasmon absorbance bands are sensitive to the dielectric properties of the stabilizing capping layers of the NPs [1, 2] or to the degree of aggregation of the NPs that leads to interparticle-coupled plasmon excitons [3, 4]. Thus, spectral changes occurring in metal NP assemblies as a result of biomolecule-induced recognition events that occur on NP surfaces and alter the surface dielectric properties of the NP or stimulate aggregation might be used for optical biosensing. Similarly, semiconductor QDs reveal size-controlled absorption and fluorescence features [5–7]. The high fluorescence quantum yields of QDs and the stability of QDs against photobleaching can be used to develop new fluorescent labels for optical biosensors [8, 9]. Alternatively, metallic NPs exhibit catalytic functions reflected by their ability to catalyze the reduction and growth of the NP seeds by the same metal or a different metal to form core–shell NPs. These properties may be applied to form NP-functionalized proteins or nucleic acids that provide hybrid systems that act as electroactive labels for amplified biosensing [10, 11] or as templates for growing nanostructures [12]. Furthermore, the coupling of biomolecules to metallic or semiconductor NPs might allow the use of the electron-conducting properties of metal NPs or the photoelectrochemical functions of semiconductor NPs to develop new electrical or photoelectrochemical biosensors. Indeed, tremendous scientific advances have been achieved in the last few years by conjugating biomolecules and NPs to functional hybrid systems. Numerous new

Nanotechnology. Volume 1: Principles and Fundamentals. Edited by Günter Schmid

ISBN: 978-3-527-31732-5

biosensors and bioanalytical paradigms were developed, and substantial progress in the use of these hybrid systems as building units of nanodevices was achieved. Several comprehensive review articles addressed different aspects and future perspectives of biomolecule–NP hybrid systems [13–18]. This chapter is aimed at summarizing the different venues where the unique optical and electronic properties of biomolecule–NP hybrid systems have been applied and to discuss future opportunities in the area. Naturally, this review is not aimed at providing full bibliographic coverage of the different topics, but it will highlight the different scientific directions that use biomolecule–NP hybrid systems, and address some specific examples.

6.2
Metal Nanoparticles for Electrical Contacting of Redox Proteins

The electrical contacting of redox proteins with electrodes is a key issue in bioelectronics. Numerous redox enzymes exchange electrons with other biological components such as other redox proteins, cofactors or molecular substrates. The exchange of electrons between the redox centers of proteins and electrodes could activate the bioelectrocatalytic functions of these proteins and thus provide a route to design different amperometric biosensors. Most of the redox proteins lack, however, direct electron transfer communication with electrodes, and hence, the bioelectrocatalytic activation of the redox enzymes is prohibited. The lack of electrical contact between the redox centers and the electrode surfaces is attributed to the spatial separation of the redox sites from the electrode by means of the protein shell [19]. Different methods to electrically communicate redox enzymes with electrodes were developed, including the application of diffusional electron mediators [20], tethering redox relays to the proteins [21–23] and the immobilization of redox proteins in electroactive polymers [24–26]. A recently developed procedure for the electrical contacting of redox proteins with electrodes involved the extraction of the native redox cofactor from the protein and the reconstitution of the resulting apo-enzyme on a surface modified with a monolayer consisting of a relay tethered to the respective cofactor units [27–30]. The reconstitution process aligned the protein on the electrode surface in an optimal orientation, while the relay units electrically contacted the cofactor sites with the conductive support, by shortening the electron transfer distances [31]. All of these methods permitted the bioelectrocatalytic activation of the respective enzymes and the development of amperometric biosensors and biofuel cells, [32, 33].

The availability of metal nanoparticles exhibiting conductivity allowed the generation of nanoparticle–enzyme hybrid systems for controlled electron transfer [34, 35]. Recently, highly efficient electrical contacting of the redox enzyme glucose oxidase (GOx) through a single Au nanoparticle (Au NP) was demonstrated [36]. The GOx–Au NP conjugate was constructed by the reconstitution of an apo-flavoenzyme, apo-glucose oxidase (apo-GOx) on a 1.4-nm Au_{55} nanoparticle functionalized with N^6-(2-aminoethyl) flavin adenine dinucleotide (FAD cofactor, amino derivative, **1**). The resulting enzyme–NP conjugate was assembled on a

thiolated monolayer by using different dithiols (**2**–**4**) as linkers [Figure 6.1(A), route a]. Alternatively, the FAD-functionalized Au nanoparticle was assembled on a thiolated monolayer associated with an electrode, and apo-GOx was subsequently reconstituted on the functional nanoparticles [Figure 6.1(A), route b]. The enzyme electrodes prepared by these two routes revealed similar protein surface coverage of about 1×10^{-12} mol cm^{-2}. The Au NP was found to act as a nanoelectrode that acted as relay units transporting the electrons from the FAD cofactor embedded in the protein to the electrode with no additional mediators, thus activating the bioelectrocatalytic functions of the enzyme. Figure 6.1(B) shows the cyclic voltammograms generated by the enzyme-modified electrode, in the presence of different concentrations of glucose. The electrocatalytic currents increase as the concentrations of glucose are elevated, and the appropriate calibration curve was extracted [Figure 6.1(B), inset]. The resulting nanoparticle-reconstituted enzyme electrodes revealed unprecedented efficient electrical communication with the electrode (electron transfer turnover rate about 5000 s^{-1}). This effective electrical contacting, far higher than the turnover rate of the enzyme with its native electron acceptor, oxygen (about 700 s^{-1}), made the enzyme electrode insensitive to oxygen or to ascorbic acid or uric acid, which are common interferents in glucose biosensing. The rate-limiting step in the electron transfer communication between the enzyme redox center and the electrode was found to be the charge transport across the dithiol molecular linker that bridges the particle to the electrode. The conjugated benzenedithiol (**4**) was found to be the most efficient electron-transporting unit among the linkers (**2**–**4**).

A similar concept was applied to electrically contact pyrroloquinoline quinone (PQQ)-dependent glucose dehydrogenase [37]. Apo-glucose dehydrogenase (apo-GDH) was reconstituted on PQQ-cofactor units (**5**) covalently linked to the amino-functionalized Au NPs [Figure 6.2(A)]. The electrocatalytic anodic currents developed by the enzyme-modified electrode, in the presence of variable concentrations of glucose, are depicted in Figure 6.2(B). The resulting electrocatalytic currents imply that the system is electrically contacted and that the Au NPs mediate the electron transfer from the PQQ-cofactor center embedded in the protein to the electrode. Using the saturation current value generated by the system [Figure 6.2(B), inset] and knowing the surface coverage of the reconstituted enzyme, 1.4×10^{-10} mol cm^{-2}, the electron transfer turnover rate between the biocatalyst and the electrode was estimated to be 1180 s^{-1}, a value that implies effective electrical communication between the enzyme and the electrode, which leads to the efficient bioelectrocatalytic oxidation of glucose.

6.3
Metal Nanoparticles as Electrochemical and Catalytic Labels

The possibility to functionalize metallic nanoparticles with different biomolecules allows the use of the biomolecule–-NP conjugates as labels for the amplified detection of biorecognition events. The NPs may be modified by their direct functionalization with the biomolecules, for example, the binding of thiolated nucleic acids to Au or Ag

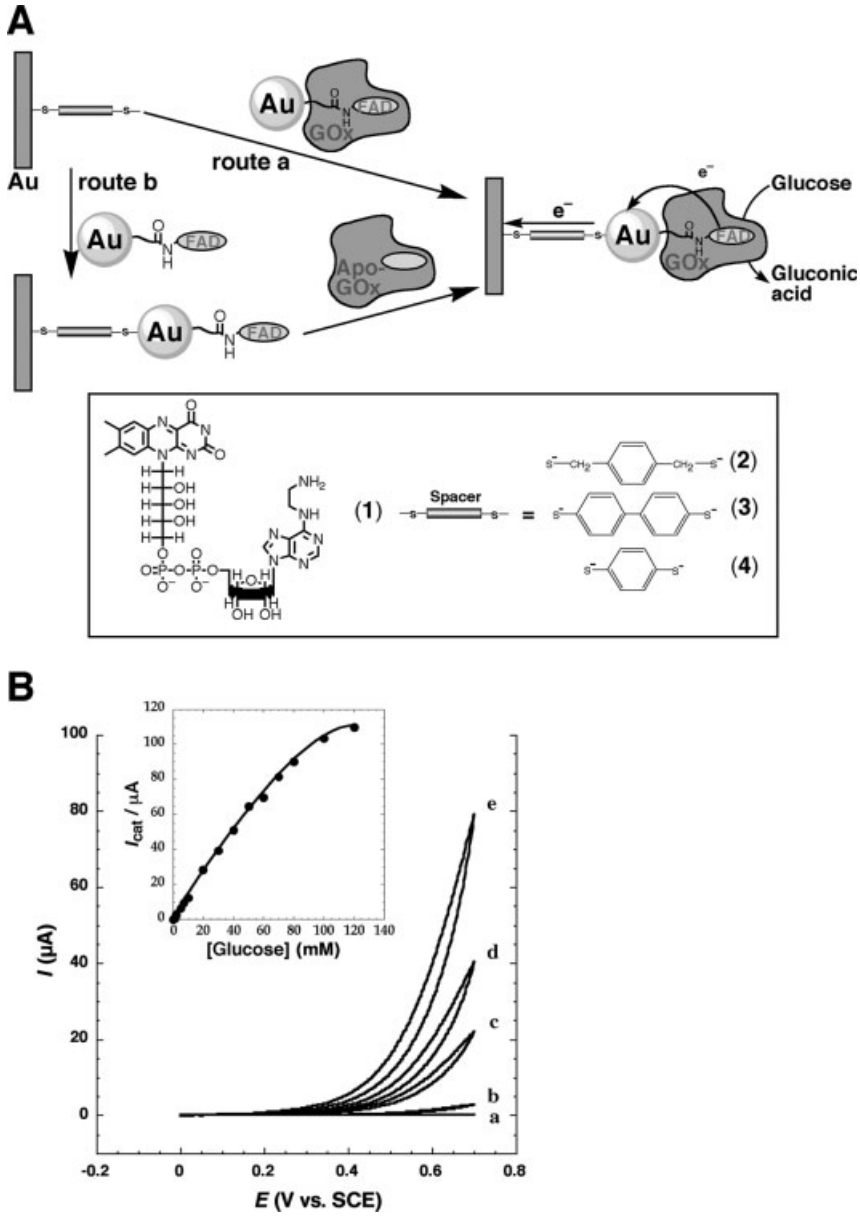


Figure 6.1 (A) The assembly of an electrically-contacted glucose oxidase (GOx) monolayer-functionalized electrode by the reconstitution of the apo-enzyme on an Au NP modified with the flavin adenine dinucleotide (FAD) cofactor (**1**). (B) Cyclic voltammograms corresponding to the bioelectrocatalyzed oxidation of different glucose concentrations by the GOx-reconstituted electrode: (a) 0, (b) 1, (c) 10, (d) 20 and (e) 50 mM. Scan rate 5 mV s^{-1}. Inset: calibration plot derived from the cyclic voltammograms at $E = 0.6$ V in the presence of different concentrations of glucose. (Reproduced from [36]. Reprinted with permission from AAAS).

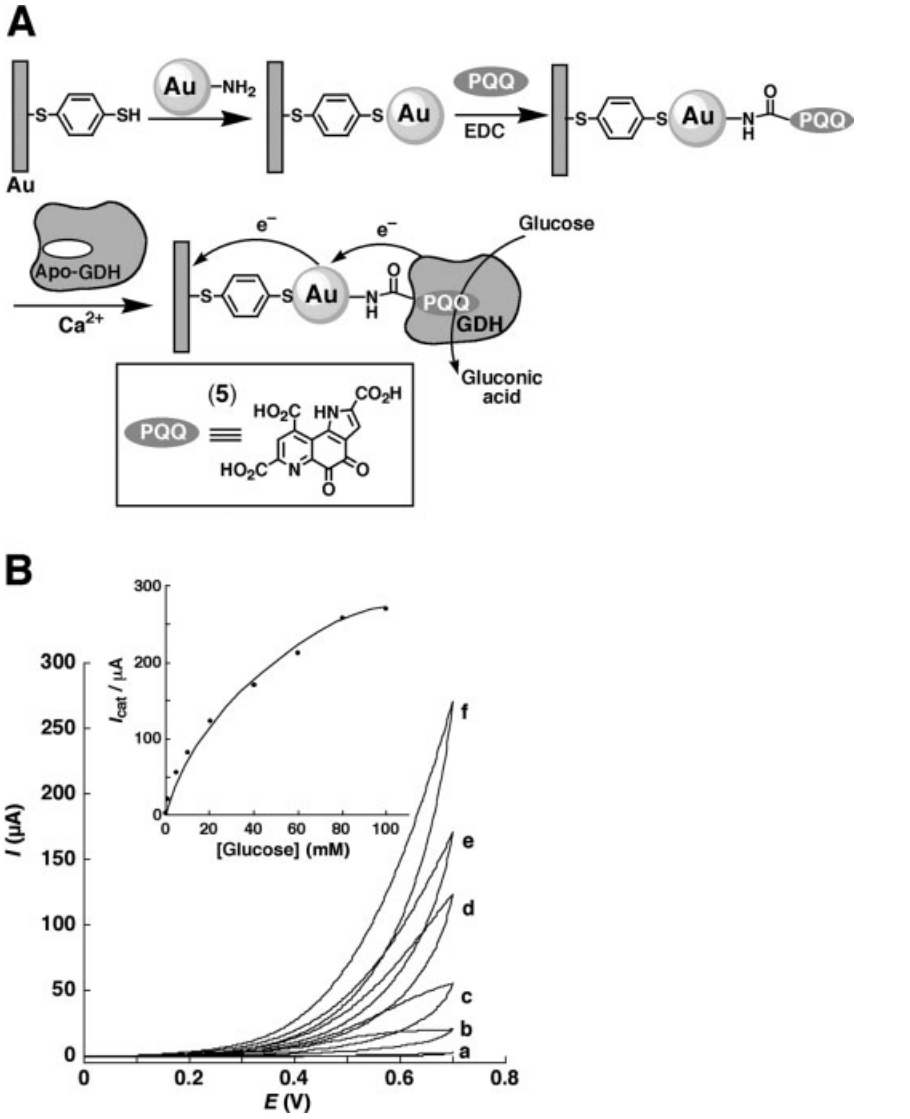


Figure 6.2 (A) The assembly of an electrically contacted glucose dehydrogenase (GDH) enzyme electrode by the reconstitution of the apo-enzyme on a PQQ (**5**)-functionalized Au NP associated with the electrode. (B) Cyclic voltammograms corresponding to the bioelectrocatalyzed oxidation of glucose by the GDH reconstituted on the PQQ-functionalized Au NPs associated with an Au electrode in the presence of different concentrations of glucose: (a) 0, (b) 1, (c) 5, (d) 20, (e) 40 and (f) 100 mM. Potential scan rate 5 mV s^{-1}. Inset: calibration plot derived from the cyclic voltammograms at $E = 0.7$ V. (Reprinted with permission from [37]. Copyright 2005 American Chemical Society).

NPs [3, 38, 39], the covalent tethering of the biomolecules to chemically functionalized NPs [40, 41] or the supramolecular binding of biomolecules to functionalized NPs, for example, the association of avidin-tagged nucleic acids with biotinylated NPs [42]. The biomolecule-modified NPs may be employed as electrochemical tracers for the amplified detection of biorecognition events [43, 44]. The chemical dissolution of the NP labels associated with the biorecognition events followed by electrochemical preconcentration of the released ions on the electrode and the subsequent stripping off of the collected metal provide a general means for the use of the particles as amplifying units for the biorecognition events [45]. Alternatively, the direct electrochemical stripping off of the NPs bound to the biorecognition complex was employed to transduce the biorecognition events [46–48]. These methods for stripping of the electrochemically pre-concentrated metals or the direct electrochemical dissolution of the metals led to 3–4 orders of magnitude improved detection limits, compared with normal pulse voltammetric techniques used to monitor DNA hybridization.

The method for the detection of DNA by the capturing the gold [49, 50] or silver [51] nanoparticles on the hybridized target, followed by the anodic stripping off of the metal tracer is depicted in Figure 6.3. Picomolar and sub-nanomolar levels of the DNA target have thus been detected. For example, the electrochemical method was employed for the Au NP-based quantitative detection of the 406-base human cytomegalovirus DNA sequence (HCMV DNA) [50]. The HCMV DNA was immobilized on a microwell surface and hybridized with the complementary oligonucleotide-modified Au nanoparticles as labels. The resulting surface-immobilized Au nanoparticle double-stranded assembly was treated with HBr–Br_2, resulting in the oxidative dissolution of the gold particles. The solubilized Au^{3+} ions were then electrochemically reduced and accumulated on the electrode and subsequently analyzed by anodic stripping voltammetry. The same approach was applied for analyzing an antigen [52] using Au nanoparticle labels and stripping voltammetry measurement. Further sensitivity enhancement can be obtained by catalytic enlargement of the gold tracer in connection with nanoparticle-promoted precipitation of gold [49] or silver [53–55]. Combining such enlargement of the metal particle tags, with the effective "built-in" amplification of electrochemical stripping analysis, paved the way to sub-picomolar detection limits. The silver-enhanced Au NP stripping method was used for the detection of DNA sequences related to the BRCA1 breast cancer gene [53]. The method showed substantial signal amplification as a result of

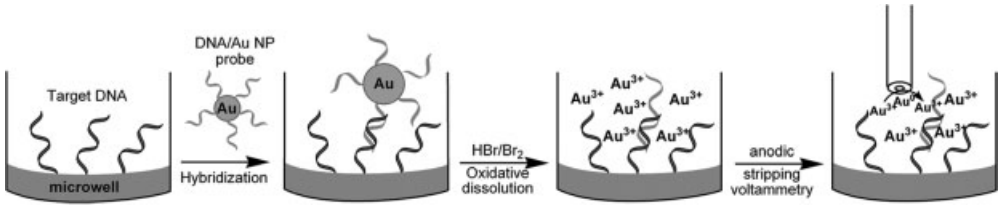


Figure 6.3 Electrochemical detection of a DNA by its labeling with a complementary nucleic acid-functionalized Au NP and the subsequent dissolution of the NPs and the electrochemical stripping of the Au^{3+} ions.

the catalytic deposition of silver on the gold tags; the silver signal was 125 times greater than that of the gold tag. A detection limit of 32 pM was achieved.

A further modification of the metal NP-induced amplified detection of DNA involved combination of magnetic particles and biomolecule-functionalized metallic NPs as two inorganic units that operate successfully in biosensing events. The formation of a biorecognition complex on the magnetic particles followed by labeling of the complex with the metallic NPs allowed the magnetic separation of the metal-labeled recognition complexes and their subsequent electrochemical detection by stripping voltammetry [49, 53]. Whereas the magnetic separation of the labeled recognition complexes enhanced the specificity of the analytical procedures, the stripping off of the metallic labels contributed to specific analysis [56, 57]. Furthermore, the use of the metallic NP labels accumulated on the magnetic NPs as catalytic seeds for the electroless enlargement of the particles by metals provided a further amplification path for the electrochemical stripping of the labels [57].

Figure 6.4 depicts the amplified detection of DNA by the application of nucleic acid-functionalized magnetic beads and Au NPs as catalytic seeds for the deposition of silver [57]. A biotin-labeled nucleic acid (**6**) was immobilized on the avidin-functionalized magnetic particles and hybridized with the complementary biotinylated nucleic acid (**7**). The hybridized assembly was then reacted with the Au-nanoparticle-avidin conjugate (**8**). Treatment of the magnetic particles–DNA–Au nanoparticle conjugate with silver ions (Ag^+) in the presence of hydroquinone results in the electroless catalytic deposition of silver on the Au nanoparticles, acting as catalytic labels. The latter process provided the amplification path since the catalytic accumulation of silver on the Au nanoparticle originates from a single DNA recognition event. The magnetic separation of the particles by an external magnet concentrated the hybridized assembly from the analyzed sample. The current originated by the voltammetric stripping off of the accumulated silver then provided the electronic

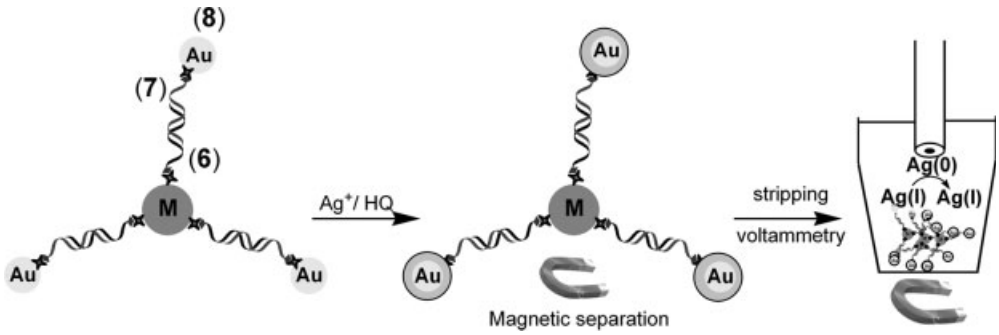


Figure 6.4 Electrochemical analysis of a DNA (**7**) by its hybridization with the complementary nucleic acid (**6**)-functionalized magnetic particle and the hybridization with the complementary nucleic acid (**8**)-modified Au NPs, followed by the deposition of Ag^0 on the particles. The analysis is performed by the magnetic separation of the particles aggregates, followed by the electrochemical stripping of the metallic NPs.

signal that transduced the analysis of the target DNA. Also, Au nanoparticle-based detection of DNA hybridization based on the magnetically induced direct electrochemical detection of the 1.4-nm Au_{67} NP tag linked to the target DNA was reported [58]. The Au_{67} NP tag was directly detected after the hybridization process, without the need for acid dissolution.

An additional method to enhance the sensitivity of electrochemical DNA detection involved the use of polymeric microparticles (carriers) loaded with numerous Au nanoparticle tags [59]. The Au NP-loaded microparticles were prepared by binding of biotinylated Au NPs to streptavidin-modified polystyrene spheres. The hybridization of target DNA immobilized on magnetic beads with the nucleic acid functionalized with Au nanoparticle-carrier polystyrene spheres, followed by the catalytic enlargement of gold labels, magnetic separation and then detection of the hybridization event by stripping voltammetry (Figure 6.5) allowed the determination of DNA targets at a sensitivity corresponding to 300 amol. A further method for the amplified detection of biorecognition complexes included the use of Au NPs as carriers of electroactive tags [60]. That is, the redox-active units capping the Au NPs linked to the biorecognition complexes allowed the amperometric transduction of the biosensing process. A detection limit of 10 amol for analyzing DNA with the functionalized NPs was reported.

The metal NP labels might be linked along double-stranded DNA, rather than be tethered by hybridization to the analyzed DNA, for the amplified electrical analysis of DNA. The generation of the metal nanoclusters along the DNA and their use for the amplified electrical analysis of DNA are depicted in Figure 6.6 and they follow the concepts used for the fabrication of metallic nanowires (see Section 6.10) [61]. The method involves the immobilization of a short DNA primer on the electrode that

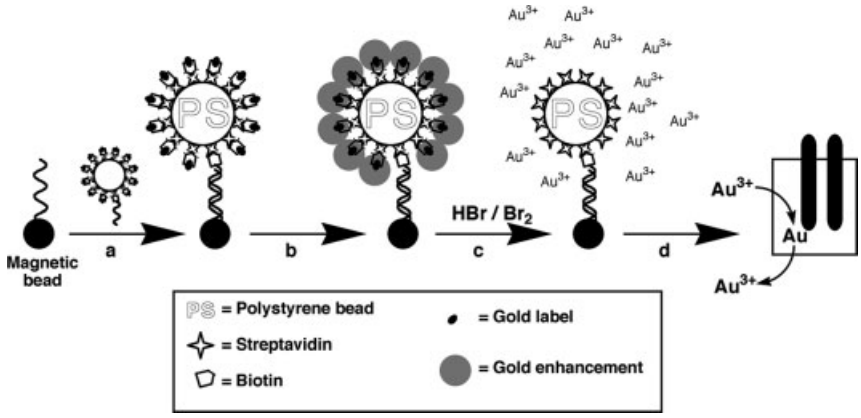


Figure 6.5 Amplified electrochemical detection of DNA by using nucleic acid–Au NP-functionalized microparticles as labels, and electroless catalytic deposition of gold on the NPs as a means of amplification: (a) hybridization of the nucleic acid–Au NP-functionalized microparticles with the target DNA, which is associated with a magnetic bead; (b) enhanced catalytic deposition of gold on the NPs; (c) dissolution of the gold clusters; (d) detection of the Au^{3+} ions by stripping voltammetry. (Reproduced with permission from [59]. Copyright 2004 Wiley-VCH).

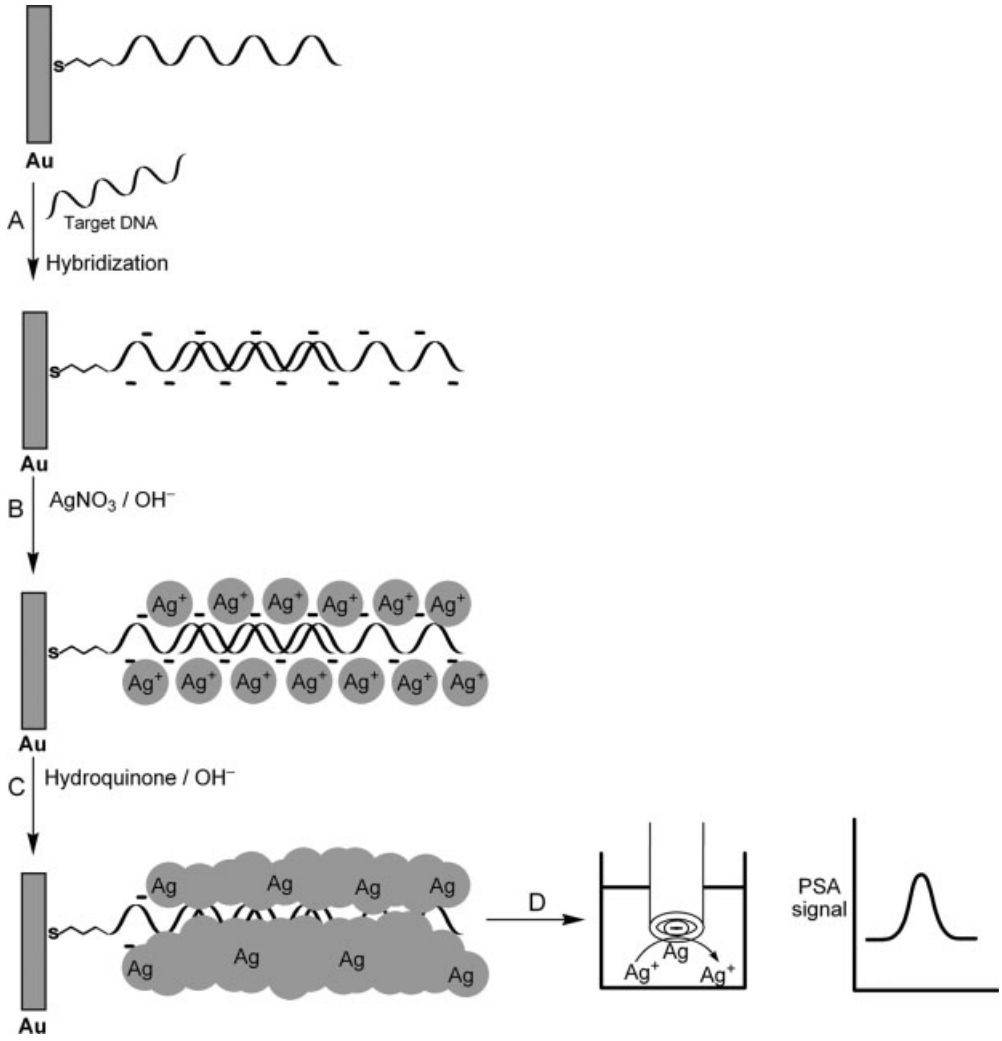


Figure 6.6 Outline of the steps involved in the amplified electrochemical detection of DNA by the deposition of catalytic silver clusters on the DNA strand: (A) hybridization of the complementary target DNA with the short DNA primer, which is covalently linked to the electrode surface through a cystamine monolayer; (B) loading of the Ag^+ ions on to the immobilized DNA; (C) reduction of Ag^+ ions by hydroquinone to form silver aggregates on the DNA backbone; (D) dissolution of the silver aggregates in acidic solution and transfer of the solution to the detection cell for stripping potentiometric measurement (PSA = potentiometric stripping analysis).

hybridizes with the target DNA (step A). The negative charges associated with the phosphate units of the long target DNA collect Ag^+ ions from the solution to form phosphate–Ag^+ complexes (step B). The bound Ag^+ ions are then reduced by hydroquinone, resulting in the formation of metallic silver aggregates along the

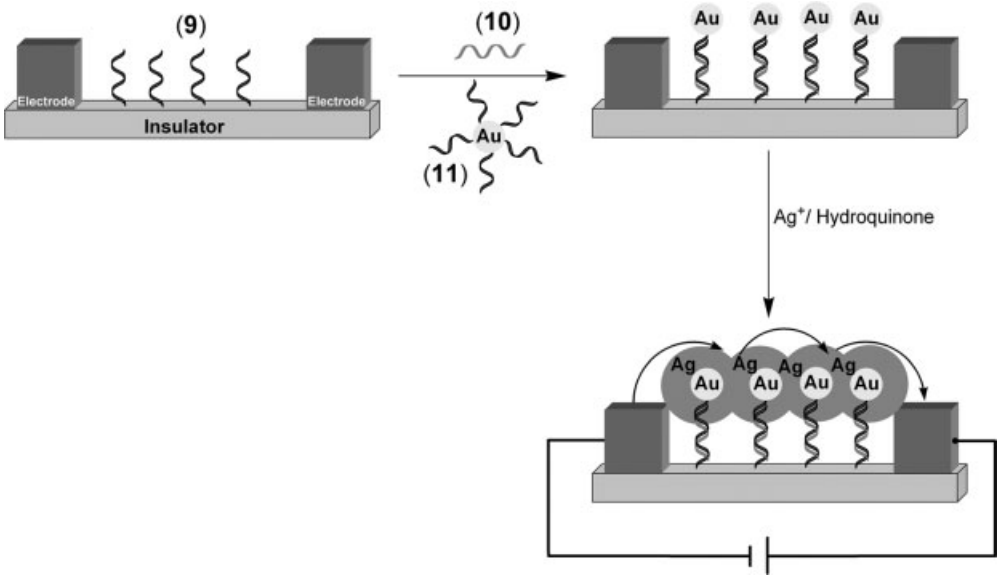


Figure 6.7 The use of a DNA–Au NP conjugate and subsequent silver deposition to connect two microelectrodes, as a means of sensing a DNA analyte.

DNA (step C). The subsequent dissolution and electrochemical stripping of the dissolved silver clusters (step D) then provide the route to detect the hybridized DNA.

The catalytic features of metal nanoparticles permit the subsequent electroless deposition of metals on the nanoparticle clusters associated along the DNA and the formation of enlarged, electrically interconnected, nanostructered wires. The formation of conductive domains as a result of biorecognition events then provides an alternative path for the electrical transduction of biorecognition events. This was exemplified by the design of a DNA detection scheme by using microelectrodes fabricated on a silicon chip [62] (Figure 6.7). The method relied on the generation of a DNA–Au NP sandwich assay within the gap separating two microelectrodes. A probe nucleic acid (**9**) was immobilized in the gap separating the microelectrodes. The target DNA (**10**) was then hybridized with the probe interface and, subsequently, the nucleic acid (**11**)-functionalized Au nanoparticles were hybridized with the free 3′-end of the target DNA, followed by silver enhancement of the Au NP labels. Catalytic deposition of silver on the gold nanoparticles resulted in electrically interconnected particles, exhibiting low resistance between the electrodes. The low resistances between the microelectrodes were controlled by the concentration of the target DNA, and the detection limit for the analysis was estimated to be about 5×10^{-13} M. A difference of 10^6 in the gap resistance was observed upon analyzing by this method the target DNA and its mutant. A related conductivity immunoassay of proteins was developed, based on Au NPs and silver enhancement [63].

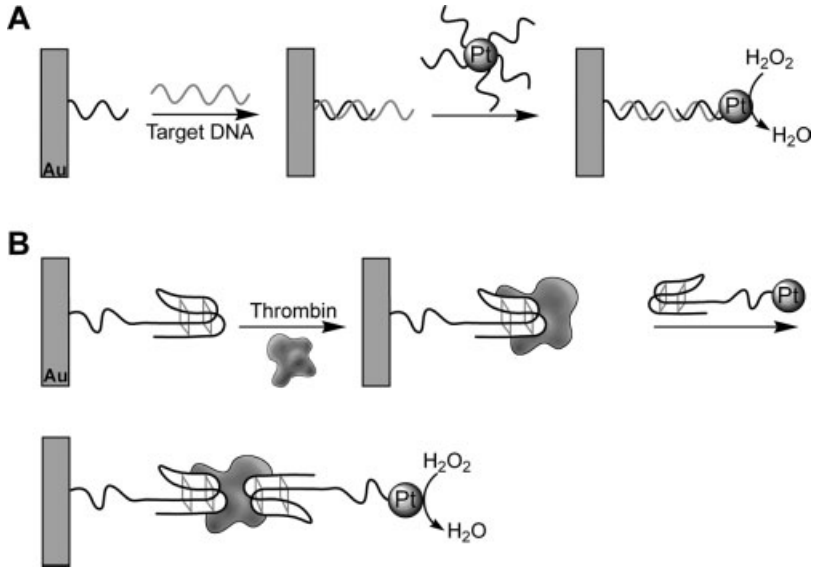


Figure 6.8 (A) Amplified electrochemical analysis of a DNA by nucleic acid-functionalized Pt NPs acting as electrocatalysts for the reduction of H_2O_2. (B) Amplified electrochemical analysis of thrombin by an aptamers monolayer-functionalized electrode and an aptamer-functionalized Pt NP labels as electrocatalysts for the reduction of H_2O_2.

A further approach for the amplified detection of DNA or proteins by employing metal nanoparticle (Pt NP) labels as catalysts for the reduction of H_2O_2 was described [64]. Nucleic acid-functionalized Pt NPs act as catalytic labels for the amplified electrochemical detection of DNA hybridization and aptamer/protein recognition events. Hybridization of the nucleic acid-modified Pt NPs to the nucleic acid–analyte DNA complex associated with an electrode permits the amperometric, amplified, detection of the DNA through the Pt NP electrocatalyzed reduction of H_2O_2 with a sensitivity limit of 1×10^{-11} M [Figure 6.8(A)]. Similarly, the association of the aptamer-functionalized Pt NPs to a thrombin aptamer–thrombin complex associated with the electrode allowed the amplified, electrocatalytic detection of thrombin with a sensitivity limit corresponding to 1×10^{-9} M [Figure 6.8(B)].

6.4
Metal Nanoparticles as Microgravimetric Labels

Nanoparticles provide a "weight label" that may be utilized for the development of microgravimetric sensing methods [quartz crystal microbalance (QCM)] that are ideal for the detection of biorecognition events. Also, the catalytic properties of metallic NPs may be employed to deposit metals on the NP-functionalized biorecognition complexes, thus allowing enhanced mass changes on the transducers

(piezoelectric crystals) and amplified biosensing. For a quartz piezoelectric crystal (AT-cut), the crystal resonance frequency changes by Δf when a mass change Δm occurs on the crystal according to the Sauerbrey equation [65]:

$$\Delta f = -2f_0^2[\Delta m/A(\mu_q\rho_q)^{1/2}] \tag{6.1}$$

where f_0 is the fundamental frequency of the quartz crystal, Δm is the mass change, A is the piezoelectrically active area, ρ_q is the density of quartz (2.648 g cm^{-3}) and μ_q is the shear modulus (2.947×10^{11} dyn cm^{-2} for AT-cut quartz). Thus, any mass changes of the piezoelectric crystals are accompanied by a change in the resonance frequency of the crystal.

The microgravimetric QCM method was applied for the amplified detection of DNA using nucleic acid-functionalized Au NPs as "weight labels" [66–68]. A target DNA molecule (**13**) was hybridized to an Au–quartz crystal that was modified with a probe oligonucleotide (**12**) and the **14**-functionalized Au NPs were hybridized to the 3′-end of the duplex DNA associated with the crystal (Figure 6.9). The subsequent secondary dendritic amplification was achieved by the interaction of the resulting interface with the target DNA (**13**) that was pretreated with the (**12**)-functionalized Au NP [67, 69]. Concentrations of DNA (**13**) as low as 1×10^{-10} M could be detected by the amplification of the target DNA by the nucleic acid-functionalized Au NP labels.

Also, the detection of DNA using nucleic acid-functionalized Au NPs and catalytic metal deposition on the NPs labels was reported [70, 71]. The Au nanoparticles act as catalytic "seeds" and catalyze the reduction of $AuCl_4^-$ and the deposition of gold on the Au NPs. Thus, the catalytic enlargement of the nanoparticles increased the mass associated with the piezoelectric crystal and provided an active amplification route for the amplified microgravimetric detection of the DNA. For example, Figure 6.10(A) depicts the amplified detection of the 7249-base M13mp18 DNA by using the catalytic

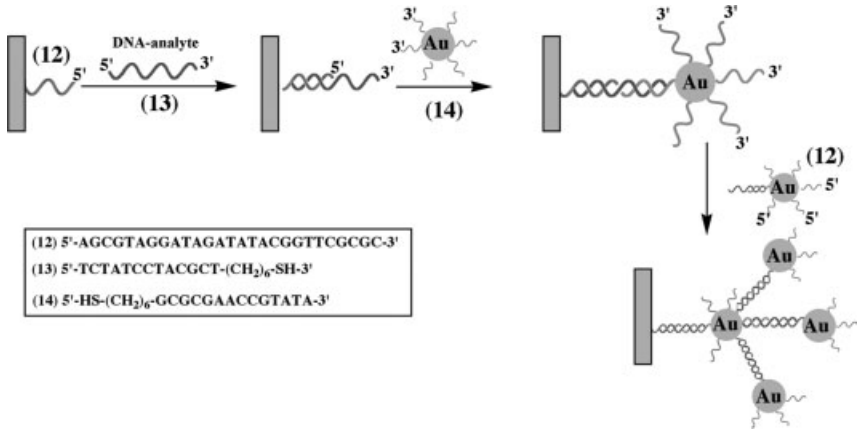
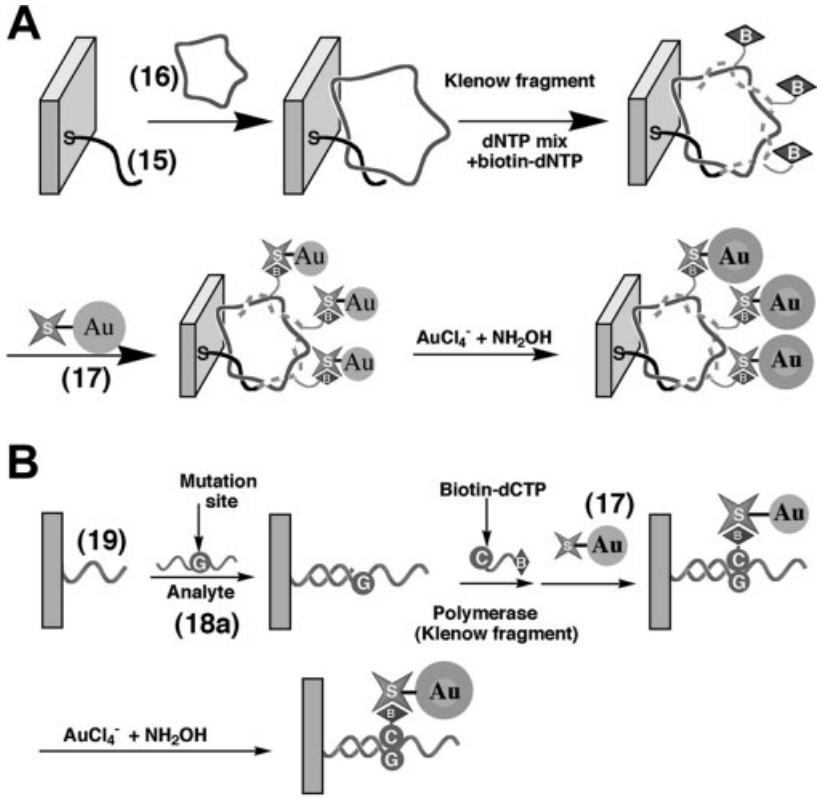


Figure 6.9 Dendritic amplified DNA sensing by the use of oligonucleotide-functionalized Au NPs, which are assembled on a quartz crystal microbalance (QCM) Au–quartz crystal (Reproduced from [67] by permission of The Royal Society of Chemistry).

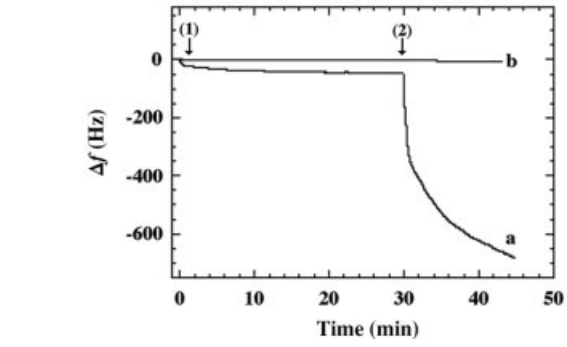
deposition of gold on an Au NP conjugate [71]. The DNA primer (**15**) was assembled on an Au–quartz crystal. After hybridization with M13mp18 DNA (**16**), the double-stranded assembly was replicated in the presence of a mixture of nucleotides (deoxynucleotide triphosphates, dNTP mixture) that included dATP, dGTP, dUTP, biotinylated dCTP (B-dCTP) and polymerase (Klenow fragment). The resulting biotin-labeled replica was then treated with the streptavidin–Au NP conjugate (Sav–Au NP) (**17**), and the resulting Au-labeled replica was subjected to the Au NP-catalyzed deposition of gold by the NH_2OH-stimulated reduction of $AuCl_4^-$. The replication process represents the primary amplification step as it increases the mass associated with the crystal and simultaneously generates a high number of biotin labels for the association of the Sav-Au NP. The binding of the conjugate represents the secondary amplification step for the analysis of M13mp18 DNA. The third step, which involves the catalyzed precipitation of the metal, led to the greatest amplification in the sensing process as a result of the increase in the mass of the Au NPs. This method enabled the M13mp18 DNA to be sensed with a detection limit of $\sim 1 \times 10^{-15}$ M.

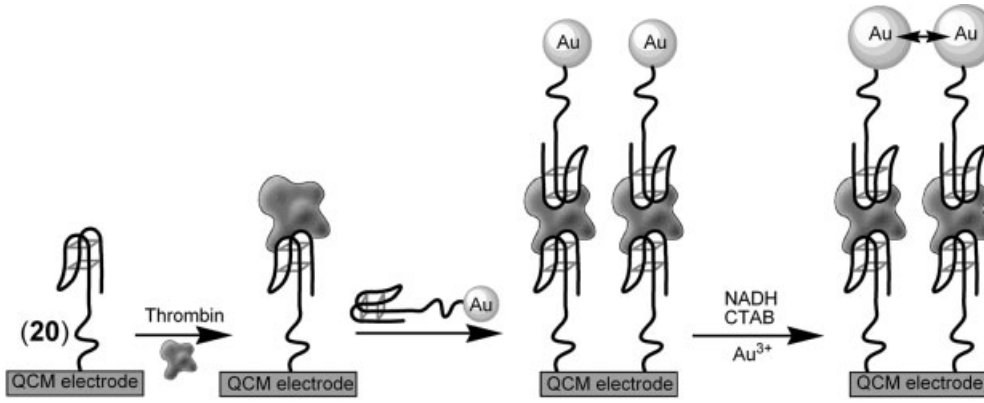
This amplification method was also applied for the analysis of a single-base mismatch in DNA as depicted in Figure 6.10(B) [70, 71]. This was exemplified with the analysis of the DNA mutant **18a**, which differs from the normal gene (**18**) by the substitution of a G base with an A base. The analysis of the mutant was performed by the immobilization of the probe DNA (**19**), which is complementary to the normal gene (**18**) and also to the mutant (**18a**) (up to one base prior to the mutation site), on the Au–quartz crystal. Hybridization of the normal gene or the mutant with this probe interface, followed by the reaction of the hybridized surfaces with biotinylated dCTP (B-dCTP) in the presence of polymerase (Klenow fragment), led to the incorporation of the biotin-labeled base only into the assembly that included the mutant **18a**. The subsequent association of the Sav–Au NP conjugate **17** followed by the catalyzed deposition of gold on the Au NPs amplified the analysis of the single-base mismatch in **18a**. Figure 6.10(C), curve a, shows the microgravimetric detection of the mutant **18a** revealing a frequency change of $\Delta f = -700$ Hz upon analyzing (**18a**), 3×10^{-9} M. The normal gene (**18**) does not alter the frequency of the crystal [Figure 6.10(C), curve b]. The mutant could be detected with a detection limit of 3×10^{-16} M.

Microgravimetric detection method was further extended to analyze proteins by aptamer-functionalized surfaces, using metal NPs as "weight labels" and as catalytic sites for enlargement of nanoparticles and the amplified sensing of the proteins. The use of metal nanoparticle labels for the amplified analysis of thrombin was reported [72]. The fact that thrombin consists of a dimer with two binding sites for aptamers [73, 74] allowed the application of aptamer-functionalized Au NPs for developing a microgravimetric QCM aptasensors for thrombin (Figure 6.11). The thiolated aptamer **20** was linked to an Au–quartz crystal. The interaction of the **20**-modified Au–quartz crystal with thrombin and the subsequent association of the aptamer **20**-functionalized Au NPs provide a primary amplification of the analysis of thrombin by applying the Au NPs as a "weight label". The secondary catalytic enlargement of the Au nanoparticles by the particle-catalyzed reduction of $AuCl_4^-$ by 1,4-dihydronicotinamide adenine dinucleotide (NADH) [75] provided a further



(18a) 5'-CTT TTC TTT TCT TTT GGA TCC GCA AGG CCA GTA ATC AAA CG-3'
(18) 5'-CTT TTC TTT TCT TTT AGA TCC GCA AGG CCA GTA ATC AAA CG-3'
(19) 5'-HS-(CH2)6-CGT TTG ATT ACT GGC CTT GCG GAT C-3'





(20) 5'-HS$(CH_2)_6$TTTTTTTTTTTTTTTGGTTGGTGTGGTTGG-3'

Figure 6.11 Microgravimetric analysis of thrombin by an aptamer (**20**)-functionalized Au–quartz crystal and the aptamer-functionalized Au NP as label and the subsequent enlargement of the NPs by the NADH-induced reduction of Au^{3+} ions. The enlarged NPs act as "weight labels" for the amplified detection of thrombin.

amplification for the sensing of thrombin. Upon analyzing thrombin at a concentration of 2×10^{-9} M, the primary amplification step resulted in a frequency change of -30 Hz, whereas the secondary amplification step altered the crystal frequency by -900 Hz.

6.5
Semiconductor Nanoparticles as Electrochemical Labels for Biorecognition Events

Other inorganic nanoparticle composites such as semiconductor NPs (or quantum dots) have been employed as labels (that substitute metallic NPs) for the amplified electrochemical detection of proteins or DNA [16, 43, 44]. The use of different semiconductor NPs allowed the parallel analysis of different targets, using the NPs as codes for different analytes. For example, CdS semiconductor nanoparticles modified with nucleic acids were employed as labels for the detection of hybridization

Figure 6.10 (A) Amplified detection of the 7249-base M13mp18 DNA (**16**) using the catalytic deposition of gold on an Au nanoparticle conjugate. (B) Analysis of a single-base mismatch in DNA (**18a**) using the catalytic deposition of gold on an Au nanoparticle conjugate. (C) Microgravimetric detection of a single-base mutant (**18a**) enhanced by the catalytic deposition of gold on an Au nanoparticle conjugate. The frequency responses were observed with a mutant DNA (**18a**) (a) and with a normal DNA (**18**) (b). Arrow (1) shows the attachment of the Sav–Au NP conjugate (**17**) and arrow (2) shows catalytic deposition of gold on the Au NPs. (Reproduced from [71] by permission of The Royal Society of Chemistry).

events of DNA [76]. Dissolution of the CdS (in the presence of 1 M HNO_3) followed by the electrochemical reduction of the Cd^{2+} to Cd^0 accumulated the metal on the electrode. The subsequent stripping off of the generated Cd^0 (to Cd^{2+}) provided the electrical signal for the DNA analysis. This method was further developed by using magnetic particles functionalized with probe nucleic acids as sensor units that hybridize with the analyte DNA and the nucleic acid-functionalized CdS NPs labels that hybridize with the single-strand domain of the analyte DNA and trace the primary formation of the probe–analyte double-stranded complex. The magnetic separation of the magnetic particle–CdS NP aggregates crosslinked by the analyte DNA, followed by dissolution of the CdS and electrochemical collection and stripping off of the Cd metal, provide the amplified electrochemical readout of the analyte DNA. In fact, this system combined the advantages of magnetic separation of the tracers CdS NPs associated with the DNA recognition events with the amplification features of the electrochemical stripping method. Highly sensitive detection of DNA is accomplished by this method (detection limit 100 fmol, reproducibility = RSD 6%) [76].

By using different semiconductor NPs as labels, the simultaneous and parallel analysis of different antibodies or different DNAs was accomplished. A model system for multiplexed analysis of different nucleic acids with semiconductor NPs was developed [77]. Three different kinds of magnetic particles were modified by three different nucleic acids (**21a–c**) and subsequently hybridized with the complementary target nucleic acids (**22a–c**). The particles were then hybridized with three different kinds of semiconductor nanoparticles, ZnS, CdS, PbS, that were functionalized with nucleic acids (**23a–c**) complementary to the target nucleic acids associated with the magnetic particles [Figure 6.12(A)]. The magnetic particles allowed the easy separation and purification of the analyte samples, whereas the semiconductor particles provided nonoverlapping electrochemical readout signals that transduced the specific kind of hybridized DNA. Stripping voltammetry of the respective semiconductor nanoparticles yielded well-defined and resolved stripping waves, thus allowing simultaneous electrochemical analysis of several DNA analytes. The same strategy was also applied for the multiplexed immunoassay of proteins [78], with simultaneous analysis of four antigens. The arsenal of inorganic labels for the parallel multiplexed analysis of biomolecules and the level of amplification were further extended by using other metal sulfide composite nanostructures. For example, InS nanorods provided an additional resolvable voltammetric wave, while the nanorods configuration of the label increased the amplification efficiency due to the higher content of stripped-off metal from the nanorod configuration as compared with a spherical NP structure [79].

This method to encode biomolecular identity by semiconductor NPs was extended for the parallel analysis of different proteins by their specific aptamers [80]. An Au electrode was functionalized with aptamers specific for thrombin and lysozyme [Figure 6.12(B)]. Thrombin and lysozyme were labeled with CdS and PbS NPs, respectively, and the NP-functionalized proteins acted as tracer labels for the analysis of the proteins. The NP-functionalized proteins were linked to the respective aptamers and subsequently interacted with the nonfunctionalized thrombin or lysozyme. The competitive displacement of the respective labeled proteins associated with the surface by the analytes, followed by dissolution of the metal sulfides associated with the

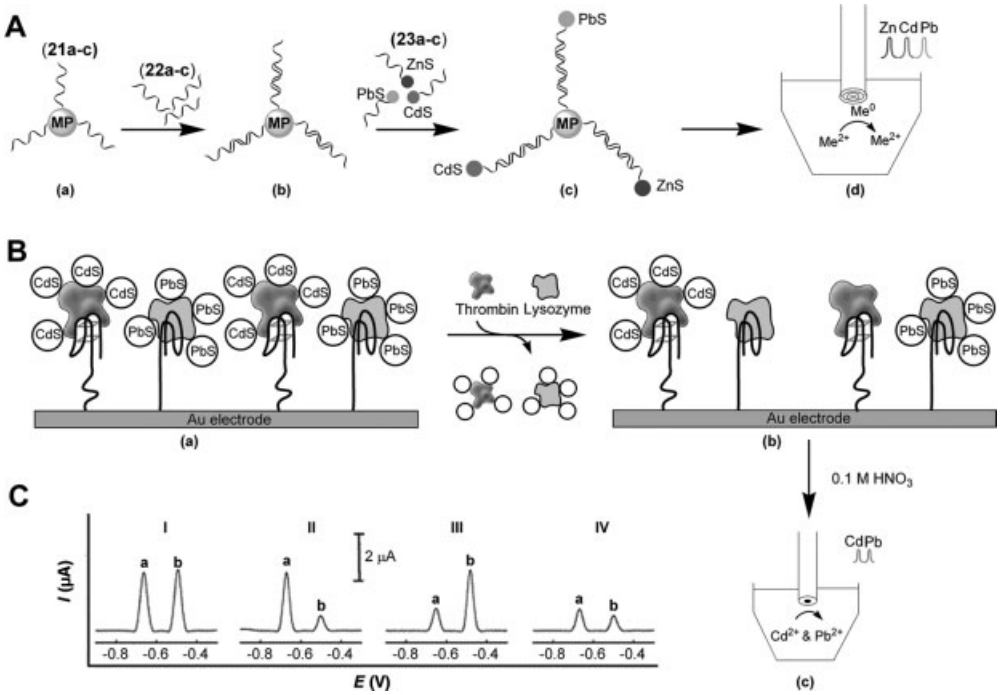


Figure 6.12 (A) Parallel electrochemical analysis of different DNAs using magnetic particles functionalized with probes for the different DNA targets and specific nucleic acid-functionalized metal sulfides as tracers. (B) Simultaneous electrochemical analysis of the two proteins, thrombin and lysozyme, using a competitive assay, where thrombin modified with CdS QDs and lysozyme modified with PbS QDs are used as tracers. (C) Square-wave stripping voltammograms corresponding to the simultaneous detection of lysozyme (a) and thrombin (b): (I) no (a), no (b); (II) 1 μg L^{-1} (a), no (b); (III) no (a), 0.5 μg L^{-1} (b); (IV) 1 μg L^{-1} (a), 0.5 μg L^{-1} (b). (Reprinted with permission from [80]. Copyright 2006 American Chemical Society).

surface and the detection of the released ions by electrochemical stripping, then allowed the quantitative detection of the two proteins [Figure 6.12(C)]. This method was further extended for coding unknown single polymorphisms (SNPs) using different encoding QDs [81]. This protocol relied on the ZnS, CdS, PbS and CuS NPs modified with four different mononucleotides and the application of the NPs to construct different combinations for specific SNPs, that yielded a distinct electronic fingerprints for the mutation sites.

6.6
Metal Nanoparticles as Optical Labels for Biorecognition Events

The unique size-controlled optical properties of the metallic NPs reflected by intense localized plasmon excitons [1, 2, 82] turn the NPs into powerful optical tags for biorecognition processes. Furthermore, the electronic interactions of the localized

plasmon with other plasmonic waves allow one not only to develop new optical amplification paths for biosensing, but also the use these electronic coupling phenomena to follow dynamic processes associated with biorecognition events. For example, surface plasmon resonance (SPR) is a common technique for following biorecognition events at metallic surfaces [83–85]. The changes in the dielectric properties of the metallic surfaces and the changes in the thickness of the dielectric films associated with the metallic surfaces alter the resonance features of the surface plasmon wave and provide the basis for SPR biosensors. The electronic coupling between an Au NP conjugated to the biorecognition complex and the plasmon wave may lead to amplification of the detection processes [86].

A colorimetric detection method for nucleic acids is based on the distance-dependent optical properties of DNA-functionalized Au NPs. The aggregation of Au NPs leads to a red shift in the surface plasmon resonance of the Au NPs as a result of an interparticle coupled plasmon exciton. Thus, the hybridization-induced aggregation of DNA-functionalized Au NPs changes the color of the solution from red to blue [3]. Changes in the optical properties of the Au NPs upon their aggregation provide a method for the sensitive detection of DNA and provide a way to design optical DNA biosensors [87–90]. Specifically, two batches of 13-nm diameter Au NPs were separately functionalized with two thiolated nucleic acids that acted as labels for the detection of the analyte DNA [Figure 6.13(A)]. Each of the NP labels is modified with nucleic acid complementary to the two ends of the analyte DNA. Since each of the nucleic acid-functionalized Au NPs includes many modifying oligonucleotides, the addition of the target DNA to a solution of the two DNA-functionalized Au NPs resulted in the crosslinking and aggregation of the nanoparticles through hybridization. Aggregation changed the color of the solution from red to purple as a result of interparticle coupled plasmon absorbance [Figure 6.13(B)]. The aggregation process was found to be temperature-dependent, and the aggregated Au NPs can reversibly dissociate upon elevation of the temperature through the melting of the double strands and reassociate with a decrease in the temperature through the rehybridization process, which results in the reversible changes of the spectrum [91]. These melting transitions, aggregation and deaggregation, occur in a narrow temperature range [Figure 6.13(C)] and allow the design of selective assays for DNA targets and high discrimination of the mismatched targets. The color changes in the narrow temperature range lies in the background of the simplest test to follow the aggregation of Au NPs, the "Northwestern" spot test [87]. This is an extremely sensitive method to discriminate between aggregated and nonaggregated gold NPs in aqueous solutions, and it relies on a detectable color change from red to blue upon aggregation. The test consists of spotting of a droplet of an aqueous solution of particles on a reversed-phase thin-layer chromatographic plate. A blue spot indicates aggregation in the presence of the target DNA, whereas a red spot indicates the presence of freely dispersed particles [Figure 6.13(C)]. The sharp melting transitions of DNA-functionalized gold nanoparticles were applied to discriminate the target DNA from DNA with single-base-pair mismatches simply by following the changes in the nanoparticle absorption as a function of temperature [87, 88]. The melting properties of DNA-linked nanoparticle aggregates are affected by a number of factors, which include DNA

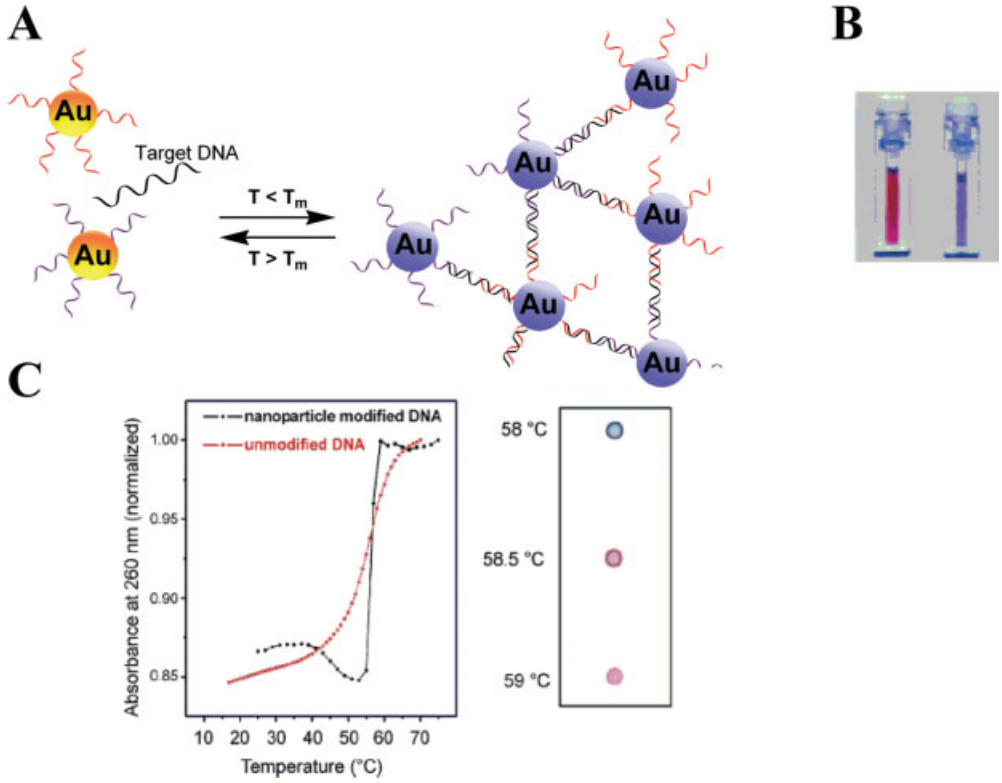
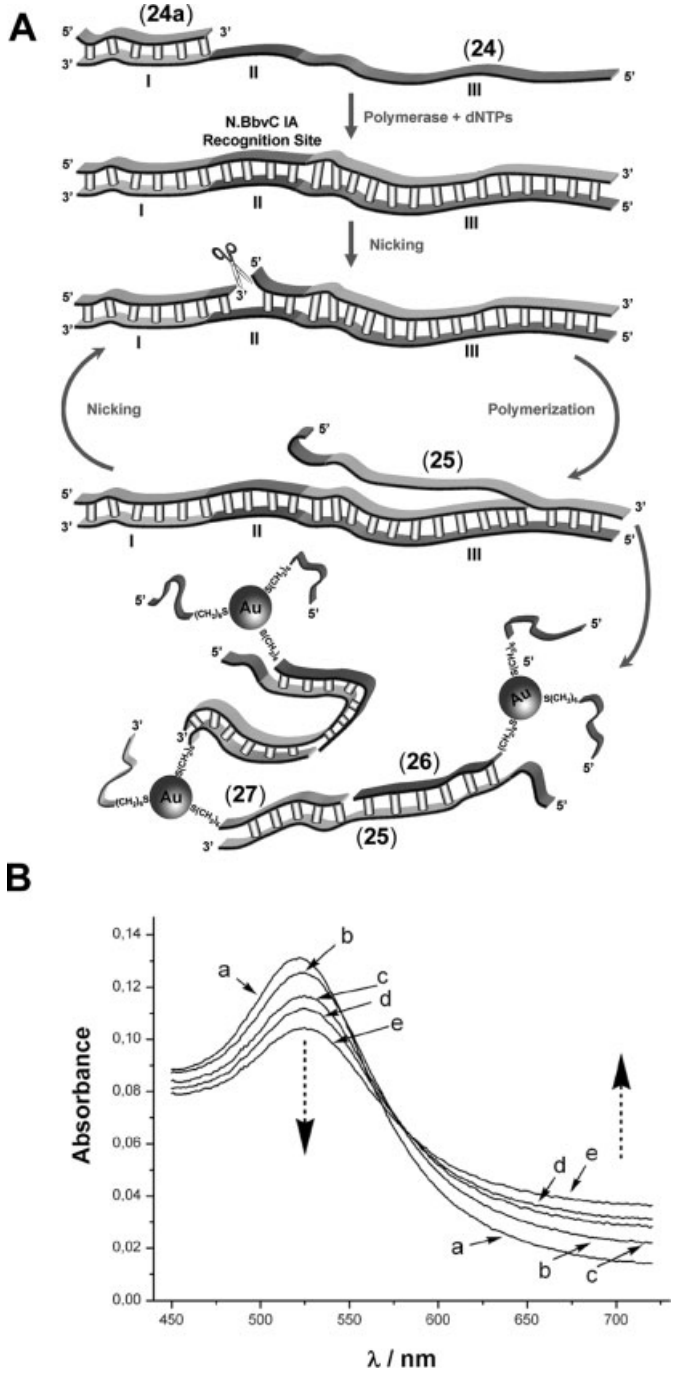


Figure 6.13 Optical detection of a target DNA through the hybridization of two kinds of nucleic acid-functionalized Au NPs complementary to the ends of the target DNA. The hybridization leads to the aggregation of the NPs (A) and to a red-to-purple color transition (B). The deaggregation of the nanoparticles is stimulated by the melting of the crosslinked DNA duplexes (C). Part (B) (Reprinted with permission from [15]. Copyright 2005 American Chemical Society). Part (C) (Reprinted from [87] with permission from AAAS).

surface density of the modifying nucleic acids, nanoparticle size, interparticle distances and salt concentration [91].

Recently, the use of DNA-based machines for the amplified detection of DNA was developed [92, 93]. The aggregation of Au NPs was used as a readout signal that follows the operation of the machine and the detection of the respective DNA [94] [Figure 6.14 (A)]. The machine consists of a nucleic acid "track" (**24**) that includes three domains, I, II and III. Domain I acts as the recognition site, and hybridization with the target DNA (**24a**) triggers, in the presence of polymerase and the dNTP mixture, the replication of the DNA track. Formation of the duplex, and specifically the formation of the duplex region II, generates the scission site for nicking enzyme Nb BbvC I. The enzyme-induced scission of the duplex activates the autonomous operation of the machine, where the replication and strand displacement of the complementary nucleic acid of region III proceed continuously. The displaced nucleic acid **25** may be considered as the "waste product". In the presence of two kinds of nucleic acids **26**- and **27**-


A
(24a)
(24)
N.BbvC IA
Recognition Site
Polymerase + dNTPs
Nicking
Nicking
Polymerization
(25)
(26)
(27)
(25)
Au
B
Absorbance
λ / nm
a
b
c
d
e

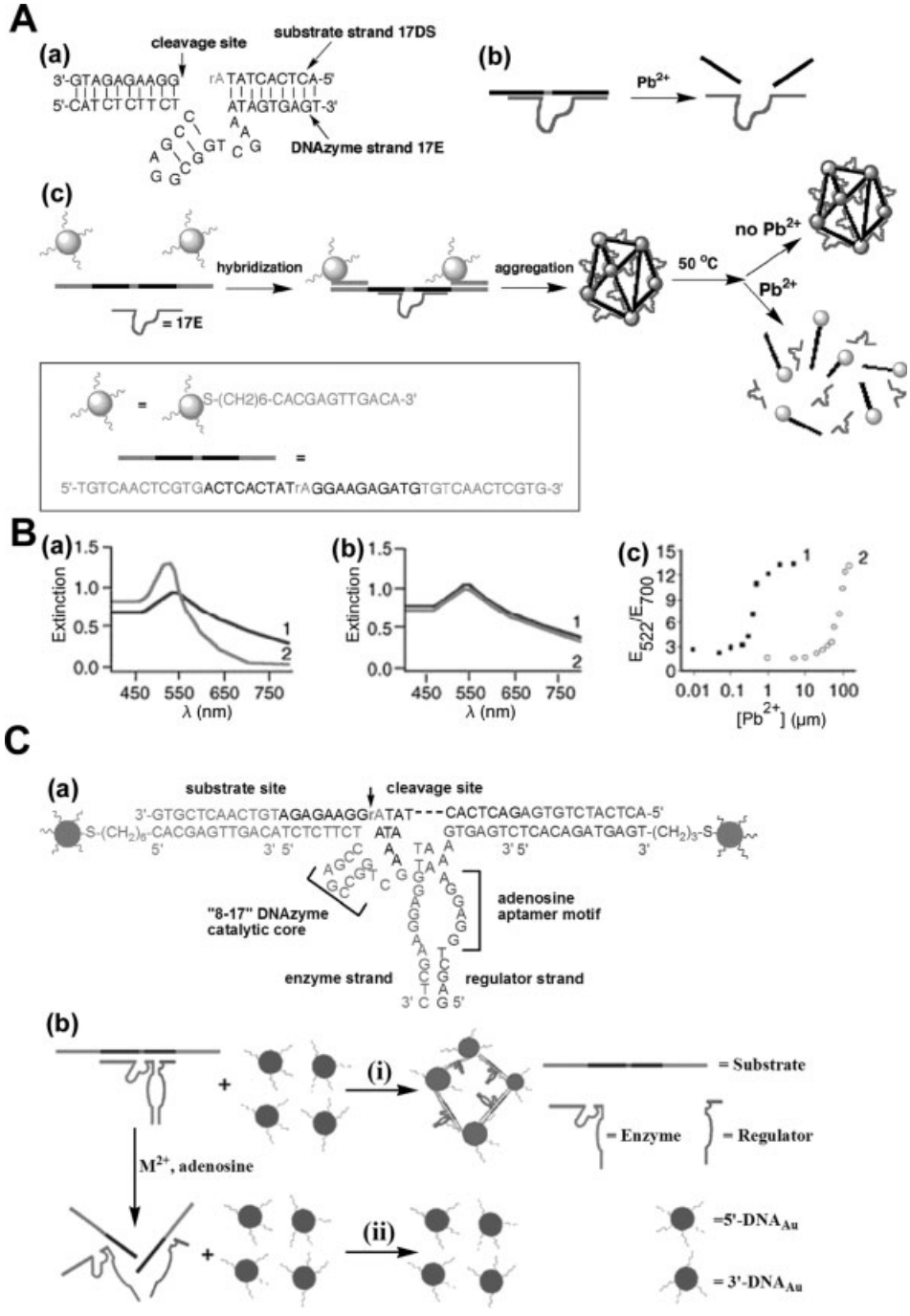
functionalized Au NPs, which are complementary to the two ends of the "waste product", aggregation of Au NPs proceeds. The color changes as readout method of aggregation of the Au NPs are depicted in Figure 6.14(B). The method allowed the optical detection of the target DNA with a sensitivity that corresponded to 1×10^{-12} M.

A different metal NP-induced analysis of DNA through the aggregation of the NPs employed the salt effect on the double layer potential of the NPs [95, 96]. Au NPs (15 nm) were functionalized with a probe nucleic acid, and the effect of the addition of the salt (up to 2.5 M) on the stability of the modified particles, as compared with unmodified (bare) NPs was examined. While the nucleic acid-functionalized Au NPs revealed stability in the presence of added salt, the nonfunctionalized Au NPs precipitated at a salt concentration of 0.1 M. The probe-functionalized Au NPs, when hybridized with the complementary target DNA, revealed a rapid red to purple color transition (<3 min) upon addition of 0.5 M NaCl. This color change was attributed to the lowering of the Au NP surface potential upon addition of the salt, which resulted in a decrease in the electrostatic repulsive interactions between the particles and consequently to shorter interparticle distances and a coupled plasmon absorbance.

The effect of salt on the stability of unmodified Au NP upon interaction with a probe nucleic acid before and after hybridization was used for the colorimetric detection of specific sequences in amplified genomic DNA [97, 98]. The method relied on the different effects of single- and double-stranded DNA on unmodified citrate-coated Au NPs. The adsorption of short ss-DNA probes on Au NPs stabilizes the citrate-coated NPs against salt-induced aggregation. The exposure of unmodified gold nanoparticles to a saline mixture containing amplified genomic DNA and short ss-DNA complementary to the regions in genomic DNA resulted in the aggregation of Au NPs and a color change from red to blue. If the short oligomers were not complementary to the regions in the genomic DNA, no color change occurred due to the stabilization of Au NPs by these oligomers. The method permitted the sequence-specific detection of label-free oligonucleotides at the level of 100 fmol and was adapted to detect single-base mismatches.

The hybridization-induced aggregation of metallic NPs was extended to analyze ions and small molecules using aptamers and DNAzymes. Aptamers are nucleic acids with specific recognition properties towards small molecules or proteins. They are prepared by the Systematic Evolution of Ligands by Exponential Enrichment (SELEX) procedure, which involves the selection and amplification of a sequence-specific nucleic acid to a target molecule from a library of 10^{15}–10^{16} nucleic acids [99–101]. Similarly, catalytic nucleic acids (DNAzymes or ribozymes) are prepared by eliciting

Figure 6.14 (A) Analysis of a DNA target (**24a**), by a nucleic acid "track" (**24**), consisting of regions I, II and III. The replication followed by nicking of the primary duplex (**24a**)/(**24**) yields displacement of the "waste product" (**25**). The displaced product (**25**) stimulates the aggregation of (**26**)- and (**27**)-functionalized Au NPs that are complementary to the two ends of (**25**). (B) Spectral changes upon analyzing different concentrations of the target DNA (**24a**) through the aggregation of the Au NPs by the "waste product" (**25**) generated by the DNA machine operated for a fixed time interval of 120 min: (a) 0, (b) 1×10^{-6}, (c) 1×10^{-7}, (d) 1×10^{-8} and (e) 1×10^{-9} M. (Reproduced with permission from [94]. Copyright 2007 Wiley-VCH).


A
(a)
cleavage site
substrate strand 17DS
3'-GTAGAGAAGG
5'-CATCTCTTCT
rATATCACTCA-5'
ATAGTGAGT-3'
DNAzyme strand 17E
(b)
Pb2+
(c)
hybridization
aggregation
50 °C
no Pb2+
Pb2+
= 17E
= S-(CH2)6-CACGAGTTGACA-3'
=
5'-TGTCAACTCGTGACTCACTATrAGGAAGAGATGTGTCAACTCGTG-3'
B
(a)
Extinction
λ (nm)
(b)
Extinction
λ (nm)
(c)
E522/E700
[Pb2+] (μm)
C
(a)
substrate site
cleavage site
3'-GTGCTCAACTGTAGAGAAGGrATAT---CACTCAGAGTGTCTACTCA-5'
S-(CH2)6-CACGAGTTGACATCTCTTCT ATA GTGAGTCTCACAGATGAGT-(CH2)3-S
"8-17" DNAzyme catalytic core
adenosine aptamer motif
enzyme strand
regulator strand
(b)
(i)
(ii)
M2+, adenosine
= Substrate
= Enzyme
= Regulator
=5'-DNAAu
= 3'-DNAAu

nucleic acids by *in vitro* selection towards transition-state analogues of chemical reactions [102] or by the selection of nucleic acids with binding affinities to metal ions [103–105] or active site analogues such as a heme [106, 107]. For example, the "8–17" DNAzyme has demonstrated high activity and specificity towards Pb^{2+} ions and revealed catalytic activity towards the specific scission of the complementary nucleic acid that included the respective cleavage site [108, 109]. Au NPs were used as optical labels for the detection of Pb^{2+} ions based on the activity of the Pb^{2+}-dependent DNAzyme that effects the separation of Au NP aggregates [110] (Figure 6.15). The method used the "8–17" DNAzyme that reveals high activity and specificity toward Pb^{2+} ions. The system consists of a substrate strand **17**DS that includes the cleavage site and is complementary to enzyme strand **17**E, which recognizes Pb^{2+} ions. The enzyme strand revealed catalytic activity in the presence of Pb^{2+} ions and resulted in scission of the substrate strand at the cleavage site. The complementary substrate strand of the DNAzyme was elongated with nucleic acid residues that are complementary to secondary nucleic acids labeled with Au NPs [Figure 6.15(A)]. Mixing of all three components of the DNAzyme – the enzyme strand, the complementary nucleic acid substrate and the nucleic acid-labeled Au NPs – resulted in the hybridization of the respective components, the aggregation of the Au NPs [Figure 6.15(A), part (c)] and the appearance of blue color, which indicated interparticle-coupled plasmon absorbance [Figure 6.15(B), part (a), curve 1]. The addition of Pb^{2+} ions resulted in the DNAzyme-assisted cleavage of the substrate strand, and this led to the separation of the Au NP aggregates and to a red color corresponding to the deaggregated Au NPs [Figure 6.15(B), part (a), curve 2]. This method allowed the colorimetric detection of Pb^{2+} ions in the concentration range 0.1–4 μM [Figure 6.15(B), part (c), curve 1]. The function of DNAzyme as an active component for the detection of Pb^{2+} ions was confirmed by using an inactive DNAzyme (**17**Ec) with poor scission activity of the substrate strand in the presence of Pb^{2+} ions [Figure 6.15(B), part (b), curves 1 and 2] and with low sensitivity [Figure 6.15(B), part (c), curve 2].

Figure 6.15 (A) DNAzyme system for the analysis of metal ions. (a) Secondary structure of the "8–17" DNAzyme system that consists of the enzyme strand **17**E and a substrate strand **17**DS. Except for a ribonucleoside adenosine (rA) at the cleavage site, all other nucleosides are deoxyribonucleosides; (b) cleavage of **17**DS by **17**E in the presence of Pb^{2+} ions; (c) DNAzyme-directed assembly of the oligonucleotide-functionalized Au NPs and the application of the assemblies for Pb^{2+} sensing. (B) UV–visible absorbance spectra of (a) the active **17**E DNAzyme–NP sensor and (b) an inactive **17**Ec DNAzyme–NP sensor; the spectra were recorded in the absence (curve 1) or presence (curve 2) of Pb^{2+} ions (5 μM); (c) calibration plots for the analysis of Pb^{2+} ions, in which the enzyme strand is the active **17**E only (curve 1) and when the **17**E:**17**Ec ratio is 1:20 (curve 2). (C) Colorimetric detection of adenosine monophosphate by a DNAzyme–aptamer conjugate, through the deaggregation of Au NPs aggregates: (a) the construct of the blocked Pb^{2+}-stimulated nucleotide cleaving DNAzyme tethered to the anti-adenosine aptamer that bridges Au NPs; (b) schematic colorimetric detection of adenosine by the DNAzyme–aptamer construct: (i) aggregated structure of blocked DNAzyme-aptamer construct; (ii) adenosine-induced release of the aptamer unit from the blocked construct followed by the activation of the DNAzyme, cleavage of the bridging units and deaggregation of the NPs. (Parts A and B reprinted with permission from [110]. Copyright 2003 American Chemical Society. Part C reprinted with permission from [111]. Copyright 2004 American Chemical Society).

The DNAzyme-controlled aggregation of Au NPs was expanded to a broader range of analytes, and a colorimetric biosensor for adenosine was tailored on the basis of the aptazyme-directed assembly of gold NPs [111]. The aptazyme is based on an "8–17" DNAzyme that is allosterically activated by the adenosine aptamer [Figure 6.15(C), part (a)]. In the absence of adenosine, an inactive aptazyme is formed, and the substrate strand operates as a linker for the assembly of a blue colored aggregate of 13-nm Au NPs [Figure 6.15(C), part (b), route (i)]. In the presence of adenosine, however, the aptazyme is activated and the substrate strand is cleaved [Figure 6.15(C), part (b), route (ii)]. This prevents the aggregation of the Au NPs, which results in a red-colored system characteristic of the individual nonaggregated nanoparticles. Concentrations of up to 1 mM of adenosine could be semiquantitatively analyzed by the extent of blue-to-red color changes or quantitatively measured by the ratio of the absorbance values at 520 and 700 nm. The addition of guanosine, cytidine or uridine (all 5 mM) instead of adenosine did not affect the aggregation of the DNA-bound Au NPs, indicating the specificity of the system towards adenosine.

Aptamer- and oligonucleotide-induced aggregation processes of Au NPs were used to develop colorimetric sensors for low molecular weight substrates such as adenosine and cocaine [112]; for example, a nucleic acid strand (**28**) that includes the specific anti-adenosine aptamer sequence and tether units I and II complementary to two kinds of **29**- and **30**-functionalized Au NPs. The mixture of (**28**) with the functionalized Au NPs resulted in the aggregation of the Au NPs through bridging of the NPs by hybridization of **28** with the functionalized NPs (Figure 6.16). In the presence of added adenosine, the adenosine–aptamer complex folds to a configuration that destabilizes the double-stranded structure, resulting in deaggregation of the particles and a purple-to-red color transition. The method allowed the detection of adenosine in the concentration range 0.3–2 mM. The generality of this method was further demonstrated by the construction of a colorimetric sensor for cocaine based on a specific cocaine aptamer [112].

The aggregation of the Au NPs is not limited to DNA detection and the process was applied to develop optical biosensing assays for sensing enzyme activities. For example, the detection of proteases (thrombin and letal factor) by the enzyme-induced cleavage of peptides was reported [113]. In this protocol, acetylated cysteine residues were added to two termini of the peptide substrate. As the cysteine units bind to Au NPs, the intact peptide bridged the NPs, resulting in their aggregation and the formation of a violet–blue color. The pretreatment of the modified peptide with target protease led to the cleavage of the peptide to the monofunctionalized cysteine residues that did not lead to the aggregation of the NPs. The assay demonstrated high sensitivity and it allowed the sensing of thrombin with a detection limit as low as 5 nM and the detection of letal factor with a detection limit of 25 nM. An additional method for the detection of proteases (thermolysin and nACT–PSA) activity was described [114]. This method is based on the protease-induced deaggregation of Au NP clusters crosslinked by π-stacking of the Fmoc residues linked to the peptide coating of the Au NPs. The cleavage of the peptide by target protease resulted in deaggregation of NPs and a color change from blue to red. Similarly, the activity of alkaline phosphatase (ALP) was monitored through the aggregation of Au NPs [115].

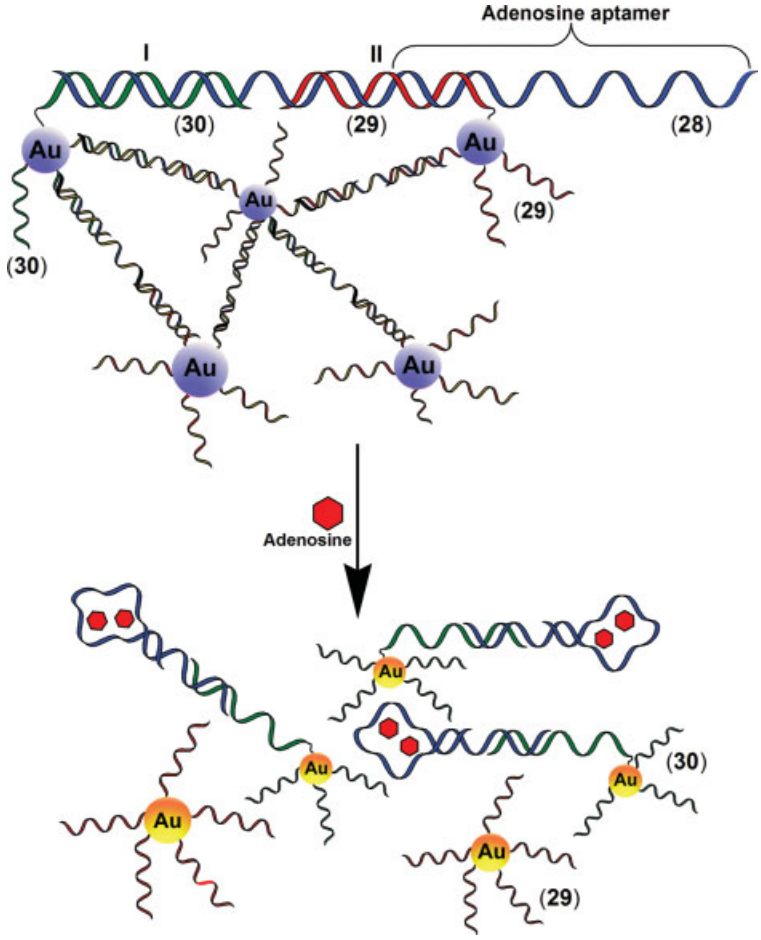


Figure 6.16 Colorimetric analysis of adenosine monophosphate through the deaggregation of Au NPs bridged by the blocked adenosine aptamer tethered to a nucleic acid (**28**), that is hybridized with nucleic acids (**29**)- and (**30**)-functionalized Au NPs. Formation of the aptamer–substrate complex separates the (**29**)-modified Au NPs and stimulates the deaggregation of the NPs.

The phosphorylated peptide H_2N-Cys-Tyr(PO_3^{2-})-Arg-OH was assembled on Au NPs through the cysteine group. Electrostatic repulsion of the negatively charged Au NPs retained the nonaggregated configuration of the NPs. The hydrolytic dephosphorylation of the peptide by ALP removed the phosphate group, thus permitting the bridging and aggregation of the NPs by the free amino group on the peptide chains.

The aggregation of Au NPs as colorimetric test for biorecognition events was extended to numerous other biorecognition complexes. For example, the process was used to detect the cholera toxin [116]. Au NPs were modified with a lactose capping

layer, and the cholera toxin (B-subunit) linked to the lactose derivative induced aggregation of the Au NPs. Upon aggregation, the color of solution of Au NPs changed from red to deep purple. The selectivity of the bioassay arises from the fact that thiolated lactose mimics the GM1 ganglioside, the native receptor of the cholera toxin. The detection limit of the assay was 54 nM. Similarly, the colorimetric sensing of platelet-derived growth factors (PDGFs) and their receptors (PDGFRs) based on aggregation of aptamer-functionalized Au NPs was also developed [117].

In addition to the absorbance features of metallic NPs that were used to follow biorecognition events in solution and on surfaces, other optical methods have been employed to detect the association of biomolecule-functionalized Au NPs on biochips. These methods included scanometric detection by light scattering, surface plasmon resonance spectroscopy, resonance-enhanced absorption by NPs and enhanced Raman scattering.

A scanometric DNA detection method was developed, and this was based on a sandwich assay format involving a DNA-functionalized glass slide, the target DNA and Au NP probes [118]. In a typical setup for scanometric detection, the modified glass slide was illuminated in the plane of the slide with white light. The slide served in such a configuration as a planar waveguide that prevents any light from reaching the microscope objective by total internal reflectance. Wherever NP probes were attached to the surface, evanescently coupled light was scattered from the slide, and the NP labels were imaged as bright, colored spots. This approach was used for the detection of target DNA molecules that were specifically bound to a DNA-functionalized surface. The resulting DNA hybrid was labeled with gold nanoparticles that allowed the scanometric detection of the DNA (Figure 6.17). At high target concentrations ($\geq$1 nM), the Au NPs on the surface could be visualized with naked eye. At low target concentrations ($\leq$100 pM), the coverage of the surface-bound Au NPs was too low, and an enhancement process was needed. Enlargement of the Au NPs by the catalytic reduction of silver ions and the deposition of silver metal on the Au NPs

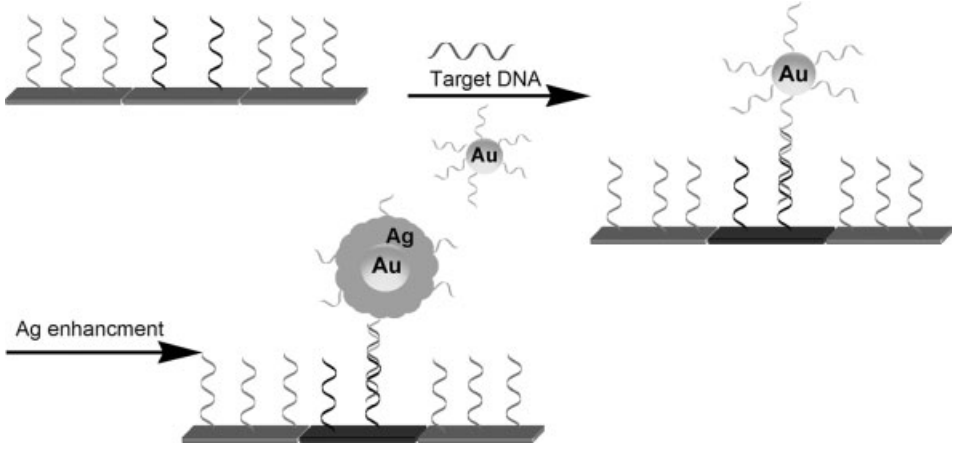


Figure 6.17 Scanometric detection of DNA on a surface using gold–silver core–shell NPs.

resulted in a 100-fold increase in the light-scattered signal and thus increased the sensitivity detection of target DNA (50 fM) [118]. This method was used to detect single-base mismatches in oligonucleotides, which were hybridized to DNA probes and were immobilized at different domains of a glass support. High sensitivities were provided by the deposition of silver, whereas the selectivity was achieved by examination of the melting properties of the spots: the mismatched spot reveals a lower melt temperature owing to its lower association constant. The scattering of light is size-dependent, and hence, by using different sized nanoparticles the simultaneous detection of different DNA sequences is feasible. Accordingly, the light scattered by DNA-functionalized 50- and 100-nm Au NP probes was used to identify two different target DNAs in solution [119]. The scanometric method was successfully applied to detect the MTHFR gene from genomic DNA at concentration as low as 200 fM without PCR amplification of the target, by the application of improved optical imaging instruments [120]. A similar approach was used to identify single nucleotide polymorphisms (SNPs) in unamplified human genomic DNA samples representing all possible genotypes for three genes involved in thrombotic disorders [121].

The scanometric method was applied as an optical detection means in a series of systems that employed nucleic acid-functionalized Au NPs as barcodes for biorecognition events such as antigen–antibody complex formation or DNA hybridization. In one system (Figure 6.18), prostate-specific antigen (PSA) was detected by nucleic acid-functionalized Au NPs that acted as signaling barcodes [122]. The Au NPs were modified with the polyclonal anti-PSA Ab that was further functionalized with thiolated nucleic acid (**31**), which were hybridized with the complementary nucleic acid (**31′**), and its sequence acted as a barcode for the sensing process. In the presence of PSA and magnetic particles functionalized with the monoclonal anti-PSA Ab, an aggregate consisting of the "sandwich" structure of the Au NPs and the magnetic particles was formed. The magnetic separation of the aggregate was followed by the thermal displacement of the barcode DNA. The released barcode nucleic acid was then amplified by the polymerase chain reaction (PCR), and the product was used to bridge the Au NPs with the complementary nucleic acid to a glass surface. The Ag-enhanced Au NPs were then analyzed by the scanometric method. Although this analytical protocol involves many steps, the PCR amplification step leads to an ultrasensitive detection method, and PSA at a level of 30 aM was analyzed.

An analogous process was used to analyze DNA (Figure 6.19) [123]. By this method, the functionalized Au NPs include the duplex DNA barcode (**32/32′**) and the nucleic acid units (**33**) that recognize the target DNA. In the presence of the target DNA (**35**), the magnetic particles modified with the nucleic acids **34** and the **32/33**-functionalized Au NPs, the magnetic particle–Au NP aggregate is formed through crosslinking the particles by **35**. The subsequent magnetic separation of the aggregate and the thermal separation of the DNA barcode were followed by scanometric detection of the released DNA code on surfaces. A PCR-like sensitivity that corresponded to 500 zM was claimed for the analysis of DNA. The DNA barcode-based sensing protocols were used to analyze protein cancer markers [124], the amyloid biomarkers for Alzheimer's disease [125] and the genes of various

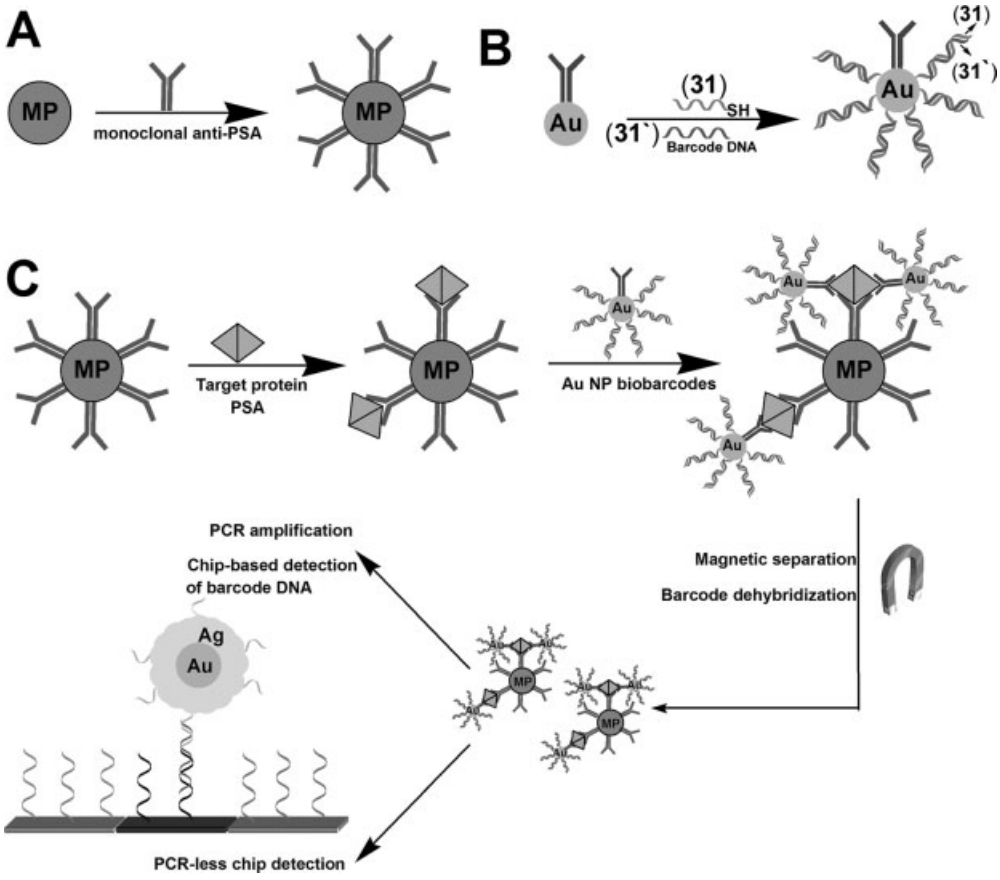


Figure 6.18 Scanometric detection of a target protein (PSA) by an immunoassay that amplifies the analysis by the PCR-induced generation of a DNA barcode: (A) preparation of the antibody-modified magnetic particles; (B) synthesis of antibody and duplex DNA (**31/31′**)-functionalized Au NPs; (C) formation of the DNA barcode-labeled Au NP immunocomplex on the magnetic particles, separation of the complex and the PCR amplification of the DNA barcode (**31′**). The amplified production is scanometrically detected on the surface.

pathogens, such as hepatitis B, Ebola virus, variola virus (smallpox, VV) and human immunodeficiency virus (HIV) [126]. Furthermore, different modifications of the method that use fluorescence detection [127] and colorimetric assay [128] have been reported.

The colorimetric scattering assay of gold nanoparticles was applied for the rapid detection of *mecA* gene in unamplified genomic DNA sequences [129]. The method was based on hybridization of DNA-functionalized Au NPs probes with the complementary sequences of target DNA in solution. The resulting solution was then spotted on a glass waveguide, which was illuminated with white light in the plane of the slide. The color of the light scattered by the oligonucleotide-functionalized Au NPs probes is different from that of the Au NPs aggregates generated by the hybridization process

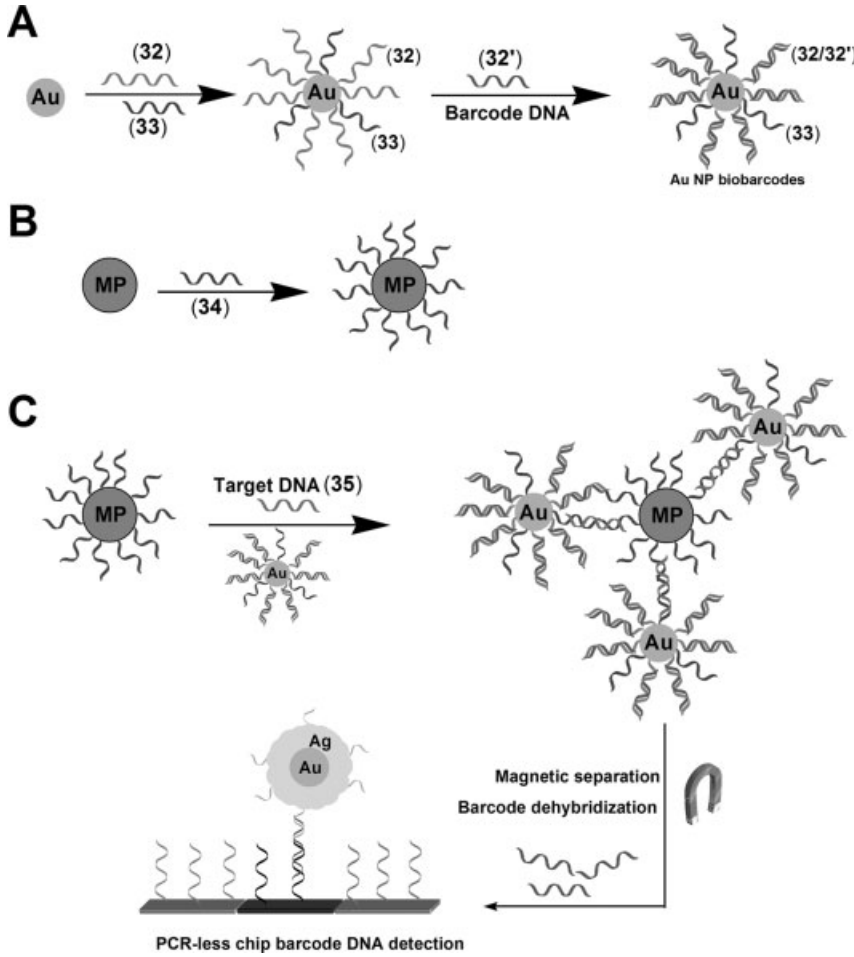


Figure 6.19 Scanometric detection of a DNA by a nucleic acid barcode-functionalized Au NP: (A) synthesis of Au NP labeled with the nucleic acid (**32**), complementary to the DNA barcode (**32′**) and the nucleic acid (**33**) complementary to the target; (B) preparation of the magnetic particles modified with the nucleic acid (**34**) complementary to the other end of the target; (C) aggregation of the functionalized Au NPs and the magnetic particles through hybridization with the target, magnetic separation of the aggregates, thermal separation of the DNA barcode and its scanometric detection.

between the probes and target DNA. In the absence of the target DNA, individual 40–50-nm Au NPs scatter green light, whereas in the presence of the target DNA aggregated nanoparticles scatter yellow to orange light. The method showed enhanced detection sensitivity (about four orders of magnitude) compared with the colorimetric assay, thus allowing the detection of zeptomole quantities of target DNA, 333 fM of synthetic DNA and 33 fM of genomic DNA. An improved light-scattering strategy using gold nanoparticles as labels for the detection of specific target DNA in a

homogeneous solution was reported [130]. The method is based on the aggregation of DNA-functionalized Au NPs through the hybridization of the target DNA with probe-labeled Au NPs and enhanced light scattering that originated from the aggregates in solution. The light-scattering assay demonstrated high sensitivity, and the human p53 gene, exon 8 DNA, was detected at a concentration as low as 0.1 pM. Moreover, the assay showed a high degree of specificity and was used for discrimination between perfectly matched targets and targets with single base-pair mismatches.

The hyper-Rayleigh scattering (HRS) technique (nonlinear light scattering) was used to monitor DNA hybridization on unmodified gold nanoparticles in a saline solution, and the target DNA was analyzed at a concentration of 10 nM [131]. The HRS assay for DNA detection was based on the differences in the electrostatic interactions between ssDNA and dsDNA with the particles. The method permitted the analysis of single-base mismatch in DNA by monitoring the HRS intensity from the different DNAs interacting with the gold nanoparticles.

Surface-enhanced Raman scattering (SERS) of nanoparticle-bound substrates allows the amplification of molecular vibrational spectra by up to 10^6-fold [132–134]. Modification of metal NPs with different Raman dyes was used to generate multiply coded NPs [135–138] and for the preparation of thousands of codes to be written and read by means of surface-enhanced Raman resonance (SERR) scattering without the need for spatial resolution of components of the code [135, 137]. The use of SERS for the analysis of biorecognition events was demonstrated with the application of Au NPs that were functionalized with Raman dyes and recognition elements [139, 140]. Formation of the complementary recognition complex on surfaces, followed by the electroless deposition of Ag on the Au NPs, allowed the enhanced readout of the biorecognition events by SERS. The concept was applied for the parallel detection of various analytes on surfaces in an array configuration [Figure 6.20(A)]. For example, six different thiol-functionalized Raman-active dyes, Cy3, Cy3.5, Cy5, TAMRA, Texas Red and Rhodamine 6G, were linked through an oligonucleotide spacer (10 adenosine units) to six different oligonucleotides and coupled to Au NPs (13 nm) to yield six different Raman dye-labeled Au NP probes [139] [Figure 6.20(B)]. These Au NP probes were then employed as labels for hybridization with the complementary targets. The capture DNA strands were spotted on a surface, and the specific binding of Au NPs by hybridization with target DNAs was followed by the silver enhancement of the Au NPs and analysis by SERS [Figure 6.20(C) and (D)]. The detection limit of this method was 20 fM. The assay demonstrated the ability to discriminate SNPs in DNA and RNA targets. A similar concept was also used to identify protein–protein and protein–small molecule interactions by using Au NPs that were functionalized with specific antibodies and encoded with specific Raman dyes [140]. Compared with colorimetric and scanometric detection methods, this method offers enhanced multiplex sensing capabilities afforded by the narrow spectral bands – fingerprints of Raman dyes. Further developments of the SERS technique allowed highly sensitive immunoassay procedures [141–143].

Au NPs have been widely employed for signal amplification of biorecognition events based on nanoparticle-enhanced SPR spectroscopy [82, 86]. The changes in the dielectric properties at thin films of metals, such as gold films, as a result of

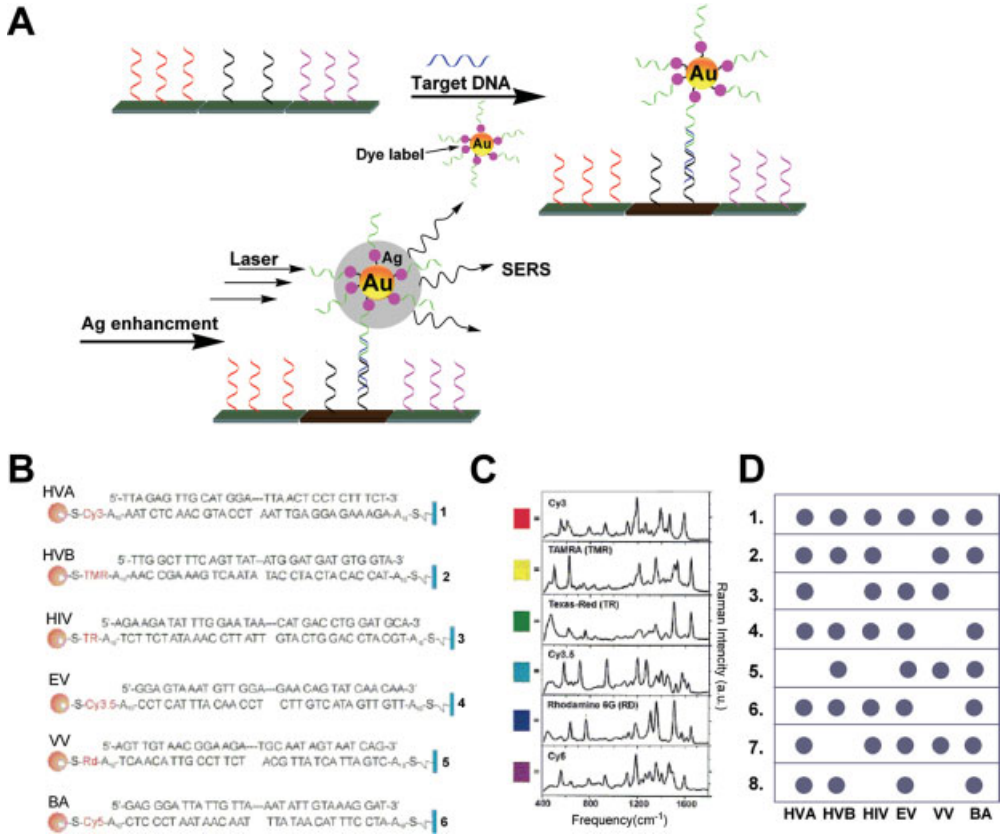


Figure 6.20 (A) Detection of a target DNA by the use of dye-functionalized Au NPs as labels for the SERS imaging. (B) Dye-labeled sequences of nucleic acids complementary to target DNAs of different pathogens. (C) Raman spectra of the dye-labeled Au NPs probes after silver enhancement upon the parallel analysis of the different pathogens on surfaces. (D) Flatbed scanner images of silver-enhanced microarrays upon the sensing of different pathogens. (Parts B, C and D reprinted from [139] with permission from AAAS).

biomolecular recognition processes, are the basis for the SPR technique. Labeling of the biorecognition complexes with Au or Ag NPs besides changing the dielectric properties, results in the electronic coupling between the localized NPs plasmon and the surface plasmon wave of the metal film. This electronic coupling significantly affects the resonance frequency of the surface wave, thus leading to the enhanced amplified optical transduction of the biorecognition events. Accordingly, Au NPs were used as labels in immunosensing [144–146] and DNA sensing [147–149] applications. The binding of Au NPs to the immunosensing interface led to a large shift in the plasmon angle, a broadening of the plasmon resonance region and an increase in the minimum reflectance, and these effects allowed the detection of the antigen with picomolar sensitivities [144]. Similarly, the sensitivity of DNA analysis was enhanced 1000-fold (10 pM) when Au NP-functionalized DNA molecules

were used as labels [147]. Also, the sensitive detection of DNA hybridization by applying catalytic growth of Au NPs as a means to enhance the SPR shifts was demonstrated [150].

The charging of Au NPs that are coupled to thin gold films affects the electronic coupling between the localized NP plasmon and the surface plasmon wave, resulting in a shift in the SPR spectra. Thus, biocatalytic electron transfer reactions that involved NP–enzyme conjugates may be monitored by SPR spectroscopy [151]. Au NPs (1.4 nm) were functionalized with N^6-(2-aminoethyl) flavin adenine dinucleotide (FAD cofactor, amine derivative; (**1**), and apo-glucose oxidase (apo-GOx) was reconstituted onto the cofactor sites. The nanoparticle–GOx conjugates were then assembled on a gold thin film (SPR electrode) by using a long-chain dithiol, HS $(CH_2)_9SH$, monolayer as a bridging linker. This yielded the biocatalytically active glucose oxidase (GOx) bound to the Au NPs in an aligned configuration [Figure 6.21 (A)]. The biocatalyzed oxidation of glucose resulted in the formation of the reduced form of the cofactor, $FADH_2$. In the absence of O_2 that acts as the natural electron acceptor for GOx, electron transfer proceeded from the reduced cofactor to the Au

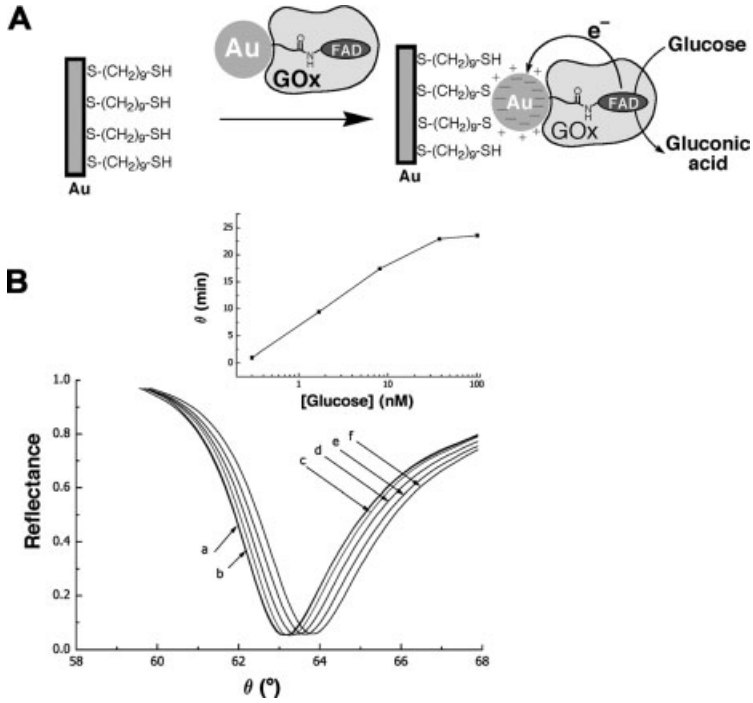


Figure 6.21 (A) Assembly of Au NP-bound reconstituted glucose oxidase (GOx) on a dithiol monolayer that is associated with an SPR-active surface and biocatalytic charging of the Au NPs in the presence of glucose. (B) SPR spectra of the Au NP–GOx hybrid system upon the addition of various concentrations of glucose: (a) 0, (b) 0.3, (c) 1.6, (d) 8, (e) 40 and (f) 100 mM. Inset: calibration plot of the SPR spectra minimum shift as a function of glucose concentration. (Reprinted with permission from [151]. Copyright 2004 American Chemical Society).

NPs, resulting in their charging. The long-chain dithiol monolayer provided a barrier for the electron tunneling from the Au NPs to the bulk Au electrode, and this preserved the electrical charge that was produced on the Au NPs. Potentiometric measurements indicated that the long-chain dithiol linkers provided a resistance layer that generated a potential gradient between the charged Au NPs and the conductive support. This potential gradient originated from the charging of the Au NPs by the biocatalytic process and thus, charge accumulated on the NPs was controlled by the biocatalytic process (the concentration of glucose). Accordingly, the plasmon coupling between the charged Au NPs and the Au support resulted in a shift of the SPR spectra, and these were dependent on the charge generated on the NPs [Figure 6.21(B)]. As the charge value was controlled by the rate of the biocatalytic reaction, the shift in the SPR spectrum was enhanced upon elevation of the glucose concentration [Figure 6.21(B), inset].

Metal nanoparticles exhibit a strong UV–visible absorption band that is not present in bulk material [1, 2, 152–154]. This absorption band, known as the localized surface plasmon resonance (LSPR), occurs when the incident photon frequency matches the collective oscillations of the conduction electrons. It is well established that the maximum extinction wavelength, λ_{max}, of the LSPR is controlled by the composition, size, shape and interparticle spacing of the nanoparticles, and also by the dielectric properties of their local environment (i.e., substrate, solvent and surface-bound molecules) [82]. The sensitivity of λ_{max} of the LSPR to the molecular environment of the NPs allowed the development of a new class of optical biosensors [155, 156] that operate in an analogous manner to their SPR counterparts by transducing small changes in the refractive index near the noble metal surface into a measurable wavelength-shift response. Triangular silver nanoparticles, ~100 nm wide and 25 nm high, fabricated by the nanosphere lithography (NSL) technique [Figure 6.22(A)], showed very unique optical properties [157, 158]. In particular, the λ_{max} of their LSPR spectrum is unexpectedly sensitive to the size, shape and local external dielectric environment of the nanoparticles (10–40 nm shift). The Ag nanotriangles were used to follow the streptavidin–biotin interactions as a model system for LSPR-based biosensors [157]. Triangular silver NPs were functionalized with biotin units and the changes in the LSPR spectra were followed upon binding of streptavidin molecules. The detection by LSPR was further amplified by the secondary coupling of biotinylated Au NPs to the streptavidin-saturated interface. A similar method was applied to design an immunosensor [158]. In addition, LSPR spectroscopy was applied for the detection and diagnosis of biomarkers, for example, the amyloid-β-derived diffusible ligands (ADDLs) that are markers for Alzheimer's disease [159, 160]. The sandwich assay composed of antibody/ADDLs/antibody was developed to amplify the LSPR response, and thus to improve the detection limit [160] [Figure 6.22(B)]. The LSPR nanosensor demonstrated sensitivity and selectivity in the detection of ultralow concentrations of ADDLs in synthetic and human samples [human brain extracts, cerebrospinal fluid (CSF)] [Figure 6.22(C)]. A detection limit corresponding to 100 fM was achieved for synthetic ADDLs. Also, modified LSPR-based biosensors consisting of gold-capped silica nanoparticle-layered substrates for monitoring DNA hybridization and antigen–antibody interactions were fabricated [161, 162].

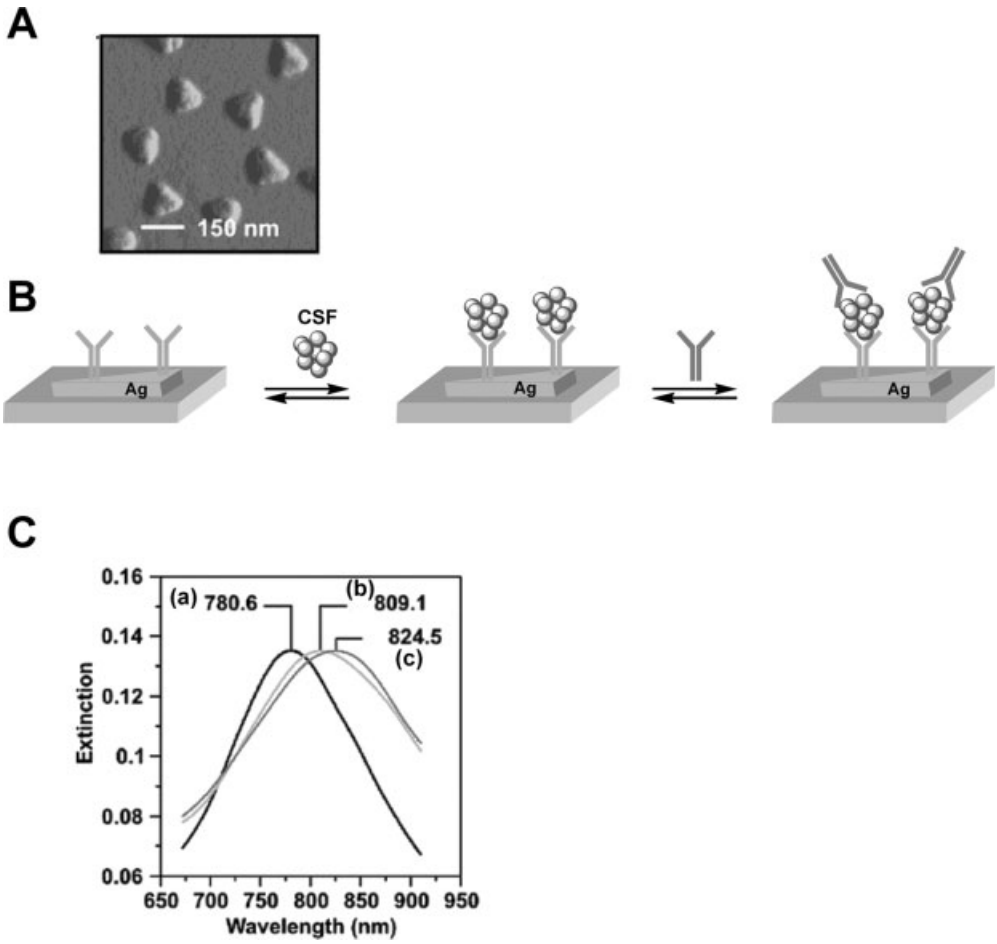


Figure 6.22 (A) AFM image of Ag^0-triangle NPs assembled on a mica surface. (B) Schematic analysis of the amyloid-β-derived diffusible ligands (ADDLs) by an immunoassay on Ag triangle NPs using LSPR. (C) LSPR curves corresponding to: (a) the capture anti-ADDL antibody-modified NPs; (b) the modified surface after treatment with the antigen (ADDL, from CSF); (c) the antibody–antigen complex after interaction with the secondary anti-ADDL antibody. (Reprinted with the permission from [160]. Copyright 2005 American Chemical Society).

6.7
Semiconductor Nanoparticles as Optical Labels

The unique optical properties of semiconductor nanoparticles or quantum dots, QDs, offer a number of advantages for biosensing, including size-controlled luminescence properties, impressive photostability and high quantum yields [5–7, 163, 164]. Efficient methods for the preparation of semiconductor nanoparticles

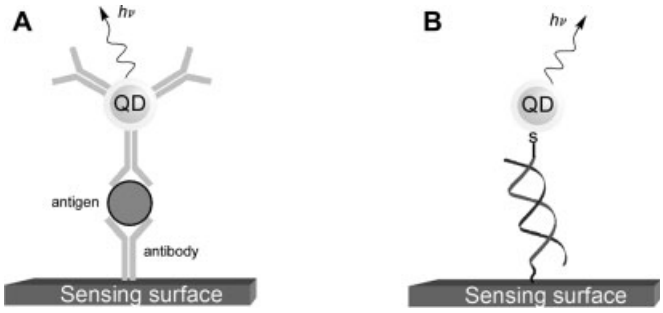


Figure 6.23 Fluorescent analysis of (A) an antigen by antibody-functionalized QDs and (B) a DNA by a nucleic acid-functionalized QD.

and their functionalization with biomolecules were developed recently [165, 166]. Unlike molecular fluorophores, which typically have very narrow excitation spectra, semiconductor QDs absorb light over a very broad spectral range. This makes it possible to excite optically a broad spectrum of quantum dot "colors" using a single excitation laser wavelength, which may enable one to probe simultaneously several markers in biosensing and assay applications. Indeed, functionalized semiconductor QDs have been used as fluorescence labels for numerous biorecognition events [9, 167–170], including their use in immunoassays for protein detection [Figure 6.23(A)] and nucleic acid detection [Figure 6.23(B)]. For example, CdSe/ZnS QDs were functionalized with avidin and these were used as fluorescent labels for biotinylated antibodies. Fluoroimmunoassays utilizing these antibody-conjugated NPs were successfully used in the detection of protein toxins (staphylococcal enterotoxin B and cholera toxin) [171, 172].

Similarly, CdSe/ZnS QDs conjugated to appropriate antibodies were applied for the multiplexed fluoroimmunoassay of toxins [Figure 6.24(A)]. Sandwich immunoassays for the simultaneous detection of the four toxins (cholera toxin, (CT), ricin, shiga-like toxin 1 (SLT), and staphylococcal enterotoxin B, (SEB)) by using different sized QDs were performed in single wells of a microtiter plate in the presence of mixture of all four QD–antibody conjugates [Figure 6.24(B)] [173], thus leading to the fluorescence that encodes for the toxin. In another example, multiplexed immunoassay formats based on antibody-functionalized QDs were used for simultaneous detection of *Escherichia coli* O157:H7 and *Salmonella typhimurium* bacteria [174] and for the discrimination between diphtheria toxin and tetanus toxin proteins [175].

Fluorescent QDs were used for the detection of single-nucleotide polymorphism in human oncogene p53 and for the multiallele detection of the hepatitis B and hepatitis C virus in microarray configurations [176]. DNA-functionalized CdSe/ZnS QDs of different sizes were used to probe hepatitis B and C genotypes in the presence of a background of human genes. The discrimination of a perfectly matched sequence of p53 gene in the presence of background oligonucleotides including different single-nucleotide polymorphism sequences was detected with true-to-false signal ratios higher than 10 (under stringent buffer conditions) at room temperature within minutes. Also, DNA–QD conjugates were used as fluorescence probes for *in situ*

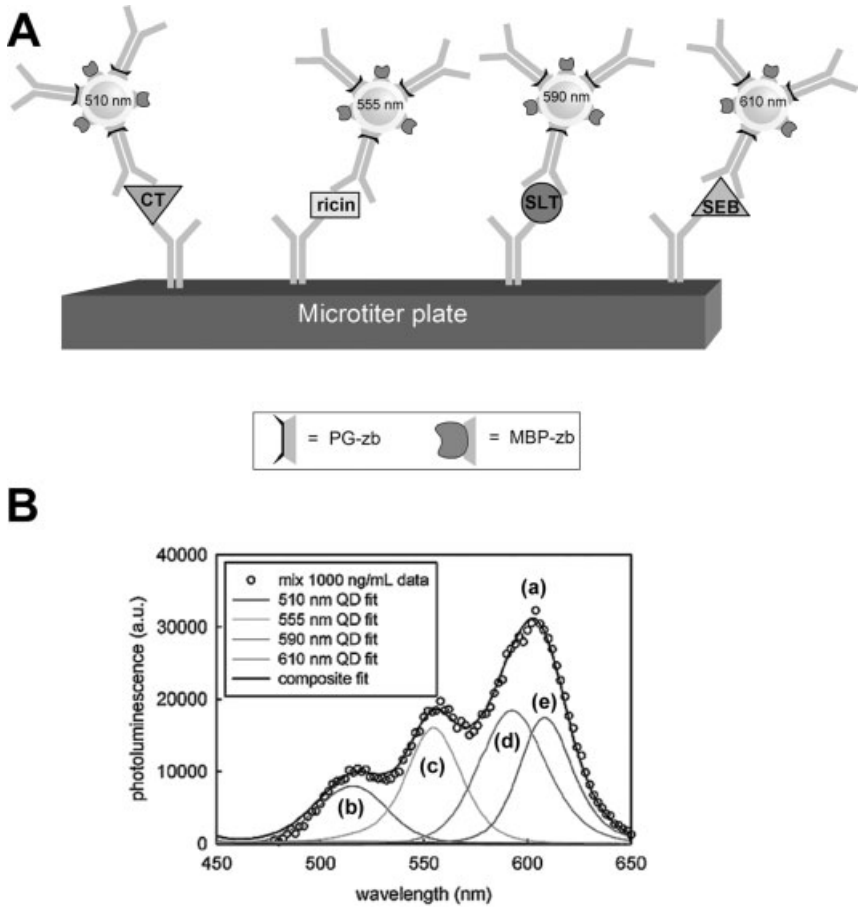


Figure 6.24 (A) Parallel optical analysis of different antigens in a well-array format using the fluorescence of different sized quantum dots. (B) Fluorescence spectrum observed upon analyzing the four analytes, 1 μg mL^{-1}, by the different sized QDs (a), and deconvoluted spectra of individual toxins: (b) CT, (c) ricin, (d) SLT and (e) SEB. (Reprinted with permission from [173]. Copyright 2004 American Chemical Society).

hybridization assays (FISH). For example, the QD-based FISH labeling method was used for the detection of Y chromosome in human sperm cells [177]. A QD-based FISH method to analyze human metaphase chromosomes was reported [178] by using QD-conjugated total genomic DNA as a probe for the detection of EBRB2/HER2/neu gene. Also, the FISH technique was used for the multiplex cellular detection of different mRNA targets [179].

Nonetheless, the superior photophysical features of semiconductor QDs are demonstrated in organic solvents, and their introduction into aqueous media, where biorecognition and biocatalytic reactions proceed, is accompanied by a severe or even complete loss of their fluorescence properties. Different methods to stabilize the

fluorescence properties of semiconductor QDs in aqueous media were reported, including their surface passivation with protective layers [180, 181] and the coating of the QDs with protecting glass films [182, 183] or polymers [184]. Although these methods preserve the photophysical properties of the QDs, the protective layers limit important useful applications of the QDs. Quantum dots could be employed as optical labels for dynamic biological processes such as biocatalyzed transformations or structurally induced biomolecular changes, for example, opening of a hairpin nucleic acid by the hybridization of DNA, using fluorescence resonance energy transfer (FRET) or electron transfer quenching as photophysical probe mechanisms [185, 186]. The sensitivity of these photophysical processes to the distance separating the donor–acceptor or chromophore–quencher pairs prevents, however, the use of fluorescent QDs passivated by relatively thick protecting layers as optical probes for dynamic bioprocesses. Hence, a very delicate nanostructuring of the capping layer is essential to allow the use of QDs as active components in energy/electron transfer reactions.

Different sensing schemes have been developed that use QDs as a FRET donor. For example, CdSe/ZnS QDs conjugated to nucleic acids have been used to follow the biocatalyzed replication of DNA [187]. CdSe/ZnS NPs were functionalized with the DNA primer **36**, which is complementary to a domain of M13mp18 DNA. Hybridization of the M13mp18 DNA with the nucleic acid-functionalized QDs, followed by the replication of the assembly in the presence of polymerase and the nucleotide (dNTP) mixture that included Texas Red-functionalized dUTP (**37**), resulted in the incorporation of the dye labels into the DNA replica [Figure 6.25(A)]. The FRET process from the semiconductor NPs to the incorporated dye units resulted in emission from the dye with concomitant quenching of the fluorescence of the QDs [Figure 6.25(B)], and it allowed the identification of the primary hybridization.

A similar approach was used to follow telomerase activity, a biocatalytic ribonucleoprotein that is a versatile marker for cancer cells. The CdSe/ZnS QDs were modified with a nucleic acid primer that is recognized by telomerase. In the presence of telomerase and the nucleotide mixture dNTPs, that included Texas Red-functionalized dUTP (**37**), the telomerization of the nucleic acid associated with the QDs was initiated, while incorporating the Texas Red-labeled nucleotide into the telomers [Figure 6.25(C)]. The FRET process from the QDs to the dye units then enabled the dynamics of the telomerization process to be followed [187].

Also, the association of maltose with the hybrid composed of maltose-binding protein was examined by the application of a CdSe/ZnS QD linked to the maltose binding protein (MBP) [188]. CdSe/ZnS QDs were functionalized with MBP, and these were interacted with a β-cyclodextrin–QSY-9 dye conjugate [Figure 6.26(A)]. The β-cyclodextrin–QSY-9 dye conjugate resulted in quenching of the luminescence of the QDs by the dye units. Addition of maltose displaced the quencher units, and this regenerated the luminescence of the QDs [Figure 6.26(B)]. This method allowed the development of a competitive QD-based sensor for maltose in solution. Similarly, a competitive QD-based assay for the detection of the explosive trinitrotoluene (TNT) was developed [189]. CdSe/ZnS QDs were functionalized with a single-chain antibody fragment that selectively binds TNT. The analogue substrate trinitrobenzene (TNB) covalently linked to the quencher dye BHQ10 was bound to the QD–antibody conjugate and quenched the QD fluorescence. In the presence of the TNT analyte, the quencher

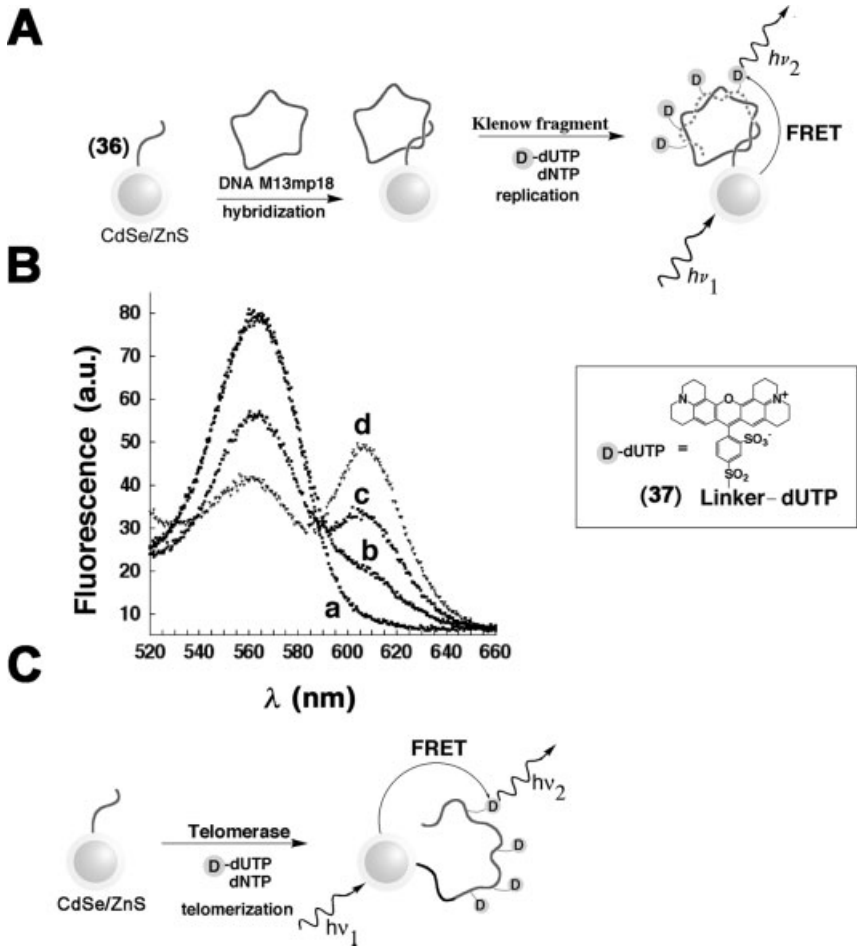


Figure 6.25 (A) Optical detection of M13 phage DNA by nucleic acid-functionalized CdSe/ZnS QDs. The replication of the analyte in the presence of the dNTP mixture that includes the Texas Red-labeled dUTP (**37**) results in the incorporation of the dye into the replica and the stimulates of a FRET process. (B) Time-dependent fluorescence changes upon incorporation of the dye (**37**) into the DNA replica and the analysis of the M13 phage DNA according to (A). (C) Optical analysis of telomerase activity by the incorporation of the Texas Red-dUTP (**37**) into the telomers associated with CdSe/ZnS QDs. (Reprinted with permission from [187]. Copyright 2003 American Chemical Society).

TNB–BHQ10 conjugate was competitively displaced. This eliminated the FRET interactions between the QD and the dye, and the fluorescence of the QDs was restored.

The hydrolytic functions of a series of proteolytic enzymes were followed by the application of QDs modified with peptides as reporter units and the FRET process as a readout mechanism [190, 191]. CdSe QDs were modified with different peptides that included specific cleavage sequences for different proteases, and quencher units were tethered to the peptide termini. The fluorescence of the QDs was quenched in

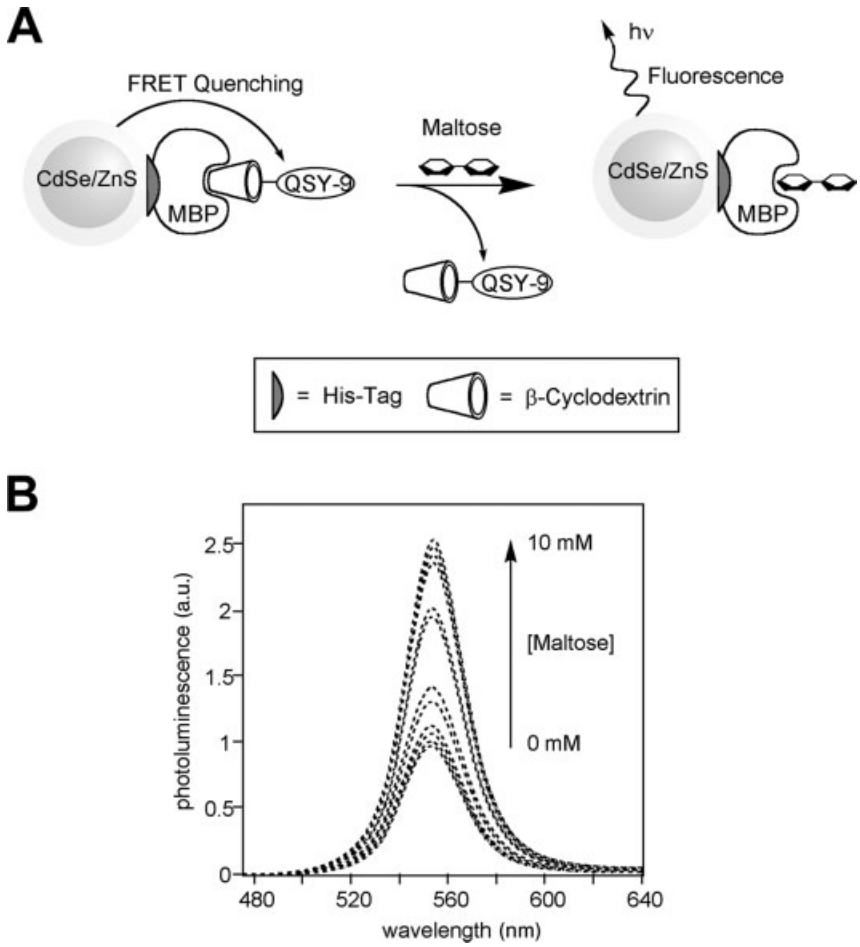


Figure 6.26 (A) Application of CdSe/ZnS QDs for the competitive assay of maltose using the maltose binding protein (MBP) as sensing material and β-cyclodextrin–QSY-9 dye conjugate, β-CD–QSY–9, as FRET quencher. (B) Fluorescence changes of the MBP-functionalized QDs upon analyzing increasing amounts of maltose. (Reprinted by permission from Macmillan Publishers Ltd.; Nature Materials [188], Copyright (2003)).

the presence of the quencher–peptide capping layer. The hydrolytic cleavage of the peptide resulted in the removal of the quencher units and this restored the fluorescence of the QDs. For example, collagenase was used to cleave the Rhodamine Red-X dye-labeled peptide (**38**) linked to CdSe/ZnS QDs [Figure 6.27(A)]. While the tethered dye quenched the fluorescence of the QD, hydrolytic scission of the dye and its corresponding removal restored the fluorescence [Figure 6.27(B)].

In a related study, the activity of tyrosinase (TR) was analyzed by CdSe/ZnS QDs [192]. The QDs were capped with tyrosine methyl ester monolayer. The tyrosinase-induced oxidation of the tyrosine groups to the respective dopaquinone units generated active quencher units that suppressed the fluorescence of the QDs

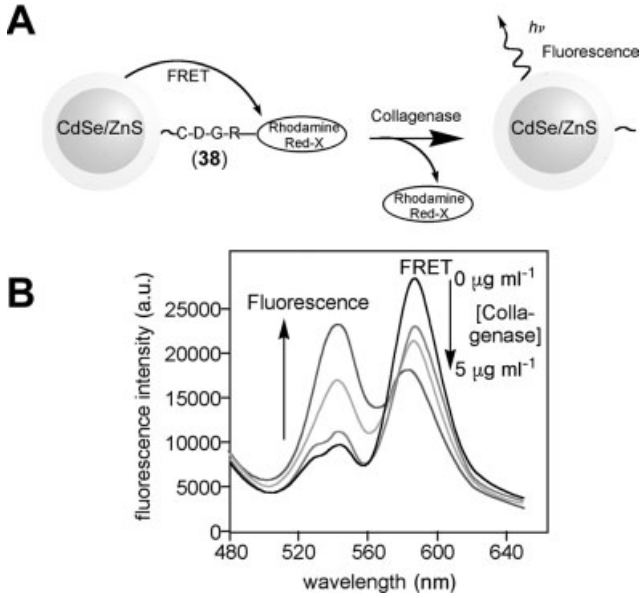


Figure 6.27 (A) Application of CdSe/ZnS QDs for the optical analysis of the protease-mediated hydrolysis of the Rhodamine Red-X-functionalized peptide (**38**). (B) Decrease in the fluorescence of the dye and the corresponding increase in the fluorescence of the QDs upon interaction with different concentrations of collagenase. (Reprinted with permission from [191]. Copyright 2006 American Chemical Society).

[Figure 6.28(A)]. The depletion of the fluorescence of the QDs upon their interaction with different concentrations of tyrosinase (TR) is displayed in Figure 6.28(B). The tyrosinase-stimulated oxidation of phenol residues was further employed to use the QDs to monitor the activity of thrombin [192]. The CdSe/ZnS QDs were functionalized with the peptide **39** that included the specific sequences for cleavage by thrombin and the tyrosine site. The tyrosinase-induced oxidation of tyrosine yields the dopaquinone units that quenched the fluorescence of the QDs [Figure 6.29(A)]. The hydrolytic scission of the peptide by thrombin cleaved off the quinone quencher units and restored the fluorescence of the QDs [Figure 6.29(B)].

The FRET process occurring within a duplex DNA structure consisting of tethered CdSe/ZnS QDs and a dye was applied to probe DNA hybridization and the DNase I cleavage of the DNA [193]. Nucleic acid **40**-functionalized CdSe/ZnS QDs were hybridized with the complementary Texas Red-functionalized nucleic acid **41** [Figure 6.30(A)]. The time-dependent resonance energy transfer from the QDs to the dye units was used to monitor the hybridization process. Treatment of the DNA duplex with DNase I resulted in the cleavage of the DNA and the recovery of the fluorescence properties of the CdSe/ZnS QDs. After cleavage of the double stranded DNA with DNase I, the intensity of the FRET band of the dye decreased and the fluorescence of CdSe/ZnS QDs increased [Figure 6.30(B)]. The luminescence properties of the QDs were only partially recovered due to the nonspecific adsorption of the dye on QDs.

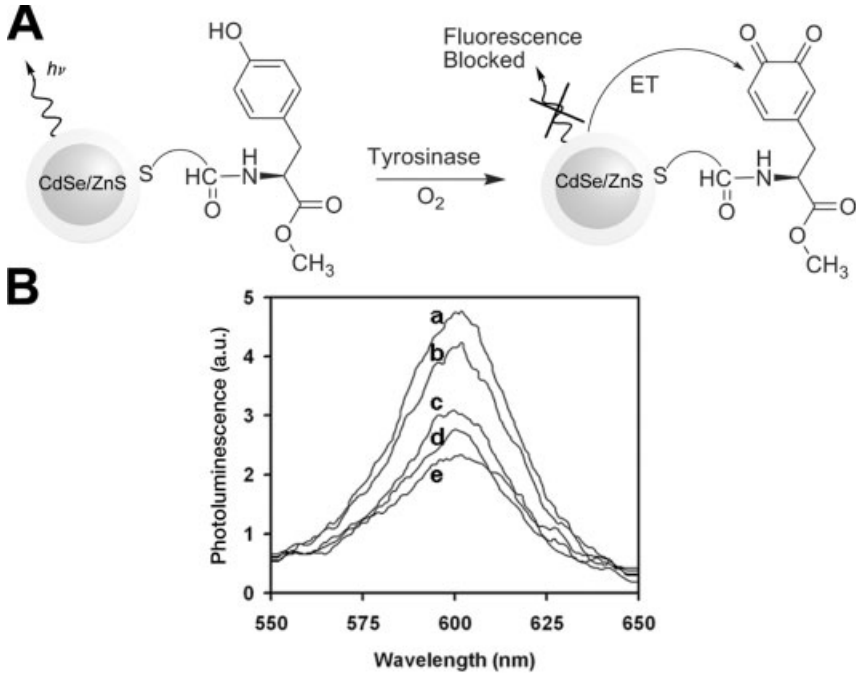


Figure 6.28 (A) Analysis of tyrosinase activity by the biocatalytic oxidation of the methyl ester tyrosine-functionalized CdSe/ZnS QDs to the dopaquinone derivative that results in the electron transfer quenching of the QDs. (B) Time-dependent fluorescence quenching of the QDs upon tyrosinase-induced oxidation of the tyrosine-functionalized QDs: (a) 0, (b) 0.5, (c) 2, (d) 5 and (e) 10 min. (Reprinted with permission from [192]. Copyright 2006 American Chemical Society).

The QDs were also used to probe the formation of aptamer–protein complexes [194]. An anti-thrombin aptamer was coupled to QDs, and the nucleic acid sequence was hybridized with a complementary oligonucleotide–quencher conjugate (Figure 6.31). The fluorescence of the QDs was quenched in the QD–quencher duplex. In the presence of thrombin the duplex was separated and the aptamer underwent a conformational change to the quadruplex structure that binds thrombin. The displacement of quencher units from the blocked aptamer activated the luminescence functions of the QDs, and a about 19-fold increase in their fluorescence was observed.

6.8 Semiconductor Nanoparticles for Photoelectrochemical Applications

The photoexcitation of the semiconductor quantum dots yields the transfer of an electron from the semiconductor valence band to its conduction band, to yield an electron–hole pair. The electron–hole recombination in the QD may lead to either thermal or radiative relaxation of the excited species. The immobilization of QDs

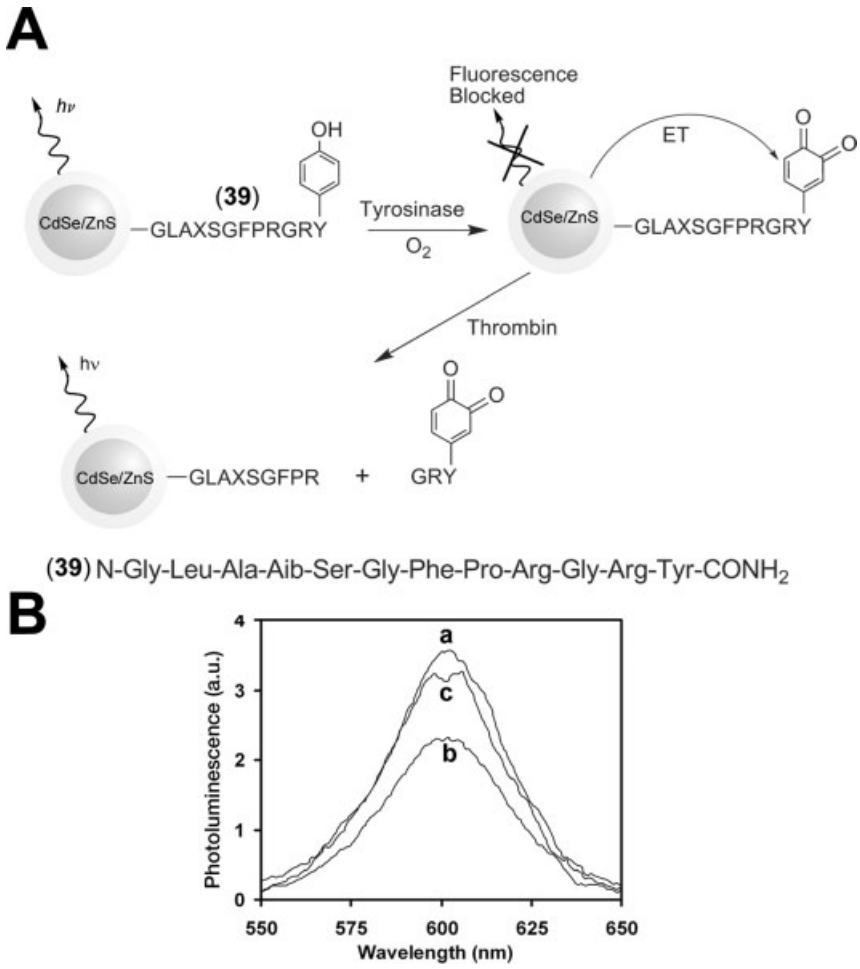


Figure 6.29 (A) Sequential analysis of tyrosinase activity and thrombin activity by the tyrosinase-induced oxidation of the tyrosine-containing peptide (**39**) associated with the CdSe/ZnS QDs that results in the electron transfer quenching of the QDs, followed by the thrombin-induced cleavage of the dopaquinone-modified peptide that restores the fluorescence of the QDs. (B) Fluorescence of: (a) the **39**-modified QDs; (b) after reaction of the QDs with tyrosinase, 10 min, and (c) after treatment of the dopaquinone-functionalized QDs with thrombin, 6 min. (Reprinted with permission from [132]. Copyright 2006 American Chemical Society).

onto electrodes permits the utilization of the photogenerated electron–hole pair for inducing the formation of photocurrents (photoelectrochemical effect). The photocurrent may be formed by the transfer of the conduction band electrons to the bulk electrode and the concomitant transfer of electrons from the electron donor to the valence band holes to yield a steady-state cathodic photocurrent. Alternatively, the photogenerated conduction band electrons may be transferred to a solution-solubilized electron acceptor, with the concomitant transfer of electrons from the

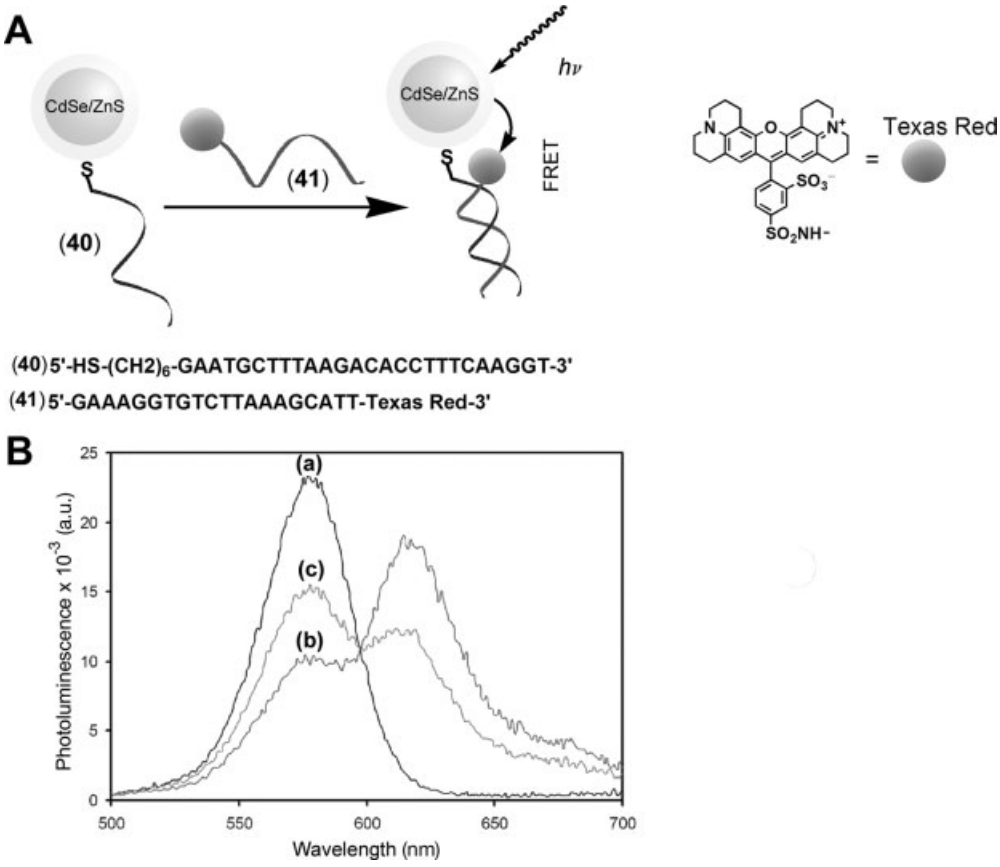


Figure 6.30 (A) Assembly of the CdSe/ZnS and Texas Red-tethered duplex DNA. (B) Fluorescence spectra of: (a) the (**40**)-functionalized CdSe/ZnS QDs; (b) the (**40**)/(**41**) duplex DNA tethered to the QDs and the Texas Red chromophore; (c) after treatment of the duplex DNA tethered to CdSe/ZnS and the dye with DNase I. (Reprinted with permission from [193]. Copyright 2005 American Chemical Society).

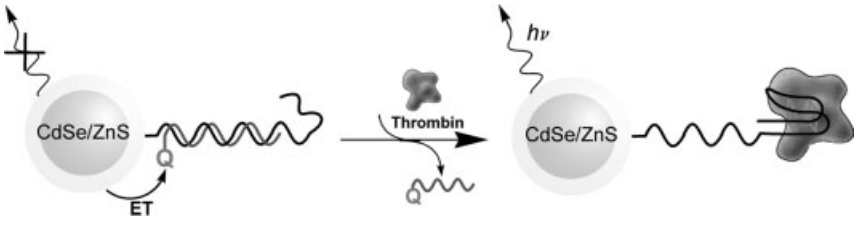


Figure 6.31 Analysis of thrombin by the protein-induced separation of the anti-thrombin aptamer blocked by a quencher-functionalized nucleic acid that restores the fluorescence of the QDs.

electrode to the valence band hole and the formation of an anodic photocurrent. The coupling of biomolecule–QD conjugates to electrode surfaces enables one not only to use the photoelectrochemical effect as a means to transduce biosensing processes, but also to tailor functional nano-architectures on surfaces that perform logic gate operations or act as switching systems. Different QD–DNA hybrid systems [195, 196] or QD–protein conjugates [197, 198] were assembled on electrodes, and the control of the photoelectrochemical properties of the QDs by the biomolecules was demonstrated. CdS nanoparticles were assembled on electrodes by double-stranded nucleic acids acting as bridging units, and the effect of a redox-active intercalator on the resulting photocurrent and its direction was demonstrated [196]. Dithiol-tethered single-stranded ssDNA (**42**) was assembled on an Au electrode and subsequently hybridized with a complementary dithiolated ssDNA (**43**) to yield a double-stranded DNA. The resulting surface was treated with CdS–NPs to yield a semiconductor nanoparticle interface linked to the electrode surface (Figure 6.32). Irradiation of the dsDNA–CdS NP-modified electrode in the presence of triethanolamine (TEOA) as electron donor resulted in an anodic low-intensity photocurrent. The observed generated photocurrent was attributed to an imperfect structure of the CdS NP–DNA assemblies that resulted from direct contact between the NPs and the electrode, rather then from charge transport through the DNA. The DNA duplex structure linking the CdS NPs to the electrode could be employed, however, as a medium to incorporate redox-active intercalators that facilitate electrical contact between the NPs and the electrode, resulting in enhanced photocurrents. Methylene blue (MB) (**44**) was intercalated into the (**42**/**43**)-dsDNA coupled to the CdS NPs [Figure 6.32(A) and (B)]. The cyclic voltammogram of the system implied that at potentials $E > -0.28$ V (vs. SCE) the intercalator exists in its oxidized form (**44**), whereas at potentials $E < -0.28$ V (vs. SCE) the intercalator exists in its reduced leuco form (**45**). Coulometric analysis of the MB redox wave, $E^{\circ} = -0.28$ V (vs. SCE), knowing the surface coverage of the

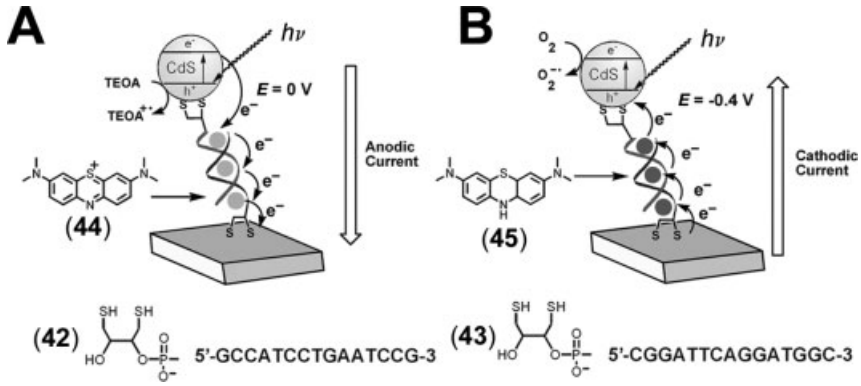


Figure 6.32 Directional electroswitched photocurrents in the CdS NP–ds-DNA–intercalator system. (A) Enhanced generation of anodic photocurrent in the presence of the oxidized methylene blue intercalator (**44**) (applied potential $E = 0$ V). (B) Enhanced generation of cathodic photocurrent in the presence of the reduced methylene blue intercalator (**45**) (applied potential $E = -0.4$ V). (Reproduced with permission from [196]. Copyright 2005 Wiley-VCH).

dsDNA, indicated that about 2–3 intercalator units were associated with the double-stranded DNA. An anodic photocurrent was generated in the system in the presence of TEOA as electron donor and MB intercalated into the dsDNA and while applying a potential of 0 V (vs. SCE) on the electrode. At this potential MB existed in its oxidized state (**44**), which acts as an electron acceptor. The resulting photocurrent was about fourfold higher than that recorded in the absence of the intercalator.

The enhanced photocurrent was attributed to the trapping of conduction band electrons by the intercalator units and their transfer to the electrode that was biased at 0 V, thus retaining the intercalator units in their oxidized form. The oxidation of TEOA by the valence band holes then led to the formation of the steady-state anodic photocurrent. Biasing the electrode at a potential of −0.4 V (vs. SCE), a potential that retained the intercalator units in their reduced state (**45**), led to blocking of the photocurrent in the presence of TEOA and under an inert argon atmosphere. This experiment revealed that the oxidized intercalator moieties with the DNA matrix played a central role in the charge transport of the conduction band electrons and the generation of the photocurrent.

Figure 6.33(A), curve b, shows the photocurrent generated by the (**42**/**43**)–dsDNA linked to the CdS NPs in the presence of the reduced intercalator **45** under conditions where the electrode was biased at −0.4 V (vs. SCE) and the system was exposed to air (oxygen). At a bias potential of −0.4 V (vs. SCE), the intercalator units exist in their reduced *leuco* form (**45**) that exhibits electron-donating properties [Figure 6.32(B)]. Photoexcitation of the CdS NPs yields electron–hole pairs in the conduction band and valence band, respectively. The transport of the conduction band electrons to oxygen with the concomitant transport of electrons from the reduced intercalator units to the

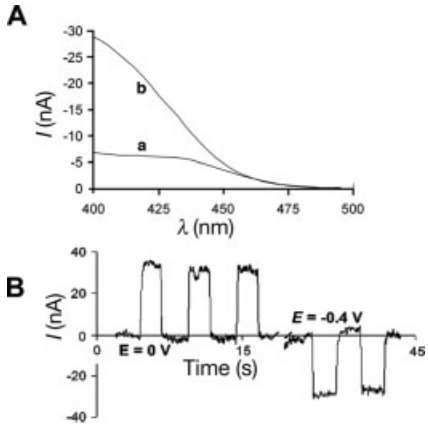


Figure 6.33 (A) Cathodic photocurrents generated in the CdS NP–dsDNA system associated with the electrode: (a) in the absence and (b) in the presence of reduced methylene blue intercalator (**45**). The data were obtained at the applied potential $E = -0.4$ V and in the presence of air. (B) Electrochemically switched anodic and cathodic photocurrents generated by the Cd NP–dsDNA–**44**/**45** systems at 0 and −0.4 V, respectively, in the presence of 20 mM TEOA and air. Photocurrents were generated with irradiation at $\lambda = 420$ nm. (Reproduced with permission from [196]. Copyright 2005 Wiley-VCH).

valence band holes completed the cycle for the generation of the photocurrent. The fact that the electrode potential retained the intercalator units in their reduced state and the infinite availability of the electron acceptor (O_2) yielded the steady-state cathodic photocurrent in the system.

The introduction of TEOA and oxygen to the electrode modified with (**42**/**43**)–DNA duplex and associated CdS NPs allowed the control of the photocurrent direction by switching the bias potential applied on the electrode. Figure 6.33(B) depicts the potential-induced switching of the photocurrent direction upon switching the electrode potential between −0.4 (cathodic photocurrents) and 0 V (anodic photocurrents), respectively. A layer-by-layer deposition of nucleic acid-functionalized CdS QDs on the electrode was followed by the photoelectrochemical transduction of the assembly process [195]. Semiconductor CdS NPs (2.6 ± 0.4 nm) were functionalized with one of the two thiolated nucleic acids **46** and **47** that are complementary to the 5′- and 3′-ends of a target DNA molecule (**48**). An array of CdS NP layers was then constructed on an Au electrode by a layer-by-layer hybridization process with the use of the target DNA **48** as crosslinker of CdS QDs functionalized with nucleic acids (**46** or **47**) complementary to the two ends of the DNA target (Figure 6.34). Illumination of the array in the presence of a sacrificial electron donor resulted in the generation of a photocurrent. The photocurrents increased with the number of generations of CdS NPs associated with the electrode, and the photocurrent action spectra followed the absorbance features of the CdS NPs, which implies that the photocurrents originated from the photoexcitation of the CdS nanoparticles. The ejection of the conduction band electrons into the electrode occurred from the QDs that were in intimate contact with the electrode support. This was supported by the fact that $Ru(NH_3)_6^{3+}$ units ($E^\circ = -0.16$ V vs SCE), which were electrostatically bound to the DNA, enhanced the photocurrent from the DNA–CdS array. The $Ru(NH_3)_6^{3+}$ units acted as charge transfer mediators that facilitated the hopping of conduction band electrons from CdS particles, which lacked contact with the electrode, due to their separation by the DNA tethers.

Enzymes or redox proteins were also linked to semiconductor QDs, and the resulting photocurrents were employed to assay the enzyme activities and to develop different biosensors. Cytochrome *c*-mediated biocatalytic transformations were coupled to CdS NPs, and the direction of the resulting photocurrent was controlled by the oxidation state of the cytochrome c mediator [198]. The CdS NPs were immobilized on an Au electrode through a dithiol linker, and mercaptopyridine units, acting as promoter units that electrically communicate between the cytochrome *c* and the NPs, were linked to the semiconductor NPs (Figure 6.35). In the presence of reduced cytochrome *c*, the photoelectrocatalytic activation of the oxidation of lactate by lactate dehydrogenase (LDH) proceeded, while generating an anodic photocurrent [Figure 6.35(A)]. Photoexcitation of the NPs resulted in the ejection of the conduction band electrons into the electrode and the concomitant oxidation of the reduced cytochrome *c* by the valence band holes. The resulting oxidized cytochrome *c* subsequently mediated the LDH-biocatalyzed oxidation of lactate. Similarly, cytochrome *c* in its oxidized form was used to stimulate the bioelectrocatalytic reduction of NO_3^- to NO_2^- in the presence of nitrate reductase (NR), while generating a cathodic

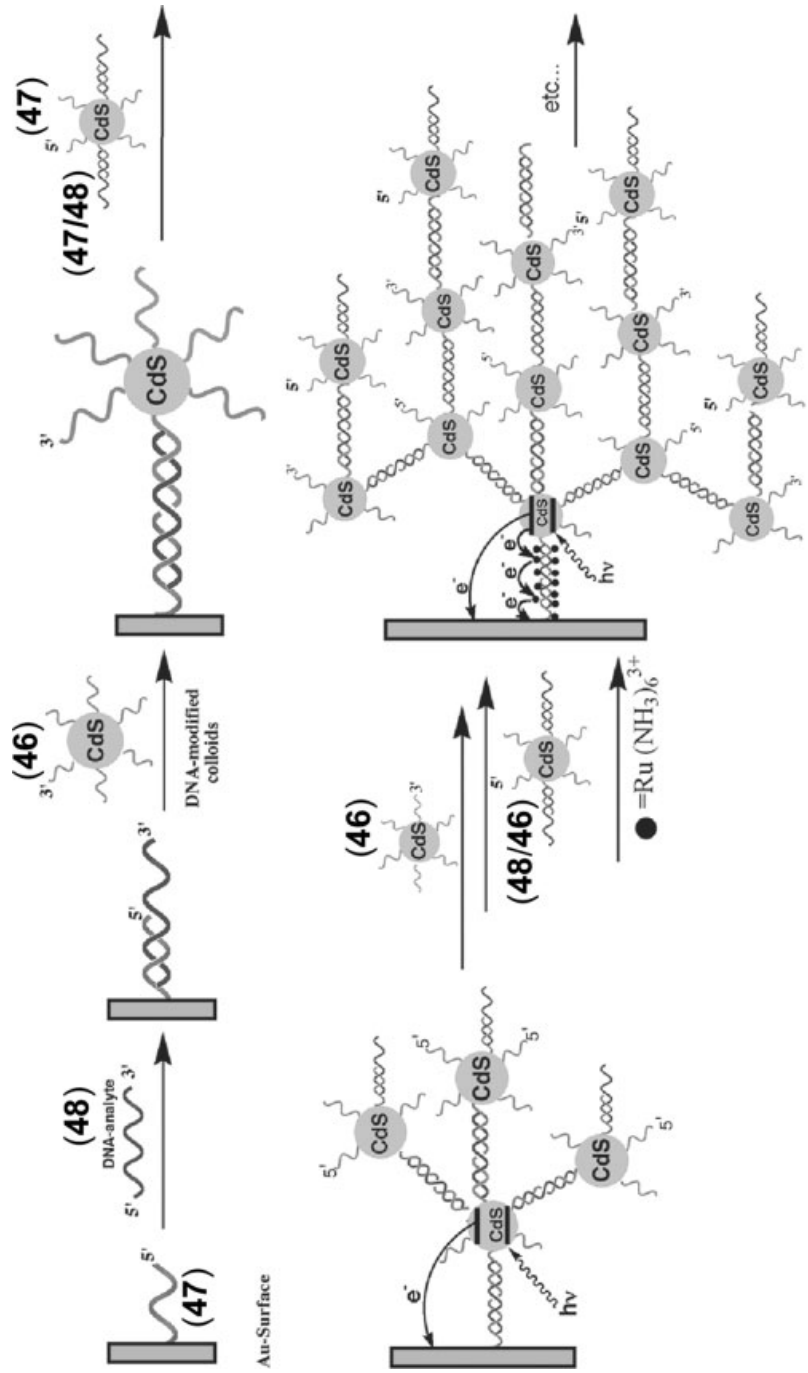


Figure 6.34 Layer-by-layer deposition of CdS NPs using **46**- and **47**-functionalized NPs and **48** as crosslinker. The association of $Ru(NH_3)_6^{3+}$ with the DNA array facilitates charge transport and enhances the resulting photocurrent. (Reproduced with permission from [195]. Copyright 2001 Wiley-VCH).

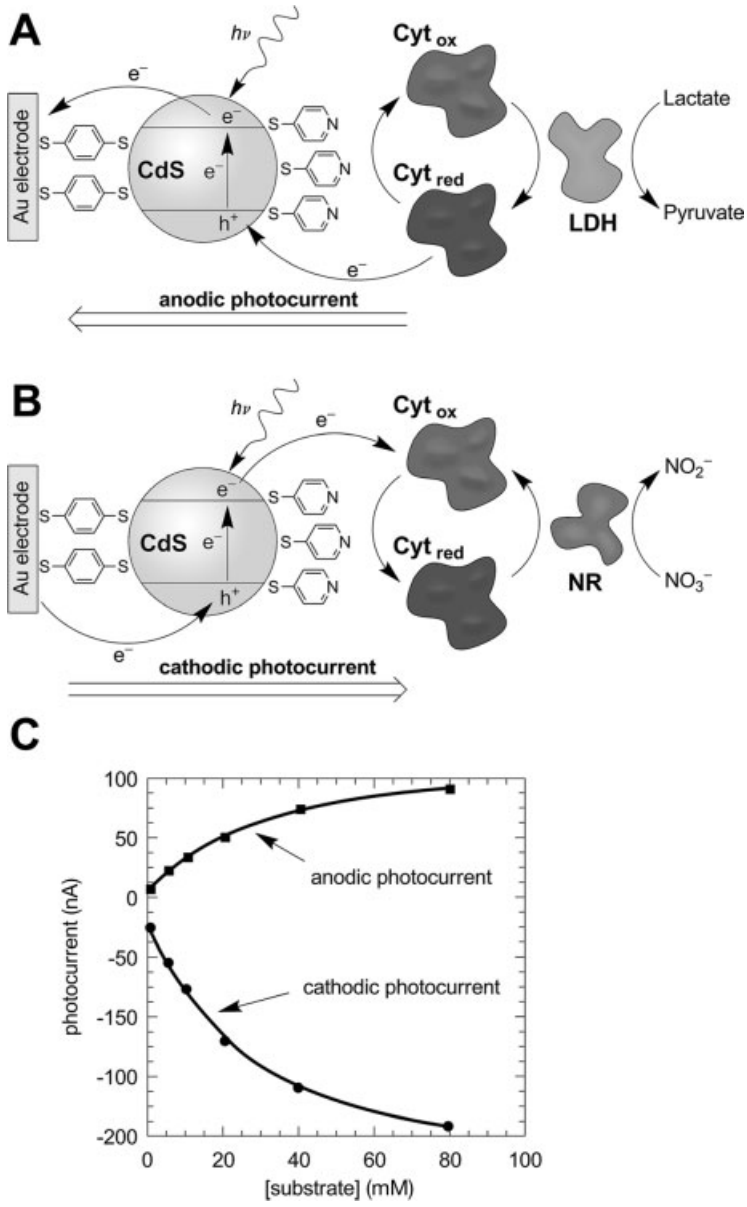


Figure 6.35 Generation of photocurrents by the photochemically induced activation of enzyme cascades by CdS NPs. (A) Photochemical activation of the cytochrome *c*-mediated oxidation of lactate in the presence of LDH. (B) Photochemical activation of the cytochrome *c*-mediated reduction of nitrate (NO_3^-) by nitrate reductase (NR). (C) Photocurrents generated by the biocatalytic cascades in the presence of various concentrations of the substrates (lactate/nitrate). (Reproduced from [198] by permission of The Royel Society of Chemistry).

photocurrent [Figure 6.35(B)]. The transfer of conduction band electrons to the oxidized, heme-containing cofactor generated the reduced cytochrome *c*, while the transfer of electrons from the electrode to the valence band holes of the NPs restored the ground state of the NPs. The cytochrome *c*-mediated biocatalyzed reduction of NO_3^- to nitrite then allowed the formation of the cathodic photocurrent, while biasing the electrode potential at 0 V vs. SCE. The photocurrents generated by the biocatalytic cascades at various concentrations of the different substrates are depicted in Figure 6.35 (C). These results demonstrated that photoelectrochemical functions of semiconductor NPs could be used to develop sensors for biocatalytic transformations. A related study employed CdSe/ZnS QDs capped with mercaptosuccinic acid as a protecting layer for the generation of photocurrents in the presence of cytochrome *c* [199].

In a different study [197], CdS NPs were assembled on an Au electrode, and the NPs were further modified with acetylcholinesterase (Figure 6.36). The biocatalyzed

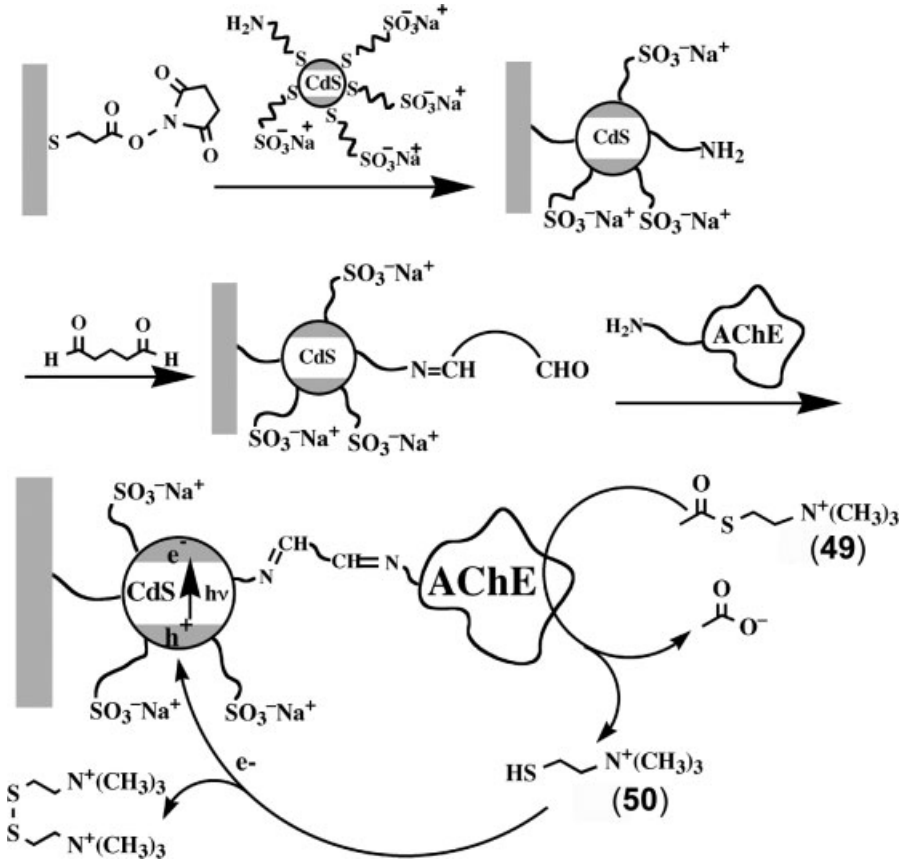


Figure 6.36 Assembly of the CdS NP–AChE hybrid system for the photoelectrochemical detection of the enzyme activity (h^+ = hole). (Reprinted with permission from [197]. Copyright 2003 American Chemical Society).

hydrolysis of acetylthiocholine (**49**) by acetylcholinesterase generated thiocholine (**50**), which acted as an electron donor for the photogenerated holes in the valence band of the CdS NPs. The resulting photocurrent was controlled by the concentration of the substrate and was depleted in the presence of inhibitors of acetylcholinesterase. The system was suggested as a potential sensor for chemical warfare agents that act as inhibitors of acetylcholinesterase.

6.9 Biomolecules as Catalysts for the Synthesis of Nanoparticles

It is well established that living organisms may synthesize NPs and even shaped metallic NPs [200–205]. For example, triangular gold nanoparticles were synthesized by using *Aloe vera* and lemongrass plant (*Cymbopogon flexuosus*) extracts as reducing agents [205, 206]. Although the mechanism of growth of the NPs is not clear, the resulting nanostructures originate from one or more products that are generated by the cell metabolism. This suggests that biomolecules might be active components for synthesizing NPs.

Indeed, a new emerging area in nanobiotechnology involves the use of biomaterials and, specifically, enzymes as active components for the synthesis and growth of particles [207]. As the enlargement of the NPs dominates their spectral properties (extinction coefficient, plasmon excitation wavelength), the biocatalytic reactions that yield the nanoparticles may be sensed by the optical properties of the generated NPs. Furthermore, enzyme–metal NP conjugates may provide biocatalytic hybrid systems that act as amplifying units for biosensing processes.

Different oxidases generate H_2O_2 upon biocatalyzed oxidation of the corresponding substrates by molecular oxygen. The H_2O_2 generated was found to reduce $AuCl_4^-$ in the presence of Au NP seeds that act as catalysts. This observation led to the development of an optical system for the detection of glucose oxidase activity and for the sensing of glucose [208] [Figure 6.37(A)]. A glass support was modified with an aminopropylsiloxane film to which negatively charged Au NPs were linked. Glucose oxidase (GOx) biocatalyzed the oxidation of glucose, and this led to the formation of H_2O_2 that acted as a reducing agent for the catalytic deposition of Au on the Au NPs associated with the glass support. The enlargement of the particles was then followed spectroscopically [Figure 6.37(B)]. Since the amount of the H_2O_2 formed is controlled by the concentration of glucose, the absorbance intensities of the resulting NPs are dominated by the concentration of glucose and the respective calibration curve [Figure 6.37(C)] was extracted. Other enzymes demonstrated, similarly, the biocatalytic generation of reducing products that grow NPs. For example, alkaline phosphatase hydrolyses *p*-aminophenol phosphate to yield *p*-aminophenol and the latter product reduces Ag^+ on Au NP seeds that act as catalytic sites [209].

Different bioactive *o*-hydroquinone derivatives such as the neurotransmitters dopamine (**51**), L-DOPA (**52**), adrenaline (**53**) and noradrenaline (**54**), were found to act as effective reducing agents of metal salts to the respective metal NPs, for example, the reduction of $AuCl_4^-$ to Au NPs without catalytic seeds [210]. For

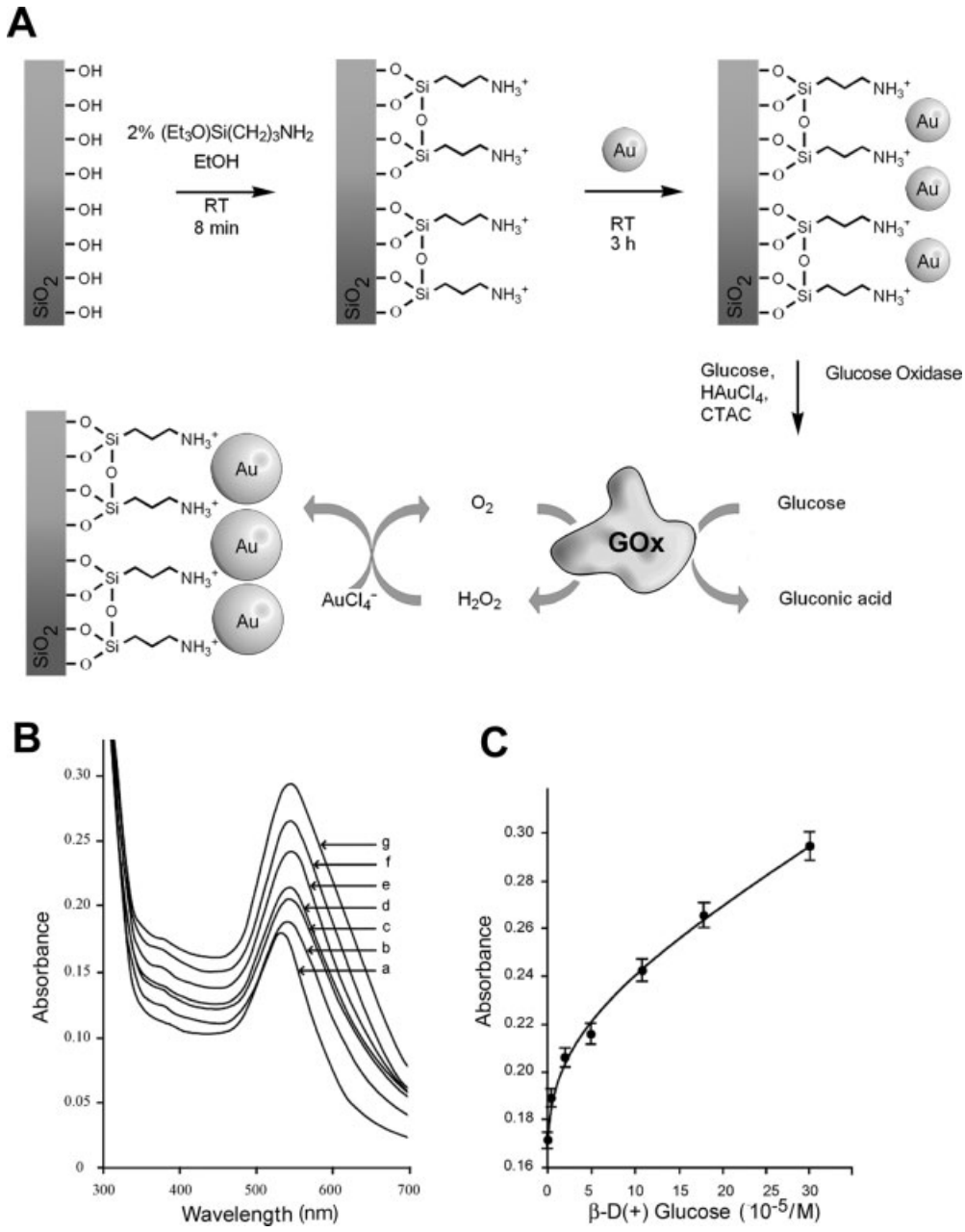


Figure 6.37 (A) Biocatalytic enlargement of Au NPs in the presence of glucose and glucose oxidase (GOx). RT, room temperature; CTAC, cetyltrimethylammonium chloride. (B) Absorbance spectra of Au NP-functionalized glass supports upon reaction with 2×10^{-4} M $HAuCl_4$ and 47 μg mL^{-1} GOx in 0.01 M phosphate buffer that includes CTAC (2×10^{-3} M) and different concentrations of β-D-(+)-glucose: (a) 0, (b) 5×10^{-6}, (c) 2×10^{-5}, (d) 5×10^{-5}, (e) 1.1×10^{-4}, (f) 1.8×10^{-4} and (g) 3.0×10^{-4} M. For all experiments, the reaction time was 10 min and the temperature was 30 °C. (C) Calibration curve corresponding to the absorbance at $\lambda = 542$ nm of the Au NP-functionalized glass supports upon analyzing different concentrations of glucose. (Reprinted with permission from [208]. Copyright 2005 American Chemical Society).

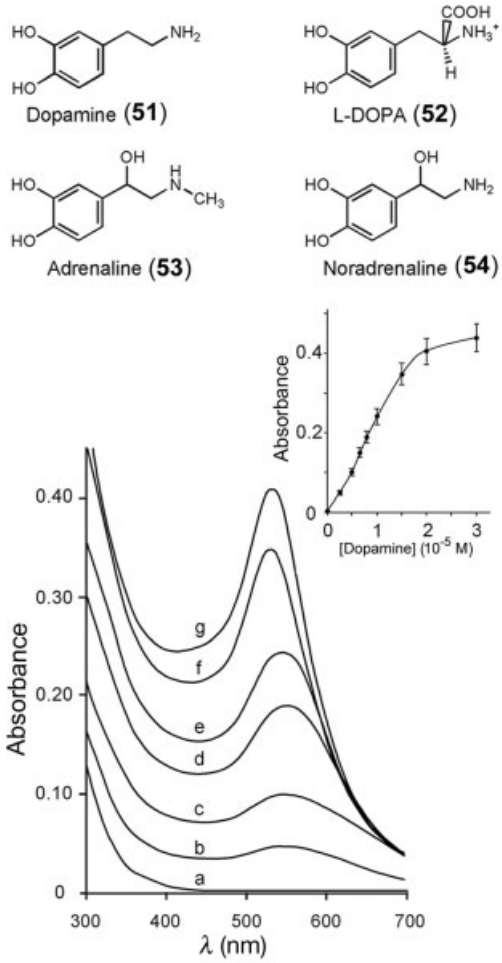


Figure 6.38 Absorbance spectra of Au NPs formed in the presence of different concentrations of dopamine (**51**): (a) 0, (b) 2.5×10^{-6}, (c) 5×10^{-6}, (d) 8×10^{-6}, (e) 1×10^{-5}, (f) 1.5×10^{-5} and (g) 2×10^{-5} M. All systems include $HAuCl_4$ (2×10^{-4} M) and CTAC (2×10^{-3} M) in 0.01 M phosphate buffer. Spectra were recorded after a fixed time interval of 2 min. Inset: calibration curve corresponding to the absorbance of the Au NP solution at $\lambda = 540$ nm formed in the presence of various concentrations of dopamine. (Reprinted with permission from [210]. Copyright 2005 American Chemical Society).

example, Figure 6.38 depicts the absorbance features of the Au NPs generated by different concentrations of dopamine (**51**). As the concentration of dopamine increased, the plasmon absorbance of the Au NPs was intensified. Although the Au NPs were enlarged, as the concentration of dopamine was elevated, the maximum absorbance of the plasmon band was blue-shifted. A detailed TEM analysis revealed that the increase in the concentration of dopamine indeed enhanced the growth of the

Au NP. Nonetheless, the enlargement process generated on the Au NP surface small (1–2 nm) Au clusters that were detached to the solution, and these acted as additional seeds for enlargement. As a result, the absorbance spectra corresponded to the enlarged particles and to the numerous small clusters/NPs, and these led to the blue shift in the plasmon absorbance. The spectral features permitted the extraction of a calibration curve corresponding to the optical detection of dopamine (Figure 6.38, inset). Analogous results were observed for the detection of the other neurotransmitters (**52**–**54**) [210].

The optical detection of L-DOPA (**52**) by means of the Au NPs enabled the following of the activity of tyrosinase. The enzyme tyrosinase is specifically expressed by melanocytes and melanoma cells and is viewed as a specific marker for these cells [211]. Tyrosinase hydroxylates tyrosine to L-DOPA using O_2 as the oxygen source [Figure 6.39(A)]. Consequently, the biocatalytically generated L-DOPA stimulated the synthesis of the Au NPs, and these acted as optical labels for the activity of the enzyme. Figure 6.39(B) shows the absorbance spectra of the Au NPs upon analyzing different amounts of tyrosinase. It was found that tyrosinase could be detected by this method with a sensitivity limit of 10 units.

The catalytic growth of metallic NPs by enzymes and the optical monitoring of the biocatalytic transformations were extended to probe enzyme inhibition [212]. The biocatalyzed hydrolysis of acetylthiocholine (**49**) (an analog of acetylcholine) yields thiocholine (**50**), and the thiol product was found to act as a reducing agent that reduces $AuCl_4^-$ on Au NP seeds. The enlargement of the NPs was used to follow the AChE

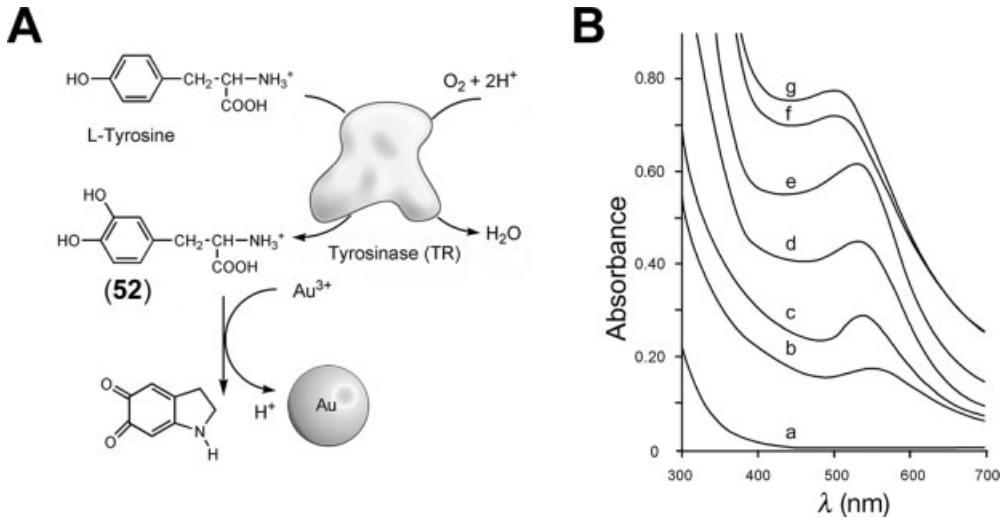


Figure 6.39 (A) Assay of tyrosinase activity through the biocatalyzed oxidation of tyrosine and the L-DOPA (**52**)-mediated formation of Au NPs. (B) Absorbance spectra of the Au NPs formed by various concentrations of tyrosinase: (a) 0, (b) 10, (c) 20, (d) 30, (e) 35, (f) 40 and (g) 60 U mL^{-1}. All systems include tyrosine (2×10^{-4} M), $HAuCl_4$ (2×10^{-4} M) and CTAC (2×10^{-3} M) in 0.01 M phosphate buffer. Spectra were recorded after a fixed time interval of 10 min. (Reprinted with permission from [210]. Copyright 2005 American Chemical Society).

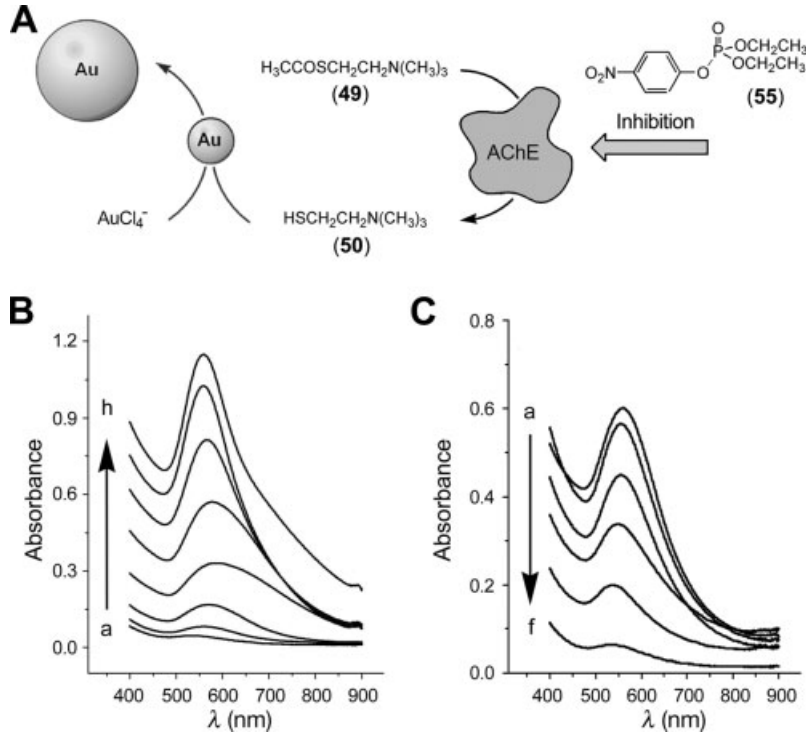


Figure 6.40 (A) AChE-mediated growth of Au NPs and its inhibition by a nerve gas analog. (B) Absorbance spectra of the Au NPs formed in the presence of AChE ($0.13\ U\ mL^{-1}$), $HAuCl_4$ (1.1×10^{-3} M), Au NP seeds (3.6×10^{-8} M) and various concentrations of acetylthiocholine (**49**): (a) 0, (b) 2.4×10^{-5}, (c) 4.8×10^{-5}, (d) 9.5×10^{-5}, (e) 1.4×10^{-4}, (f) 1.9×10^{-4}, (g) 2.4×10^{-4} and (h) 3.8×10^{-4} M. Spectra were recorded after 5 min of particle growth. (C) Absorbance spectra of the Au NPs formed in the presence of AChE and acetylthiocholine in the presence of different concentrations of the inhibitor (**55**): (a) 0, (b) 1.0×10^{-8}, (c) 5.0×10^{-8}, (d) 1.0×10^{-7}, (e) 2.0×10^{-7} and (f) 4×10^{-7} M. All systems include AChE ($0.13\ U\ mL^{-1}$), $HAuCl_4$ (1.1×10^{-3} M), Au NP seeds (3.6×10^{-8} M) and acetylthiocholine (1.4×10^{-3} M). In all experiments, AChE was incubated with the respective concentration of the inhibitor for a time interval of 20 min, and the absorbance spectra of the enlarged Au NPs were recorded after 5 min of biocatalytic growth. (Reprinted with permission from [212]. Copyright 2005 American Chemical Society).

activity [Figure 6.40(A)]. The fact that the enzyme controls the absorbance of the Au NPs and thus their degree of enlargement [Figure 6.40(B)] suggested that upon inhibition of AChE the growth of the particles would be blocked. Indeed, the addition of paraoxon (**55**), a well-established AChE irreversible inhibitor that mimics the functions of organophosphate nerve gases, to the biocatalytic system that synthesizes the nanoparticles resulted in the inhibition of the growth of the Au NPs [Figure 6.40(C)]. Also, an electron-transfer mediator, Os(II) bispyridine-4-picolinic acid, was used for the enlargement of Au NP seeds through the biocatalytic oxidation of glucose by GOx [213]. A similar mediated redox mechanism was used to follow the inhibition of AChE through blocking the enlargement of the Au NPs [213].

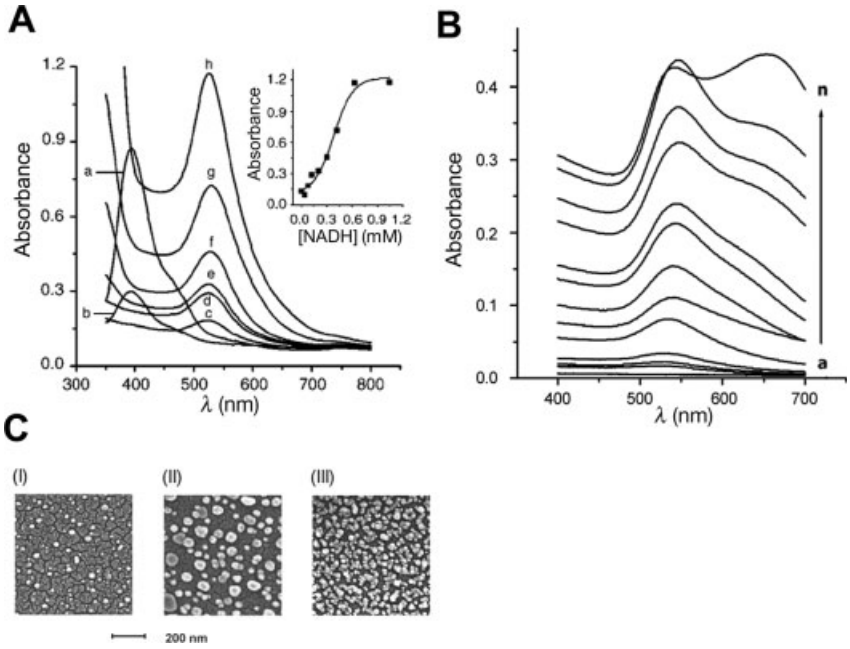


Figure 6.41 (A) Absorption spectra corresponding to the growth of Au NPs (4.0×10^{-10} M) in solution in the presence of increasing concentrations of NADH: (a) 0, (b) 4.2×10^{-5}, (c) 8.4×10^{-5}, (d) 12.6×10^{-5}, (e) 21×10^{-5}, (f) 30×10^{-5}, (g) 42×10^{-5} and (h) 63×10^{-5} M. Inset: calibration curve for the analysis of NADH by the resulting Au NPs at $\lambda = 524$ nm. (B) Absorbance changes of the Au NP–aminopropylsiloxane-functionalized glass slides upon enlargement with different concentrations of NADH: (a) 0, (b) 14×10^{-5}, (c) 27×10^{-5}, (d) 34×10^{-5}, (e) 41×10^{-5}, (f) 44×10^{-5}, (g) 48×10^{-5}, (h) 51×10^{-5}, (i) 54×10^{-5}, (j) 58×10^{-5}, (k) 61×10^{-5}, (l) 65×10^{-5}, (m) 68×10^{-5} and (n) 1.36×10^{-3} M. All systems included $HAuCl_4$ (1.8×10^{-4} M) and CTAB (7.4×10^{-2} M). (C) SEM images of NADH-enlarged Au particles generated on an Au NP-functionalized glass support using $HAuCl_4$ (1.8×10^{-4} M), CTAB (7.4×10^{-2} M) and NADH: (I) 2.7×10^{-4}; (II) 5.4×10^{-4} and (III) 1.36×10^{-3} M. (Reproduced with permission from [75]. Copyright 2004 Wiley-VCH).

1,4-Dihydronicotinamide adenine dinucleotide (phosphate) (NADH or NADPH) cofactors were found to enlarge Au NP seeds by the reduction of $AuCl_4^-$. The enzyme/cofactor-mediated enlargement or synthesis of metal NPs thus provides a means to follow the activities of enzymes or their substrates. Numerous biocatalyzed transformations involve NAD^+-dependent enzymes and the coupling of these biocatalysts to the growth of Au NPs allowed the development of different biosensors [75]. Figure 6.41(A) shows the absorbance spectra of a system consisting of Au NP seeds, $AuCl^{4-}$ and cetyltrimethylammonium bromide (CTAB) upon addition of various concentrations of NADH. The plasmon absorbance of the generated Au NPs is intensified as the concentration of NADH increases. Accordingly, the growth of the Au NPs allowed the quantitative optical analysis of NADH [see the calibration curve in Figure 6.41(A), inset]. The enlargement of the Au NP seeds was also examined on

surfaces. Citrate-stabilized Au NP seeds were immobilized on an aminopropylsiloxane film associated with a glass support, and the functionalized surface was treated with different concentrations of NADH in the presence of $AuCl_4^-$. The growth of the Au NPs was followed spectroscopically [Figure 6.41(B)]. The plasmon absorbance of the Au NPs increased as the concentration of NADH was elevated, consistent with the growth of the particles. The SEM images of the particles confirmed the growth of the NPs on the glass support, [Figure 6.41(C)]. The particles generated by 2.7×10^{-4}, 5.4×10^{-4} and 1.36×10^{-3} M NADH are shown in images (I), (II) and (III), which reveal the generation of Au NPs with an average diameter of 13 ± 1, 40 ± 8 and 20 ± 5 nm, respectively. The NPs generated by the high concentration of NADH (III) reveals a 2D array of enlarged particles that touch each other, consistent with the spectral features of the surface.

The quantitative growth of Au NPs by NADH allowed the assay of NAD^+-dependent enzymes and their substrates. This was demonstrated by the application of the enzyme lactate dehydrogenase (LDH) and lactate as substrate [75] [Figure 6.42(A)]. The absorbance spectra of the NPs enlarged in the presence of different concentrations of lactate, are shown in Figure 6.42(B). From the derived calibration curve, lactate

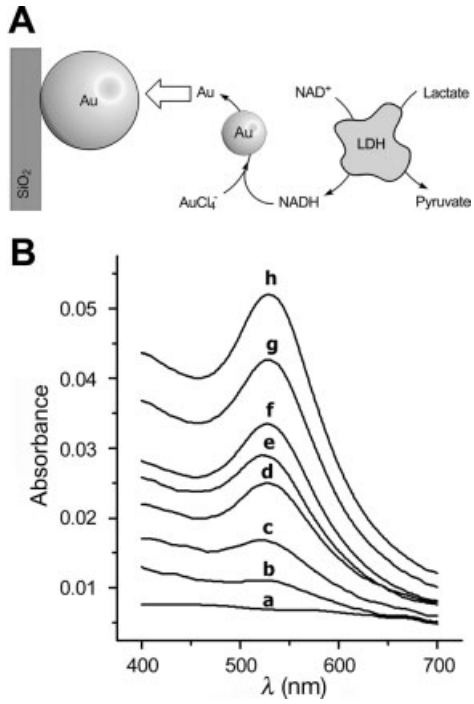


Figure 6.42 (A) Biocatalytic enlargement of Au NPs by the NAD^+–LDH–lactate system. (B) Spectral changes of the Au NP-functionalized glass supports upon interaction with the growth solution consisting of $HAuCl_4$ (1.8×10^{-4} M), CTAB (7.4×10^{-2} M) and lactate–LDH-generated NADH formed within 30 min in the presence of various concentrations of lactate: (a) 0, (b) 2.9×10^{-3}, (c) 3.6×10^{-3}, (d) 5.1×10^{-3}, (e) 5.8×10^{-3}, (f) 6.6×10^{-3}, (g) 7.3×10^{-3}, (h) 9.8×10^{-3} M. (Reproduced with permission from [75]. Copyright 2004 Wiley-VCH).

could be analyzed with a sensitivity limit corresponding to 3×10^{-3} M. Also, the reduced cofactor NADH was also employed to reduce Cu^{2+} ions to Cu^0 metal deposited on core Au NPs [214].

The NADH-induced growth of Au NPs was further developed by demonstrating that the reduced cofactor was able to synthesize, under controlled conditions, shaped NPs (in the form of dipods, tripods and tetrapods). These shaped particles exhibit unique optical properties (color) as a result of their high aspect ratio and the existence of a longitudinal plasmonic exciton [215]. It was demonstrated that in basic aqueous solution (pH = 11) that included ascorbate and NADH as reducing agents and CTAB as surfactant, the rapid formation of the shaped Au NPs was observed. The Au NP shapes consisted of dipods (12%), tripods (45%), tetrapods (13%) and spherical particles (30%) (Figure 6.43). The shaped particles were made of "arms" about 20–25 nm long and with a width of 2–5 nm (Figure 6.43). The development of the shaped particles was controlled by the concentration of NADH. At low NADH concentrations, "embryonic-type" shapes were observed, and at high cofactor concentrations, well-shaped particles were detected [Figure 6.43(A)–(C)]. A series of experiments indicated that the particles were formed in two steps. In the primary step, ascorbate reduced, at pH 11, $AuCl_4^-$ to Au NP seeds. These particles acted as catalysts for the rapid NADH-mediated reduction of $AuCl_4^-$ to Au^0 that was deposited on the seeds. High-resolution transmission electron microscopy (HRTEM) allowed

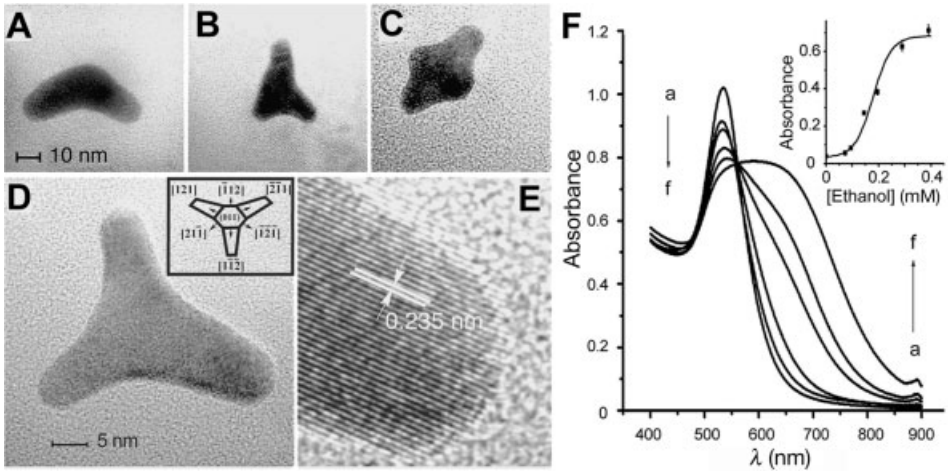


Figure 6.43 (A–C) Typical TEM images of dipod-, tripod- and tetrapod-shaped Au NPs formed in the presence of 4.0×10^{-6} M NADH, respectively. (D) HRTEM image of a representative tripod-shaped Au NP formed in the presence of, 4.0×10^{-6} M NADH. (E) HRTEM analysis of the tripod Au NP. The inset in (D) shows the crystal planes and the tripod growth directions extracted from the HRTEM analysis. (F) Absorbance spectra corresponding to the Au NPs formed by the biocatalytic generation of NADH in the presence of alcohol dehydrogenase (AlcDH) and variable concentrations of ethanol: (a) 0, (b) 7.3×10^{-5}, (c) 9.4×10^{-5}, (d) 1.5×10^{-4}, (e) 2.0×10^{-4} and (f) 2.9×10^{-4} M. Inset: calibration curve corresponding to the absorbance of the shaped Au NPs at $\lambda = 680$ nm formed in the presence of different concentrations of ethanol. (Reprinted with permission from [215]. Copyright 2005 American Chemical Society).

the detailed analysis of the growth directions of the shaped NPs, and this enabled the growth mechanism to be proposed. For example, HRTEM images of tripod particles are given, Figure 6.43(D) and (E). The lattice planes exhibit an inter-planar distance of 0.235 nm that corresponds to the {1 1 1} type planes of crystalline gold. The pods, separated by 120°, revealed a crystallite orientation of [0 1 1] with growth directions of the pods of type <211>; namely the pods are extended the direction $[1\bar{1}\bar{2}]$, $[\bar{2}\bar{1}1]$ and [1 2 1] [Figure 6.43(D), inset]. The shaped Au NPs revealed a red-shifted plasmon absorbance band at $\lambda = 680$ nm, consistent with a longitudinal plasmon exciton in the "rod-like" structures. The blue color of the shaped Au NPs distinctly differs from the red spherical Au NPs.

The fact that the degree of shaping and the absorbance spectra of the resulting Au NPs were controlled by the concentration of NADH allowed the development of an ethanol biosensor based on the shape-controlled synthesis of NPs. The biocatalyzed oxidation of ethanol by AlcDH yields NADH, and the biocatalytically generated NADH acted as an active reducing agent for the formation of blue shaped Au NPs. As the amount of NADH is controlled by the concentration of ethanol, the extent of the structurally developed shaped nanoparticles exhibiting the red-shifted longitudinal plasmon was then controlled by the substrate concentration. Figure 6.43(F) shows the absorbance spectra resulting from the gradual development of the shaped Au NPs formed in the presence of different concentrations of ethanol. The derived calibration curve corresponding to the optical analysis of ethanol by the shaped Au NPs is depicted in Figure 6.43(F), inset.

6.10
Biomolecule Growth of Metal Nanowires

The synthesis of nanowires is one of the challenging topics in nanobiotechnology [216, 217]. The construction of objects at the molecular or supramolecular level could be used to generate templates or "seeds" for nanometer-sized wires. Nanowires are considered as building blocks for the self-assembly of logic and memory circuits in future nanoelectronic devices [218]. Thus, the development of methods to assemble metal or semiconductor nanowires is a basic requirement for the construction of nanocircuits and nanodevices. Furthermore, the nanowires should include functional sites that allow their incorporation into structures of higher complexity and hierarchical functionality. The use of biomaterials as templates for the generation of nanowires and nanocircuitry is particularly attractive. Biomolecules exhibit dimensions that are comparable to those of the envisaged nano-objects. In addition to the fascinating structures of biomaterials that may lead to new inorganic or organic materials, templates of biological origin may act as "factories" for the production of "molds" for nanotechnology. The replication of DNA, the synthesis of proteins and the self-assembly of protein monomers into sheets, tubules and filaments all represent biological processes for the high-throughput synthesis of biomolecular templates for nanotechnology. Specifically, the coupling of NPs to biomolecules might yield new materials that combine the self-assembly properties of the biomolecules with the catalytic functions of the NPs. That is, the catalytic

enlargement of the NPs associated with the biomolecules might generate the metallic nanowires.

Among the different biomaterials, DNA is of specific interest as a template for the construction of nanowires [219–221]. The ease of synthesis of DNA of controlled lengths and predesigned shapes, together with the information stored in the base sequence, introduce rich structural properties and addressable structural domains for the binding of the reactants that form nanowires. Also, numerous biocatalysts such as endonucleases, ligase, telomerase and polymerase can "cut", "paste", elongate or replicate DNA and may be considered as nanotools for shaping the desired DNA and, eventually, for the generation of nanocircuits. In addition, the intercalation of molecular components into DNA and the binding of cationic species such as metal ions to the phosphate units of nucleic acids allow the assembly of chemically active functional complexes, which may be used as precursors for the formation the nanowires.

One of the early examples that demonstrated the synthesis of Ag nanowires [222] is depicted in Figure 6.44. Two microelectrodes, which were positioned opposite one

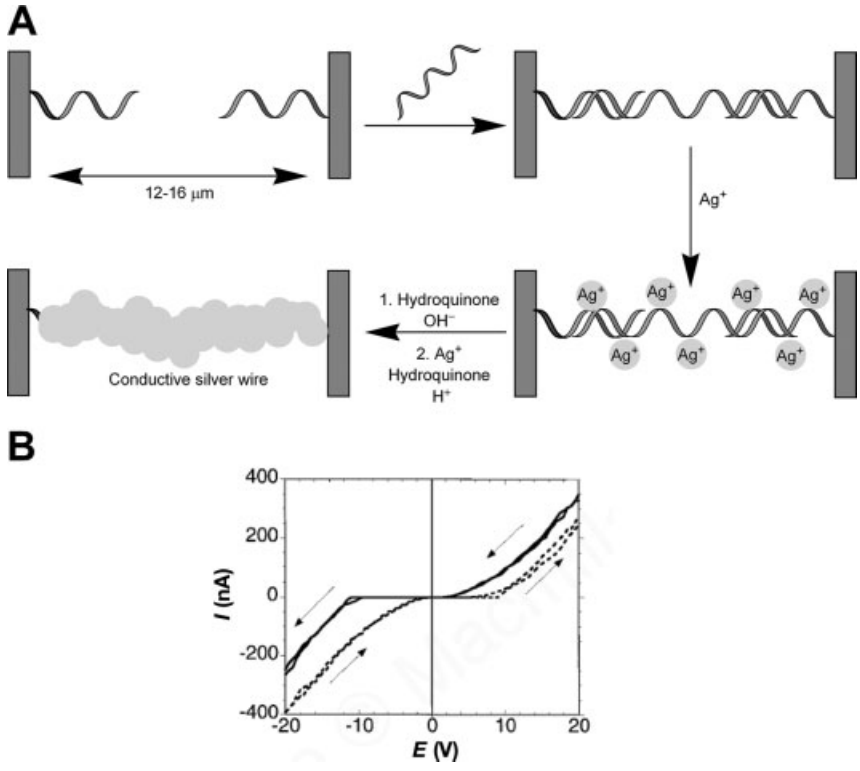


Figure 6.44 (A) Construction of a nanowire that bridges two microelectrodes by the deposition of Ag^+ ions on a bridging DNA strand followed by the chemical reduction of the Ag^+ ions to the metallic agglomerate. (B) Current versus voltage (*I*–*V*) curves obtained with the structures produced. (Reprinted with permission from Macmillan Publishers Ltd: Nature [222]. Copyright 1998).

another with a 12–16-μm separation gap, were functionalized with 12-base oligonucleotides that were then bridged with a 16-μm long λ-DNA [Figure 6.44(A)]. The resulting phosphate units of the DNA bridge were loaded with Ag^+ ions by ion-exchange and the bound Ag^+ ions were reduced to Ag metal with hydroquinone. The resulting small Ag aggregates along the DNA backbone were further reduced by reducing Ag^+ under acidic conditions and catalytic deposition of Ag on the Ag aggregates formed metallic nanowires. The resulting Ag nanowires exhibited dimensions corresponding to micrometer long and about 100 nm wide nanowires. The conduction properties of the nanowires revealed non-ohmic behavior and threshold potentials were needed to activate electron transport through the wires [Figure 6.44 (B)]. The potential gap in which no current passes through the nanowires was attributed to the existence of structural defects in the nanowires. The "hopping" of electrons across these barriers requires an overpotential that is reflected by the break voltage. This study was extended with the synthesis of many other metallic nanowires on DNA templates, and Cu [223], Pt [224] and Pd [225] nanowires were generated on DNA backbones.

The use of DNA and metallic NPs as templates and NPs as building blocks of nanowires was further demonstrated by using telomers synthesized by HeLa cancer cell extracts [226]. The constant repeat units that exist in the telomers provided addressable domains for the self-assembly of the NPs and the subsequent synthesis of the nanowires. By one approach, the primer **56** was telomerized in the presence of the dNTPs nucleotide mixture, which contained (aminoallyl)-dNTPs (**57**) [Figure 6.45(A)]. The resulting amine-containing telomers were treated with Au NPs (1.4 nm) (**58**), which were functionalized with the single *N*-hydroxysuccinimidyl ester groups, to yield Au NP-modified telomers. The Au NP-decorated DNA wires were then enlarged by electroless gold deposition to generate metallic nanowires [Figure 6.45(B) and (C)]. A second approach involved the telomerase-induced generation of telomers that included constant repeat units. The hybridization of the Au NPs functionalized with nucleic acids complementary to the telomer repeat domains, followed by electroless enlargement of the NPs, yielded Au nanowires.

The binding of proteins, such as RecA, to DNA has been used as a means for the patterning of nanoscale DNA-based metal wires with nonconductive or semiconductive gaps [227]. The metallization of the free, non-protein-coated DNA segments permitted the sequence-specific biomolecular lithography of DNA-based nanowires [228]. A single-stranded nucleic acid sequence was complexed with RecA and was carried to a double-stranded duplex DNA, a process that led to the nanolithographic patterned insulation of the DNA template by the protein [Figure 6.46(A)]. A carbon nanotube was then specifically attached to the protein patch using a series of antigen–antibody recognition processes: The anti-RecA antibody (Ab) was bound to the protein linked to the DNA duplex and the biotinylated anti-antibody was then linked to the RecA Ab. The latter Ab was used to bind specifically the streptavidin-coated carbon nanotubes on the protein patch. Ag^+ ions were bound to the free DNA domains and these were reduced to Ag clusters that were associated with the DNA assembly. The dsDNA was modified with aldehyde groups prior to this process to allow the chemical reduction of the electrostatically bound Ag^+ ions to Ag clusters. The

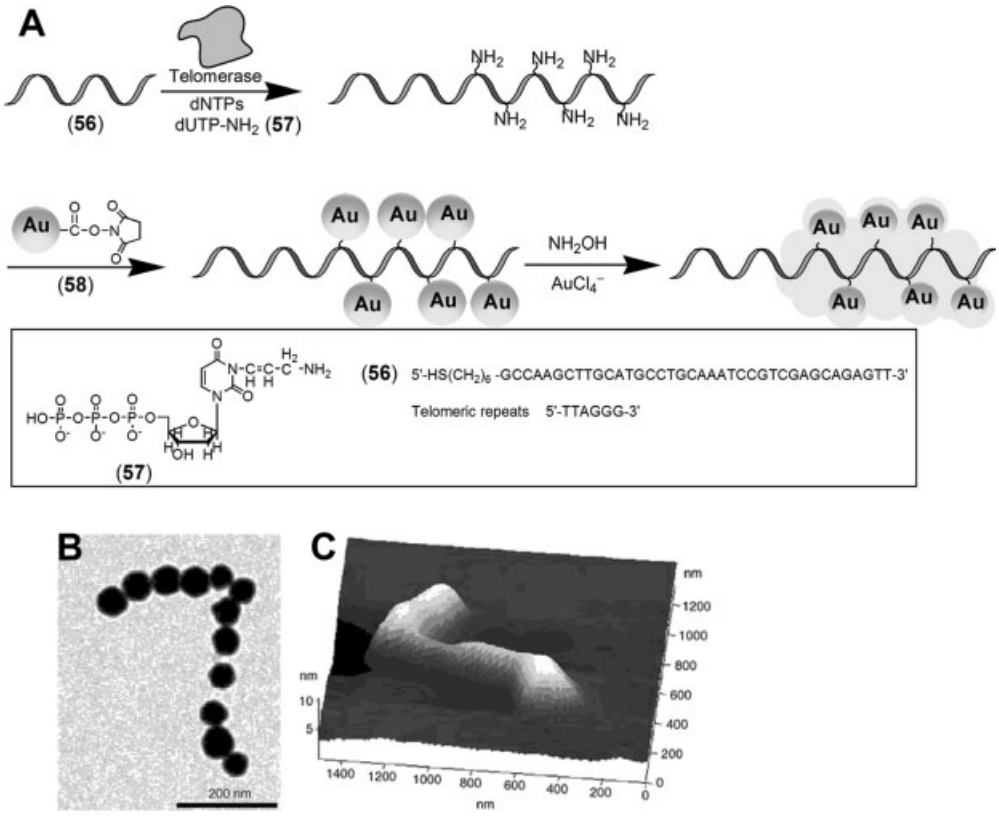


Figure 6.45 Assembly of Au nanowires on a telomer template. (A) Covalent attachment of Au NPs to amine groups, which were introduced into the telomer structure during the telomerization step, followed by catalytic enlargement of the NPs. (B) TEM and (C) AFM images of an Au nanowire that was generated according to the procedure outlined in (A). (Reprinted with permission from [226]. Copyright 2004 American Chemical Society).

subsequent catalytic electroless enlargement of the Ag^0 nanoclusters with gold generated the device where the two Au contacts were bridged by the carbon nanotubes. A representative SEM image of the device is depicted in Figure 6.46(B). The resulting device deposited on a Si substrate acted as a field-effect nanotransistor, where the gold contacts bridged by the CNT acted as the nanoscale source and drain electrodes, and the current flow through the CNT was gated by the applied potential on the Si support [Figure 6.46(C)]. Other DNA–protein arrays based on streptavidin and biotinylated DNA were used to organize tailored NP nanostructures that were subsequently metallized to composite architectures of nanowires [229].

Proteins may be applied as templates for the deposition of metal nanoparticles and for the formation of metal nanowires. For example, aromatic short-chain peptides, such as the Alzheimer's diphenylalanine β-amyloid (**59**), form well-ordered nano-

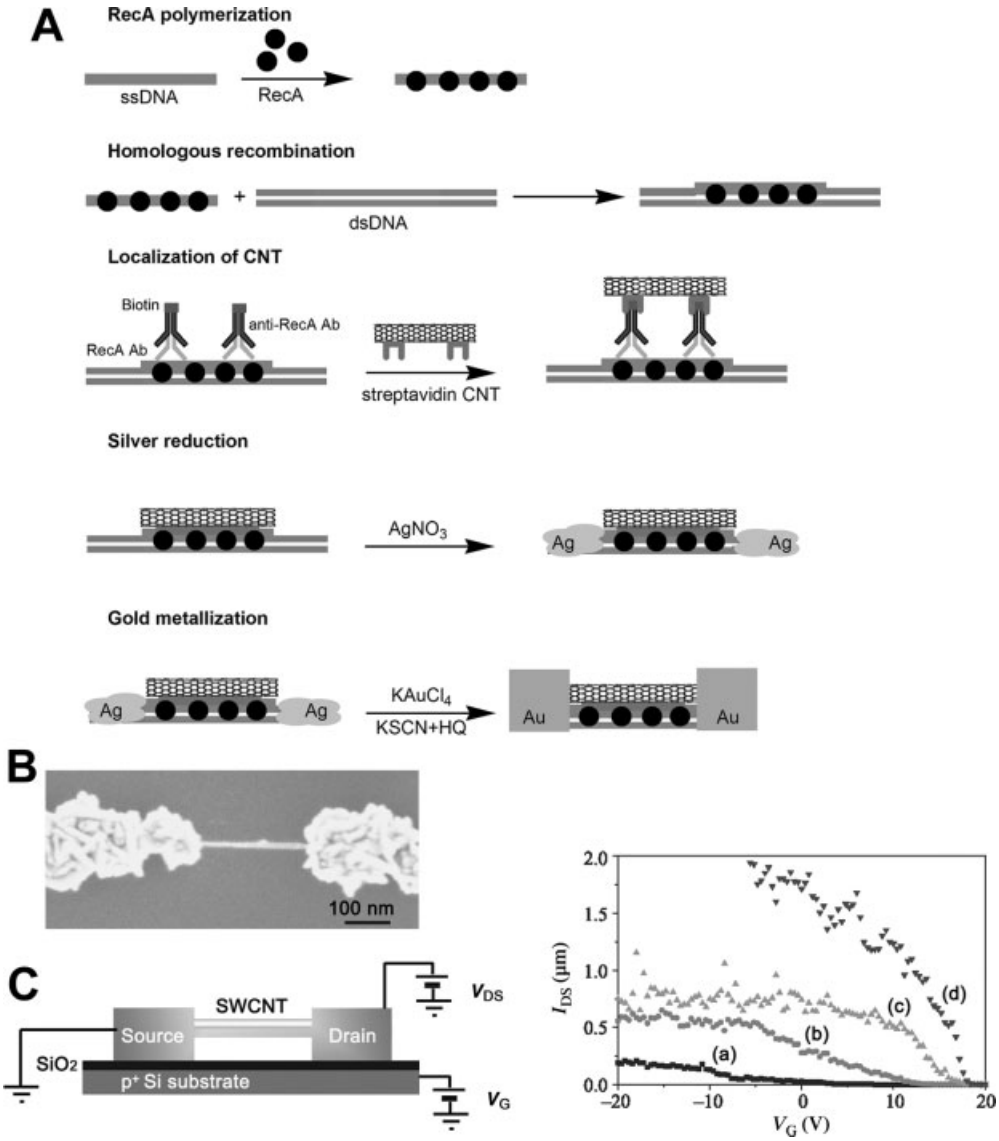


Figure 6.46 (A) Construction of a DNA-templated CNT FET. (B) SEM image of a rope of CNTs and metallic wires contacting it. (C) Electrical circuit and electrical characterization of the DNA-templated CNT FET. The drain–source current is given versus gate voltage for different values of drain–source bias: V_{DS} = (a) 0.5, (b) 1, (c) 1.5 and (d) 2 V. (Parts B and C reproduced from [227] with permission from AAAS).

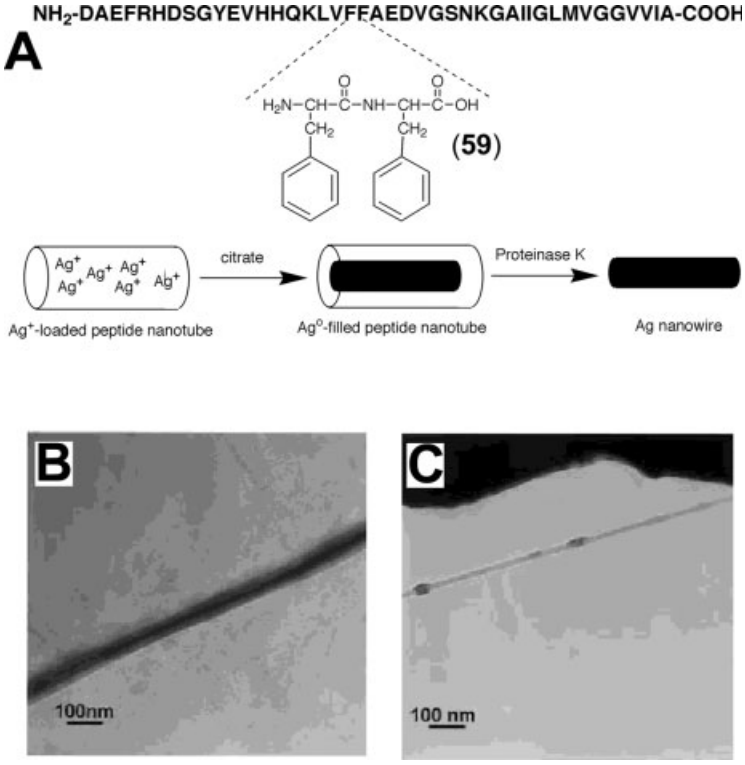


Figure 6.47 (A) Formation of a silver nanowire inside a channel of a short-chain diphenylalanine peptide tube. (B) TEM image of the peptide template, which is filled with a metallic silver nanowire. (C) TEM image of the silver nanowire after the degradation of the peptide template in the presence of proteinase K. (Parts B and C reproduced from [230] with permission from AAAS).

tubes that were used as templates for growing Ag nanowires [230] [Figure 6.47(A)]. The peptide nanotubes were loaded with Ag^+ ions, which were reduced with citrate to yield metallic silver nanowires inside the peptide nanotubes [Figure 6.47(B)]. The peptide coating was then removed by enzymatic degradation in the presence of proteinase K to yield micrometer-long Ag nanowires with a diameter of 20 nm [Figure 6.47(C)]. Upon application of D-phenylalanine-based peptide fibrils, which are resistant to proteinase K, the peptide coating of the Ag nanowires was preserved. Peptide nanotubes were also used to generate coaxial metal–insulator–metal nanotubes [231] [Figure 6.48(A)]. The D-phenylalanine peptide **59** assembled nanotube was interacted with Ag^+ ions and the resulting intra-tube-associated ions were reduced to form the Ag^0 nanowire. The resulting composite was further modified by tethering of a thiol-functionalized peptide to the nanowire peptide coating. The association of Au NPs with the thiol groups followed by deposition of gold on the Au NP seeds generated the resulting coaxial metal nanowires [Figure 6.48(B)].

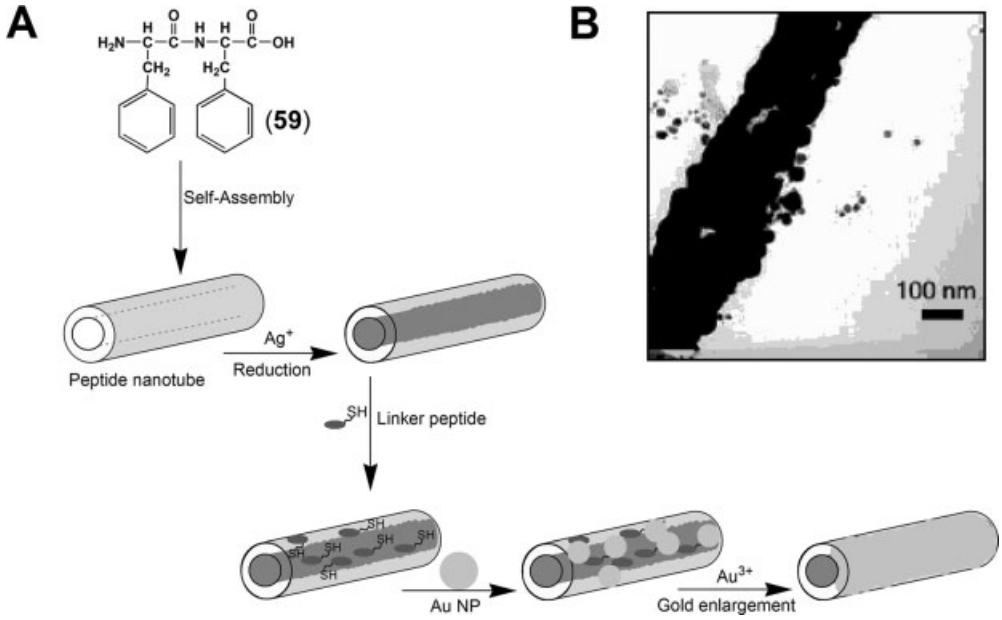


Figure 6.48 (A) Stepwise synthesis of coaxial Ag/Au nanowires in a peptide template. (B) TEM image of the coaxial wire formed in the peptide template. (Reproduced with permission from [231]. Copyright 2006 American Chemical Society).

The specific assembly of protein subunits into polymeric structures could provide a means for the patterning of the generated metal nanowires. The f-actin filament provides specific binding for the biomolecular motor protein myosin, which forms a complex with the filament, where the motility of myosin along the filament is driven by ATP [232, 233]. The f-actin filament is formed by the reversible association of g-actin subunits in presence of ATP, Mg^{2+} and Na^{+} ions. Accordingly, the f-actin filament was used as a template for the formation of metallic nanowires [234]. The f-actin filament was covalently modified with Au NPs (1.4 nm) that were functionalized with single *N*-hydroxysuccinimidyl ester groups, and the Au NP-functionalized g-actin subunits were then separated and used as versatile building blocks for the formation of the nanostructure. The Mg^{2+}/Na^{+}/ATP-induced polymerization of the functionalized monomers yielded the Au NP-functionalized filaments [Figure 6.49(A)], and electroless catalytic deposition of gold on the Au NP-functionalized f-actin filament yielded 1–3-μm long gold wires of height 80–150 nm. The gold wires revealed metallic conductivity with a resistance similar to that of bulk gold. By the sequential polymerization of naked actin filament units on the preformed Au NP–actin wire and the subsequent electroless catalytic deposition of gold on the Au NPs, patterned actin–Au wire–actin filaments were generated [Figure 6.49(C)]. A related approach was applied to yield the inverse Au wire–actin–Au wire patterned filaments [Figure 6.49(B)]. The nanostructure consisting of actin–Au nanowire–actin was used

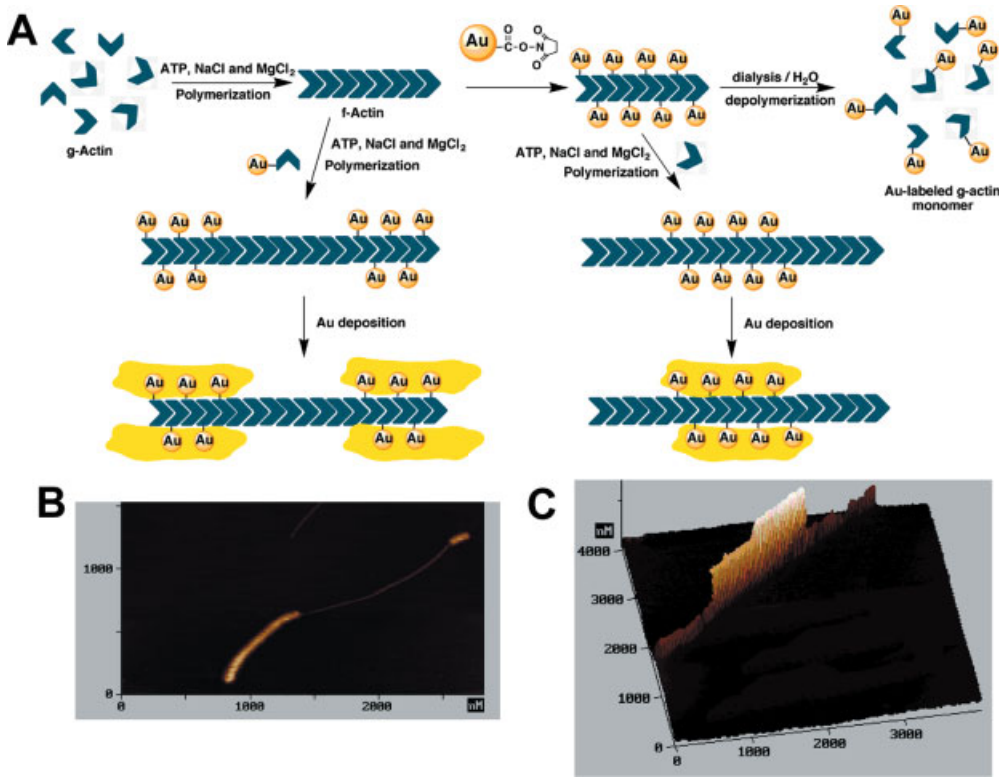


Figure 6.49 (A) Assembly of patterned actin-based Au nanowires: Left: an Au wire–actin–Au wire filament; Right: an actin–Au wire–actin filament. (B) AFM image of the Au wire–actin–Au wire filament. (C) AFM image of the actin–Au wire–actin filament. (Reprinted by permission from Macmillan Publishers Ltd: Nature Materials 234, Copyright (2004)).

as a nanotransporter driven by an external fuel. The actin–Au nanowire–actin nanostructure was deposited on a myosin-modified glass surface. Addition of ATP resulted in the motility of the nanostructures on the surface of $250 \pm 50\,\mathrm{nm\,s^{-1}}$. Thus, such metallic nano-objects conjugated to motor proteins were suggested as potential nanotransporters, where chemicals deposited on the Au cargo are carried by the proteins.

In contrast with the use of proteins as passive templates for the growth of nanowires, one can use enzymes and NPs as active hybrid systems for the synthesis of nanocircuitry and for the preparation of patterned nanostructures. The participation of enzymes in growing metal NPs and particularly the biocatalytic enlargement of metal NP seeds suggests that metal NP–enzyme hybrids could be used as active components for the synthesis of metallic nanowires. The flavoenzyme glucose oxidase was functionalized with Au NPs (1.4 nm, average loading of 12 NPs per

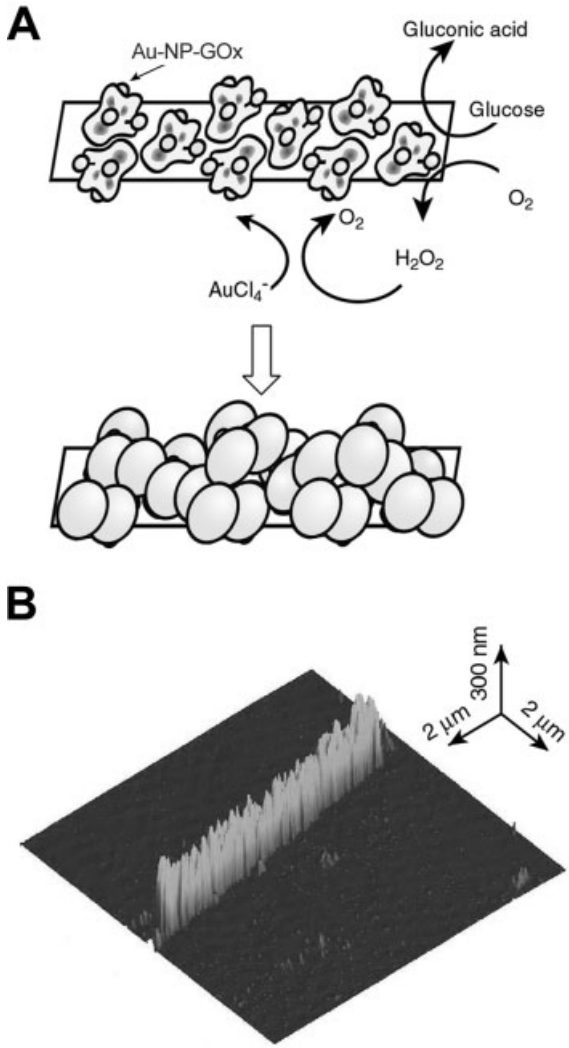


Figure 6.50 (A) Generation of an Au nanowire by the biocatalytic enlargement of an Au NP functionalized GOx line deposited on a silicon support by DPN. (B) Atomic force microscopy (AFM) image of the Au nanowire generated by the Au NP-functionalized GOx "biocatalytic ink". (Adapted with permission from [209]. Copyright 2006 Wiley-VCH).

enzyme unit) and the biocatalytic–NP hybrid material was used as a template for the stepwise synthesis of metallic nanowires [209] [Figure 6.50(A)]. The enzyme–Au NP hybrid was used as a "biocatalytic ink" for the pattering of Si surfaces using dip-pen nanolithography (DPN). The subsequent glucose-mediated generation of H_2O_2 and the catalytic enlargement of the NPs resulted in the catalytic growth of the particles,

and this yielded micrometer-long Au metallic wires exhibiting heights and widths in the region of 150–250 nm, depending on the biocatalytic "development" time interval [Figure 6.50(B)].

This process is not limited to redox enzymes, and NP-functionalized biocatalysts that transform a substrate to an appropriate reducing agent may be similarly used as "biocatalytic inks" for the generation of metallic nanowires. This was exemplified with the application of Au NP-functionalized alkaline phosphatase, AlkPh (average loading of 10 NPs per enzyme unit) as "biocatalytic ink" for the generation of Ag nanowires [209] [Figure 6.51(A)]. The alkaline phosphatase-mediated hydrolysis of *p*-aminophenol phosphate (**60**) yielded *p*-aminophenol (**61**), which reduced Ag^{+} to Ag^{0} on the Au NP seeds associated with the enzyme. This allowed the enlargement of the Au particles acting as a core for the formation of a continuous silver wires exhibiting height of 30–40 nm, on the protein template [Figure 6.51 (B)]. The method allowed the stepwise, orthogonal formation of metal nanowires composed of different metals and controlled dimensions [Figure 6.51(C)]. Furthermore, the biocatalytic growth of the nanowires exhibited a self-inhibition mechanism and upon coating of the protein with the metal no further enlargement occurred. This self-inhibition mechanism is specifically important since the dimensions of the resulting nanowires are controlled by the sizes of the biocatalytic templates.

6.11 Conclusions and Perspectives

Nanobiotechnology, and particularly, biomolecule–NP hybrid systems, that represent the functional elements of nanobiotechnology, provide an emerging interdisciplinary scientific discipline. This chapter has addressed these scientific areas by discussing numerous biomolecule–NP conjugates as functional units for biosensing, building elements of nanocircuitry and devices and composite materials for intracellular diagnostics or medical therapeutics. Biomolecule–NP hybrid systems combine the evolution-optimized recognition and reactivity properties of biomolecules with the unique electronic, catalytic and optical properties of metallic or semiconductor NPs. The biomolecule–NP hybrids provide new materials that reveal unique properties and functions as a result of the composite structures and the quantum size effects accompanying the different nano-assemblies.

The analytical applications of biomolecule–NP systems have seen tremendously advances in the last decade. The electronic properties of metallic NPs were used to electrically communicate redox proteins with electrodes and to develop amperometric biosensors. The catalytic functions of metallic NPs were broadly applied to develop amplification methods for biosensing events, and the localized plasmonic features of NPs were extensively used to develop new optical methods that probe biorecognition events and design novel biosensor configurations. Optical phenomena such as the surface-enhanced fluorescence or Raman signals by dye-modified metallic NPs, the coupling of the localized plasmon of metallic NPs with surface plasmonic waves,

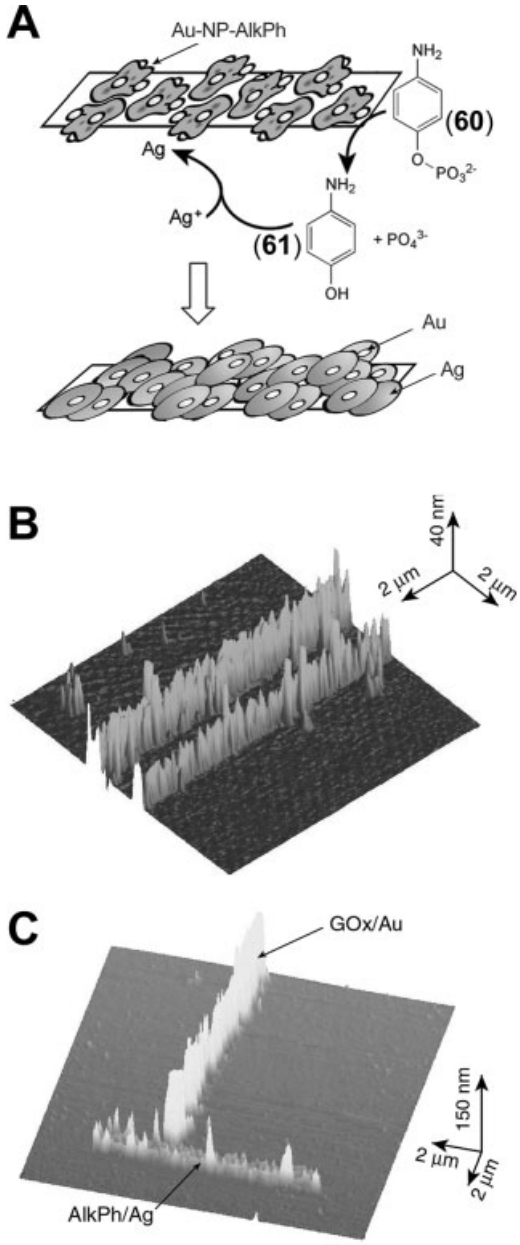


Figure 6.51 (A) Biocatalytic enlargement of Au NP-modified AlkPh with silver and fabrication of silver lines on a support using DPN and the "biocatalytic ink". (B) AFM image of Ag nanowires generated on the Au NP–AlkPh template deposited on the silicon support by DPN after 40 min of enlargement in the Ag growth solution. (C) AFM image of Au and Ag nanowires generated by deposition and enlargement of the Au NP–GOx template, followed by the passivation of the Au nanowire with mercaptoundecanoic acid and the subsequent deposition of Au NP–AlkPh template and its enlargement to the Ag nanowire. (Adapted with permission from [209]. Copyright 2006 Wiley-VCH).

the interparticle plasmon coupling in metallic NP aggregates and the reflectance of NPs were creatively used to image biorecognition events and to develop optical biosensors. Similarly, the coupling of biomolecules to semiconductor NPs (quantum dots) demonstrated the utility of these systems to assemble new optical and photoelectrochemical biosensor systems. The size-controlled emission properties of semiconductor QDs, as a result of the quantum confinement of the electronic levels in the nanoparticles, allow the multiplexed analysis of different targets and the design of high-throughput analyses in array formats. The development of photoelectrochemically based biosensors by the use of biomolecule–semiconductor hybrid systems highlights the bridge between the optical and electronic applications of biomolecule–NP hybrids for biosensing.

The development of biomolecule–NP-based sensors reached the level of practical applicability, and various analytical systems for clinical diagnostics, the analysis of food products, environmental pollutants and homeland security biosensors are expected to emerge from these hybrid nanocomposites. Some new challenging research directions in the analytical applications of biomolecule–NP hybrid systems can be identified, however. The unique optical properties of metal or semiconductor nanorods [235, 236] pave the way to new optical applications of these nano-objects. Composite nanorods may act as optical barcodes for biosensing events and could provide a rich library of labels for the parallel analysis of complex mixtures. Similarly, metal NP–semiconductor quantum dot (dumbbell) nanostructures [237] may provide new composites to assemble biomolecule hybrid systems of tailored electronic and optical functions.

Significant progress in the use of biomolecules as templates for the synthesis of nanostructures and nanocircuitry has been achieved in recent years. DNA and proteins act as efficient nanoscale templates for the synthesis of metallic or semiconductor nanowires. Although the scientific approaches to construct the nanowires are innovative, important challenges in this field are still ahead of us. The non-ohmic behavior of the conductance through nanowires and the difficulties in electrically wiring the nanowires to micro-contacts requires a fundamental experimental and theoretical understanding of charge transport phenomena across these nanostructures. The use of enzymes as catalysts that grow nanowires of predesigned structures [209] represents an important path to the further development of nanocircuitry and devices.

Numerous other applications of biomolecule–NP hybrid systems may be envisaged. The use of such systems in nanomedicine [238] has attracted increasing interest, and different applications such as photodynamic anticancer therapy [239], targeted delivery of radioisotopes [240], drug delivery [241] and gene therapy [242] have been demonstrated.

Physicists, chemists, biologists and material scientists have already recognized the limitless scientific opportunities and challenges in the applications of biomolecule–NP hybrid systems. The rapid progress in the field suggests that these interdisciplinary efforts will lead to exciting new science of immense practical and technological utility.

References

1 Mulvaney, P. (1996) *Langmuir*, **12**, 788–800.

2 Alvarez, M.M., Khoury, J.T., Schaaff, T.G., Shafigullin, M.N., Vezmar, I. and Whetten, R.L. (1997) *The Journal of Physical Chemistry B*, **101**, 3706–3712.

3 Mirkin, C.A., Letsinger, R.L., Mucic, R.C. and Storhoff, J.J. (1996) *Nature*, **382**, 607–609.

4 Aslan, K., Luhrs, C.C. and Perez-Luna, V.H. (2004) *The Journal of Physical Chemistry B*, **108**, 15631–15639.

5 Brus, L.E. (1991) *Applied Physics A*, **53**, 465–474.

6 Alivisatos, A.P. (1996) *Science*, **271**, 933–937.

7 Grieve, K., Mulvaney, P. and Grieser, F. (2000) *Current Opinion in Colloid and Interface Science*, **5**, 168–172.

8 Chan, W.C.W., Maxwell, D.J., Gao, X., Bailey, R.E., Han, M. and Nie, S. (2002) *Current Opinion in Biotechnology*, **13**, 40–46.

9 Sapsford, K.E., Pons, T., Medintz, I.L. and Mattoussi, H. (2006) *Sensors*, **6**, 925–953.

10 Katz, E., Willner, I. and Wang, J. (2004) *Electroanalysis*, **16**, 19–44.

11 Wang, J. (2003) *Analytica Chimica Acta*, **500**, 247–257.

12 Patolsky, F., Weizmann, Y., Lioubashevski, O. and Willner, I. (2002) *Angewandte Chemie-International Edition*, **41**, 2323–2327.

13 Daniel, M.-C. and Astruc, D. (2004) *Chemical Reviews*, **104**, 293–346.

14 Katz, E. and Willner, I. (2004) *Angewandte Chemie-International Edition*, **43**, 6042–6108.

15 Rosi, N.L. and Mirkin, C.A. (2005) *Chemical Reviews*, **105**, 1547–1562.

16 Wang, J. (2005) *Small*, **1**, 1036–1043.

17 Medintz, I.L., Uyeda, H.T., Goldman, E.R. and Mattoussi, H. (2005) *Nature Materials*, **4**, 435–446.

18 Pellegrino, T., Kudera, S., Liedl, T., Javier, A.M., Manna, L. and Parak, W.J. (2005) *Small*, **1**, 48–63.

19 Willner, I. (2002) *Science*, **298**, 2407–2408.

20 Bartlett, P.N., Tebbutt, P. and Whitaker, R.C. (1991) *Progress in Reaction Kinetics*, **16**, 55–155.

21 Degani, Y. and Heller, A. (1987) *Journal of Physical Chemistry*, **91**, 1285–1289.

22 Schuhmann, W., Ohara, T.J., Schmidt, H.-L. and Heller, A. (1991) *Journal of the American Chemical Society*, **113**, 1394–1397.

23 Willner, I., Riklin, A., Shoham, B., Rivenzon, D. and Katz, E. (1993) *Advanced Materials*, **5**, 912–915.

24 Heller, A. (1992) *Journal of Physical Chemistry*, **96**, 3579–3587.

25 Bu, H., Mikkelsen, S.R. and English, A.M. (1995) *Analytical Chemistry*, **67**, 4071–4076.

26 Willner, I., Katz, E., Lapidot, N. and Baüerle, P. (1992) *Bioelectrochemistry and Bioenergetics*, **29**, 29–45.

27 Willner, I., Heleg-Shabtai, V., Blonder, R., Katz, E., Tao, G., Bückmann, A.F. and Heller, A. (1996) *Journal of the American Chemical Society*, **118**, 10321–10322.

28 Zayats, M., Katz, E. and Willner, I. (2002) *Journal of the American Chemical Society*, **124**, 2120–2121.

29 Raitman, O.A., Katz, E., Bückmann, A.F. and Willner, I. (2002) *Journal of the American Chemical Society*, **124**, 6487–6496.

30 Raitman, O.A., Patolsky, F., Katz, E. and Willner, I. (2002) *Chemical Communications*, 1936–1937.

31 Willner, I. and Katz, E. (2000) *Angewandte Chemie-International Edition*, **39**, 1180–1218.

32 Murphy, L. (2006) *Current Opinion in Chemical Biology*, **10**, 177–184.

33 Heller, A. (2004) *Physical Chemistry Chemical Physics*, **6**, 209–216.

34 Willner, B., Katz, E. and Willner, I. (2006) *Current Opinion in Biotechnology*, **17**, 589–596.

35 Niemeyer, C.M. (2003) *Angewandte Chemie-International Edition*, **42**, 5796–5800.

36 Xiao, Y., Patolsky, F., Katz, E., Hainfeld, J.F. and Willner, I. (2003) *Science*, **299**, 1877–1881.

37 Zayats, M., Katz, E., Baron, R. and Willner, I. (2005) *Journal of the American Chemical Society*, **127**, 12400–12406.

38 Park, S.-J., Lazarides, A.A., Mirkin, C.A., Brazis, P.W., Kannewurf, C.R. and Letsinger, R.L. (2000) *Angewandte Chemie-International Edition*, **39**, 3845–3848.

39 Parak, W.J., Pellegrino, T., Micheel, C.M., Gerion, D., Williams, S.C. and Alivisatos, A.P. (2003) *Nano Letters*, **3**, 33–36.

40 Ghosh, S.S., Kao, P.M., McCue, A.W. and Chappelle, H.L. (1990) *Bioconjugate Chemistry*, **1**, 71–76.

41 Dubertret, B., Calame, M. and Libchaber, A.J. (2001) *Nature Biotechnology*, **19**, 365–370.

42 Niemeyer, C.M. (2001) *Chemistry – A European Journal*, **7**, 3188–3195.

43 Merkoçi, A., Aldavert, M., Marín, S. and Alegret, S. (2005) *Trends in Analytical Chemistry*, **24**, 341–349.

44 Merkoçi, A. (2007) *FEBS Journal*, **274**, 310–316.

45 Wang, J. (1985) *Stripping Analysis*, VCH, Weinheim.

46 Gonzáles-García, M.B. and Costa-García, A. (1995) *Bioelectrochemistry and Bioenergetics*, **38**, 389–395.

47 González-García, M.B., Fernández-Sánchez, C. and Costa-García, A. (2000) *Biosensors & Bioelectronics*, **15**, 315.

48 Ozsoz, M., Erdem, A., Kerman, K., Ozkan, D., Tugrul, B., Topcuoglu, N., Ekren, H. and Taylan, M. (2003) *Analytical Chemistry*, **75**, 2181–2187.

49 Wang, J., Xu, D., Kawde, A.-N. and Polsky, R. (2001) *Analytical Chemistry*, **73**, 5576–5581.

50 Authier, L., Grossirod, C., Brossier, P. and Limoges, B. (2001) *Analytical Chemistry*, **73**, 4450–4456.

51 Cai, H., Xu, Y., Zhu, N., He, P. and Fang, Y. (2002) *Analyst*, **127**, 803–808.

52 Dequire, M., Degrand, C. and Limoges, B. (2000) *Analytical Chemistry*, **72**, 5521–5528.

53 Wang, J., Polsky, R. and Xu, D. (2001) *Langmuir*, **17**, 5739–5741.

54 Lee, T.M.-H., Li, L.-L. and Hsing, I.-M. (2003) *Langmuir*, **19**, 4338–4343.

55 Li, L.-L., Cai, H., Lee, M.-H., Barford, J. and Hsing, I.-M. (2004) *Electroanalysis*, **16**, 81–87.

56 Wang, J. and Kawde, A.-N. (2002) *Electrochemistry Communications*, **4**, 349–352.

57 Wang, J., Xu, D. and Polsky, R. (2002) *Journal of the American Chemical Society*, **124**, 4208–4209.

58 Pumera, M., Castañeda, M.T., Pividori, M.I., Eritja, R., Merkoçi, A. and Alegret, S. (2005) *Langmuir*, **21**, 9625–9629.

59 Kawde, A. and Wang, J. (2004) *Electroanalysis*, **16**, 101–107.

60 Wang, J., Li, J., Baca, A.J., Hu, J., Zhou, F., Yan, W. and Pang, D.W. (2003) *Analytical Chemistry*, **75**, 3941–3945.

61 Wang, J., Rincón, O., Polsky, R. and Dominguez, E. (2003) *Electrochemistry Communications*, **5**, 83–86.

62 Park, S.-J., Taton, T.A. and Mirkin, C.A. (2002) *Science*, **295**, 1503–1506.

63 Velev, O.D. and Kaler, E.W. (1999) *Langmuir*, **15**, 3693–3698.

64 Polsky, R., Gill, R., Kaganovsky, L. and Willner, I. (2006) *Analytical Chemistry*, **78**, 2268–2271.

65 Buttry, D.A. and Ward, M.D. (1992) *Chemical Reviews*, **92**, 1355–1379.

66 Zhou, X.C., O'Shea, S.J. and Li, S.F.Y. (2000) *Chemical Communications*, 953–954.

67 Patolsky, F., Ranjit, K.T., Lichtenstein, A. and Willner, I. (2000) *Chemical Communications*, 1025–1026.

68 Liu, T., Tang, J. and Jiang, L. (2004) *Biochemical and Biophysical Research Communications*, **313**, 3–7.

69 Han, S., Lin, J., Satjapipat, M., Baca, A.J. and Zhou, F. (2001) *Chemical Communications*, 609–610.

70 Willner, I., Patolsky, F., Weizmann, Y. and Willner, B. (2002) *Talanta*, **56**, 847–856.

71 Weizmann, Y., Patolsky, F. and Willner, I. (2001) *Analyst*, **126**, 1502–1504.

72 Pavlov, V., Xiao, Y., Shlyahovsky, B. and Willner, I. (2004) *Journal of the American Chemical Society*, **126**, 11768–11769.

73 Bock, L.C., Griffin, L.C., Latham, J.A., Vermass, E.H. and Toole, J.J. (1992) *Nature*, **355**, 564–566.

74 Tasset, D.M., Kubik, M.F. and Steiner, W. (1997) *Journal of Molecular Biology*, **272**, 688–698.

75 Xiao, Y., Pavlov, V., Levine, S., Niazov, T., Markovitch, G. and Willner, I. (2004) *Angewandte Chemie-International Edition*, **43**, 4519–4522.

76 Wang, J., Liu, G., Polsky, R. and Merkoçi, A. (2002) *Electrochemistry Communications*, **4**, 722–726.

77 Wang, J., Liu, G. and Merkoçi, A. (2003) *Journal of the American Chemical Society*, **125**, 3214–3215.

78 Liu, G., Wang, J., Kim, J., Jan, M. and Collins, G. (2004) *Analytical Chemistry*, **76**, 7126–7130.

79 Wang, J., Liu, G. and Zhu, Q. (2003) *Analytical Chemistry*, **75**, 6218–6222.

80 Hansen, J.A., Wang, J., Kawde, A.-N., Xiang, Y., Gothelf, K.V. and Collins, G. (2006) *Journal of the American Chemical Society*, **128**, 2228–2229.

81 Wang, J., Lee, T. and Liu, G. (2005) *Journal of the American Chemical Society*, **127**, 38–39.

82 Hutter, E. and Fendler, J.H. (2004) *Advanced Materials*, **16**, 1685–1706.

83 Karlsson, R. (2004) *Journal of Molecular Recognition*, **17**, 151–161.

84 Englebienne, P., Hoonacker, A.V. and Verhas, M. (2003) *Spectroscopy*, **17**, 255–273.

85 Shankaran, D.R., Gobi, K.V. and Miura, N. (2007) *Sensors and Actuators B*, **121**, 158–177.

86 Schultz, D.A. (2003) *Current Opinion in Biotechnology*, **14**, 13–22.

87 Elghanian, R., Storhoff, J.J., Mucic, R.C., Letsinger, R.L. and Mirkin, C.A. (1997) *Science*, **277**, 1078–1081.

88 Storhoff, J.J., Elghanian, R., Mucic, R.C., Mirkin, C.A. and Letsinger, R.L. (1998) *Journal of the American Chemical Society*, **120**, 1959–1964.

89 Reynolds, R.A., III, Mirkin, C.A. and Letsinger, R.L. (2000) *Journal of the American Chemical Society*, **122**, 3795–3796.

90 Souza, G.R. and Miller, J.H. (2001) *Journal of the American Chemical Society*, **123**, 6734–6735.

91 Jin, R., Wu, G., Li, Z., Mirkin, C.A. and Schatz, G.C. (2003) *Journal of the American Chemical Society*, **125**, 1643–1654.

92 Beissenhirtz, M.K. and Willner, I. (2006) *Organic & Biomolecular Chemistry*, **4**, 3392–3401.

93 Weizmann, Y., Beissenhirtz, M.K., Cheglakov, Z., Nowarski, R., Kotler, M. and Willner, I. (2006) *Angewandte Chemie-International Edition*, **45**, 7384–7388.

94 Beissenhirtz, M.K., Elnathan, R., Weizmann, Y. and Willner, I. (2007) *Small*, **3**, 375–379.

95 Sato, K., Hosokawa, K. and Maeda, M. (2003) *Journal of the American Chemical Society*, **125**, 8102–8103.

96 Sato, K., Hosokawa, K. and Maeda, M. (2007) *Analytical Sciences*, **23**, 17–20.

97 Li, H. and Rothberg, L. (2004) *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 14036–14039.

98 Li, H. and Rothberg, L. (2004) *Journal of the American Chemical Society*, **126**, 10958–10961.

99 Ellington, A.D. and Szostak, J.W. (1990) *Nature*, **346**, 818–822.

100 Tuerk, C. and Gold, L. (1990) *Science*, **249**, 505–510.

101 Wilson, D.S. and Szostak, J.W. (1999) *Annual Review of Biochemistry*, **68**, 611–647.

102 Schlatterer, J.C., Stuhlmann, F. and Jäschke, A. (2003) *ChemBioChem*, **4**, 1089–1092.

103 Breaker, R.R. and Joyce, G.F. (1994) *Chemistry & Biology*, **1**, 223–229.

104 Cuenoud, B. and Szostak, J.W. (1995) *Nature*, **375**, 611–614.

105 Santoro, S.W., Joyce, G.F., Sakthivel, K., Gramatikova, S. and Barbas, C.F. III (2000) *Journal of the American Chemical Society*, **122**, 2433–2439.

106 Travascio, P., Witting, P.K., Mauk, A.G. and Sen, D. (2001) *Journal of the American Chemical Society*, **123**, 1337–1348.

107 Travascio, P., Li, Y. and Sen, D. (1998) *Chemistry & Biology*, **5**, 505–517.

108 Santoro, S.W. and Joyce, G.F. (1997) *Proceedings of the National Academy of Sciences of the United States of America*, **94**, 4262–4266.

109 Faulhammer, D. and Famulok, M. (1996) *Angewandte Chemie-International Edition in English*, **35**, 2837–2841.

110 Liu, J. and Lu, Y. (2003) *Journal of the American Chemical Society*, **125**, 6642–6643.

111 Liu, J. and Lu, Y. (2004) *Analytical Chemistry*, **76**, 1627–1632.

112 Liu, J. and Lu, Y. (2006) *Angewandte Chemie-International Edition*, **45**, 90–94.

113 Guarise, C., Pasquato, L., De Filillis, V. and Scrimin, P. (2006) *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 3978–3982.

114 Laromaine, A., Koh, L., Murugesan, M., Ulijn, R.V. and Stevens, M.M. (2007) *Journal of the American Chemical Society*, **129**, 4156–4157.

115 Choi, Y., Ho, N.-H. and Tung, C.-H. (2007) *Angewandte Chemie-International Edition*, **46**, 707–709.

116 Schofield, C.L., Field, R.A. and Russell, D.A. (2007) *Analytical Chemistry*, **79**, 1356–1361.

117 Huang, C.-C., Huang, Y.-F., Cao, Z., Tan, W. and Chang, H.-T. (2005) *Analytical Chemistry*, **77**, 5735–5741.

118 Taton, T.A., Mirkin, C.A. and Letsinger, R.L. (2000) *Science*, **289**, 1757–1760.

119 Taton, T.A., Lu, G.L. and Mirkin, C.A. (2001) *Journal of the American Chemical Society*, **123**, 5164–5165.

120 Storhoff, J.J., Marla, S.S., Bao, P., Hagenow, S., Mehta, H., Lucas, A., Garimella, V., Patno, T., Buckingham, W., Cork, W. and Müller, U.R. (2004) *Biosensors & Bioelectronics*, **19**, 875–883.

121 Bao, P., Huber, M., Wei, T.-F., Marla, S.S., Storhoff, J.J. and Müller, U.R. (2005) *Nucleic Acids Research*, **33**, e15.

122 Nam, J.-M., Thaxton, C.S. and Mirkin, C.A. (2003) *Science*, **301**, 1884–1886.

123 Nam, J.-M., Stoeva, S.I. and Mirkin, C.A. (2004) *Journal of the American Chemical Society*, **126**, 5932–5933.

124 Stoeva, S.I., Lee, J.-S., Smith, J.E., Rosen, S.T. and Mirkin, C.A. (2006) *Journal of the American Chemical Society*, **128**, 8378–8379.

125 Georganopoulou, D.G., Chang, L., Nam, J.M., Thaxton, C.S., Mufson, E.J., Klein, W.L. and Mirkin, C.A. (2005) *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 2273–2276.

126 Stoeva, S.I., Lee, J.-S., Thaxton, C.S. and Mirkin, C.A. (2006) *Angewandte Chemie-International Edition*, **45**, 3303–3306.

127 Oh, B.-K., Nam, J.-M., Lee, S.W. and Mirkin, C.A. (2006) *Small*, **2**, 103–108.

128 Nam, J.-M., Wise, A.R. and Groves, J.T. (2005) *Analytical Chemistry*, **77**, 6985–6988.

129 Storhoff, J.J., Lucas, A.D., Garimella, V., Bao, Y.P. and Müller, U.R. (2004) *Nature Biotechnology*, **22**, 883–887.

130 Du, B.-A., Li, Z.-P. and Liu, C.-H. (2006) *Angewandte Chemie-International Edition*, **45**, 8022–8025.

131 Ray, P.C. (2006) *Angewandte Chemie-International Edition*, **45**, 1151–1154.

132 Kneipp, K., Kneipp, H., Itzkan, I., Dasari, R.R. and Feld, M.S. (1999) *Chemical Reviews*, **99**, 2957–2975.

133 Moskovits, M. (2005) *Journal of Raman Spectroscopy*, **36**, 485–496.

134 Haynes, C.L., McFarland, A.D. and Van Duyne, R.P. (2005) *Analytical Chemistry*, **77**, 338A–346A.

135 Faulds, K., Smith, W.E. and Graham, D. (2004) *Analytical Chemistry*, **76**, 412–417.

136 Graham, D., Faulds, K. and Smith, W.E. (2006) *Chemical Communications*, 4363–4371.

137 Faulds, K., Stewart, L., Smith, W.E. and Graham, D. (2005) *Talanta*, **67**, 667–671.

138 Faulds, K., Fruk, L., Robson, D.C., Thompson, D.G., Enright, A., Smith, W.E. and Graham, D. (2006) *Faraday Discussions*, **132**, 261–268.

139 Cao, Y.C., Jin, R. and Mirkin, C.A. (2002) *Science*, **297**, 1536–1540.

140 Cao, Y.C., Jin, R., Nam, J.-M., Thaxton, C.S. and Mirkin, C.A. (2003) *Journal of the American Chemical Society*, **125**, 14676–14677.

141 Grubisha, D.S., Lipert, R.J., Park, H.-Y., Driskell, J. and Porter, M.D. (2003) *Analytical Chemistry*, **75**, 5936–5943.

142 Driskell, J.D., Kwarta, K.M., Lipert, R.J. and Porter, M.D. (2005) *Analytical Chemistry*, **77**, 6147–6154.

143 Xu, S., Ji, X., Xu, W., Li, X., Wang, L., Bai, Y., Zhao, B. and Ozaki, Y. (2004) *Analyst*, **129**, 63–68.

144 Lyon, L.A., Musick, M.D. and Natan, M.J. (1998) *Analytical Chemistry*, **70**, 5177–5183.

145 Englebienne, P., Hoonacker, A.V. and Verhas, M. (2001) *Analyst*, **126**, 1645–1651.

146 Huang, L., Reekmans, G., Saerens, D., Friedt, J.-M., Frederix, F., Francis, L., Muyldermans, S., Campitelli, A. and van Hoof, C. (2005) *Biosensors & Bioelectronics*, **21**, 483–490.

147 He, L., Musick, M.D., Nicewarner, S.R., Salinas, F.G., Benkovic, S.J., Natan, M.J. and Keating, C.D. (2000) *Journal of the American Chemical Society*, **122**, 9071–9077.

148 Yao, X., Li, X., Toledo, F., Zurita-Lopez, C., Gutova, M., Momand, J. and Zhou, F. (2006) *Analytical Biochemistry*, **354**, 220–228.

149 Sato, Y., Sato, K., Hosokawa, K. and Maeda, M. (2006) *Analytical Biochemistry*, **355**, 125–131.

150 Yang, X., Wang, Q., Wang, K., Tan, W. and Li, H. (2007) *Biosensors & Bioelectronics*, **22**, 1106–1110.

151 Lioubashevski, O., Chegel, V., Patolsky, F., Katz, E. and Willner, I. (2004) *Journal of the American Chemical Society*, **126**, 7133–7143.

152 El-Sayed, M.A. (2001) *Accounts of Chemical Research*, **34**, 257–264.

153 Yonzon, C.R., Zhang, X., Zhao, J. and Van Duyne, R.P. (2007) *Spectroscopy*, **22**, 42–56.

154 Moores, A. and Goettmann, F. (2006) *New Journal of Chemistry*, **30**, 1121–1132.

155 Haes, A.J. and Van Duyne, R.P. (2004) *Analytical and Bioanalytical Chemistry*, **379**, 920–930.

156 Zhao, J., Zhang, X., Yonzon, C.R., Haes, A.J. and Van Duyne, R.P. (2006) *Nanomedicine*, **1**, 219–228.

157 Haes, A.J. and Van Duyne, R.P. (2002) *Journal of the American Chemical Society*, **124**, 10596–10604.

158 Riboh, J.C., Haes, A.J., McFarland, A.D., Yonzon, C.R. and Van Duyne, R.P. (2003) *The Journal of Physical Chemistry B*, **107**, 1772–1780.

159 Haes, A.J., Hall, W.P., Chang, L., Klein, W.L. and Van Duyne, R.P. (2004) *Nano Letters*, **4**, 1029–1034.

160 Haes, A.J., Chang, L., Klein, W.L. and Van Duyne, R.P. (2005) *Journal of the American Chemical Society*, **127**, 2264–2271.

161 Endo, T., Kerman, K., Nagatani, N., Takamura, Y. and Tamiya, E. (2005) *Analytical Chemistry*, **77**, 6976–6984.

162 Endo, T., Kerman, K., Nagatani, N., Hiepa, H.M., Kim, D.-K., Yonezawa, Y. and Tamiya, E. (2006) *Analytical Chemistry*, **78**, 6465–6475.

163 Alivisatos, A.P. (2004) *Nature Biotechnology*, **22**, 47–52.

164 Nirmal, M. and Brus, L.E. (1999) *Accounts of Chemical Research*, **32**, 407–414.

165 Niemeyer, C.M. (2001) *Angewandte Chemie-International Edition*, **40**, 4128–4158.

166 Katz, E., Shipway, A.N., and Willner, I. (2003) in *Nanoscale Materials*, (eds L.M.

Liz-Marzan and P. Kamat), Kluwer, Dordrecht, Chapter 2, pp. 5–78.

167 Michalet, X., Pinaud, F.F., Bentolila, L.A., Tsay, J.M., Doose, S., Li, J.J., Sundaresan, G., Wu, A.M., Gambhir, S.S. and Weiss, S. (2005) *Science*, **307**, 538–544.

168 Costa-Fernandez, J.M., Pereiro, R. and Sanz-Medel, A. (2006) *Trends in Analytical Chemistry*, **25**, 207–218.

169 Klostranec, J.M. and Chan, W.C.W. (2006) *Advanced Materials*, **18**, 1953–1964.

170 Murphy, C.J. (2002) *Analytical Chemistry*, **74**, 520A–526A.

171 Goldman, E.R., Balighian, E.D., Mattoussi, H., Kuno, M.K., Mauro, J.M., Tran, P.T. and Anderson, G.P. (2002) *Journal of the American Chemical Society*, **124**, 6378–6382.

172 Goldman, E.R., Anderson, G.P., Tran, P.T., Mattoussi, H., Charles, P.T. and Mauro, J.M. (2002) *Analytical Chemistry*, **74**, 841–847.

173 Goldman, E.R., Clapp, A.R., Anderson, G.P., Uyeda, H.T., Mauro, J.M., Medintz, I.L. and Mattoussi, H. (2004) *Analytical Chemistry*, **76**, 684–688.

174 Yang, L. and Li, Y. (2006) *Analyst*, **131**, 394–401.

175 Hoshino, A., Fujioka, K., Manabe, N., Yamay, S., Goto, Y., Yasuhara, M. and Yamamoto, K. (2005) *Microbiology and Immunology*, **49**, 461–470.

176 Gerion, D., Chen, F., Kannan, B., Fu, A., Parak, W.J., Chen, D.J., Majumdar, A. and Alivisatos, A.P. (2003) *Analytical Chemistry*, **75**, 4766–4772.

177 Pathak, S., Choi, S.-K., Arnheim, N. and Thompson, M.E. (2001) *Journal of the American Chemical Society*, **123** 4103–4104.

178 Xiao, Y. and Barker, P.E. (2004) *Nucleic Acids Research*, **32**, e28.

179 Chan, P.M., Yuen, T., Ruf, F., Gonzalez-Maeso, J. and Sealfon, S.C. (2005) *Nucleic Acids Research*, **33**, e161.

180 Mattoussi, H., Mauro, J.M., Goldman, E.R., Anderson, G.P., Sunder, V.C., Mikulec, F.V. and Bawendi, M.G. (2000) *Journal of the American Chemical Society*, **122**, 12142–12150.

181 Warren, C.W. and Nie, S. (1998) *Science*, **281**, 2016–2018.

182 Gerion, D., Pinaud, F., Willimas, S.C., Parak, W.J., Zanchet, D., Weiss, S. and Alivisatos, A.P. (2001) *The Journal of Physical Chemistry B*, **105**, 8861–8871.

183 Bruchez, M., Jr. Moronne, M., Gin, P., Weiss, S. and Alivisatos, A.P. (1998) *Science*, **281**, 2013–2016.

184 Gao, X., Cui, Y., Levenson, R.M., Chung, L.W.K. and Nie, S. (2004) *Nature Biotechnology*, **22**, 969–976.

185 Clapp, A.R., Medintz, I.L. and Mattoussi, H. (2006) *ChemPhysChem*, **7**, 47–57.

186 Sapsford, K.E., Berti, L. and Medintz, I.L. (2006) *Angewandte Chemie-International Edition*, **45**, 4562–4588.

187 Patolsky, F., Gill, R., Weizmann, Y., Mokari, T., Banin, U. and Willner, I. (2003) *Journal of the American Chemical Society*, **125**, 13918–13919.

188 Medintz, I.L., Clapp, A.R., Mattoussi, H., Goldman, E.R., Fisher, B. and Mauro, J.M. (2003) *Nature Materials*, **2**, 630–638.

189 Goldman, E.R., Medintz, I.L., Whitley, J.L., Hayhurst, A., Clapp, A.R., Uyeda, H.T., Deschamps, J.R., Lassman, M.E. and Mattoussi, H. (2005) *Journal of the American Chemical Society*, **127**, 6744–6751.

190 Medintz, I.L., Clapp, A.R., Brunel, F.M., Tiefenbrunn, T., Uyeda, H.T., Chang, E.L., Deschamps, J.R., Dawson, P.E. and Mattoussi, H. (2006) *Nature Materials*, **5**, 581–589.

191 Shi, L., De Paoli, V., Rosenzweig, N. and Rosenzweig, Z. (2006) *Journal of the American Chemical Society*, **128**, 10378–10379.

192 Gill, R., Freeman, R., Xu, J., Willner, I., Winograd, S., Shweky, I. and Banin, U. (2006) *Journal of the American Chemical Society*, **128**, 15376–15377.

193 Gill, R., Willner, I., Shweky, I. and Banin, U. (2005) *The Journal of Physical Chemistry B*, **109**, 23715–23719.

194 Levy, M., Cater, S.F. and Ellington, A.D. (2005) *ChemBioChem*, **6**, 2163–2166.

195 Willner, I., Patolsky, F. and Wasserman, J. (2001) *Angewandte Chemie-International Edition*, **40**, 1861–1864.

196 Gill, R., Patolsky, F., Katz, E. and Willner, I. (2005) *Angewandte Chemie-International Edition*, **44**, 4554–4557.

197 Pardo-Yissar, V., Katz, E., Wasserman, J. and Willner, I. (2003) *Journal of the American Chemical Society*, **125**, 622–623.

198 Katz, E., Zayats, M., Willner, I. and Lisdat, F. (2006) *Chemical Communications*, 1395–1397.

199 Stoll, C., Kudera, S., Parak, W.J. and Lisdat, F. (2006) *Small*, **2**, 741–743.

200 Klaus, T., Joerger, R., Olsson, E. and Granqvist, C.G. (1999) *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 13611–13614.

201 Gardea-Torresdey, J.L., Parsons, J.G., Gomez, E., Peralta-Videa, J., Troiani, H.E., Santiago, P. and Yacaman, M.J. (2002) *Nano Letters*, **2**, 397–401.

202 Ahmad, A., Senapati, S., Khan, M., Kumar, R. and Sastry, M. (2003) *Langmuir*, **19**, 3550–3553.

203 Shankar, S.S., Ahmad, A. and Sastry, M. (2003) *Biotechnology Progress*, **13**, 1627–1631.

204 Ahmad, A., Mukherjee, P., Mandal, D., Senapati, S., Khan, M.I., Kumar, R. and Sastry, M. (2002) *Journal of the American Chemical Society*, **124**, 12108–12109.

205 Rai, A., Singh, A., Ahmad, A. and Sastry, M. (2006) *Langmuir*, **22**, 736–741.

206 Chandran, S.P., Chaudhary, M., Pasricha, R., Ahmad, A. and Sastry, M. (2006) *Biotechnology Progress*, **22**, 577–583.

207 Willner, I., Baron, R. and Willner, B. (2006) *Advanced Materials*, **18**, 1109–1120.

208 Zayats, M., Baron, R., Popov, I. and Willner, I. (2005) *Nano Letters*, **5**, 21–25.

209 Basnar, B., Weizmann, Y., Cheglakov, Z. and Willner, I. (2006) *Advanced Materials*, **18**, 713–718.

210 Baron, R., Zayats, M. and Willner, I. (2005) *Analytical Chemistry*, **77**, 1566–1571.

211 Angeletti, C., Khomitch, V., Halaban, R. and Rimm, D.L. (2004) *Diagnostic Cytopathology*, **31**, 33–37.

212 Pavlov, V., Xiao, Y. and Willner, I. (2005) *Nano Letters*, **5**, 649–653.

213 Xiao, Y., Pavlov, V., Shlyahovsky, B. and Willner, I. (2005) *Chemistry – A European Journal*, **11**, 2698–2704.

214 Shlyahovsky, B., Katz, E., Xiao, Y., Pavlov, V. and Willner, I. (2005) *Small*, **1**, 213–216.

215 Xiao, Y., Shlyahovsky, B., Popov, I., Pavlov, V. and Willner, I. (2005) *Langmuir*, **21**, 5659–5662.

216 Gazit, E. (2007) *FEBS Journal*, **274**, 317–322.

217 Baron, R., Willner, B. and Willner, I. (2007) *Chemical Communications*, 323–332.

218 Kovtyukhova, N.I. and Mallouk, T.E. (2002) *Chemistry – A European Journal*, **8**, 4354–4363.

219 Seeman, N.C. (1998) *Annual Review of Biophysics and Biomolecular Structure*, **27**, 225–248.

220 Seeman, N.C. (2005) *Trends in Biochemical Sciences*, **30**, 119–125.

221 Gu, Q., Cheng, C., Gonela, R., Suryanarayanan, S., Anabathula, S., Dai, K. and Haynie, D.T. (2006) *Nanotechnology*, **17**, R14–R25.

222 Braun, E., Eichen, Y., Sivan, U. and Ben-Yoseph, G. (1998) *Nature*, **391**, 775–778.

223 Monson, C.F. and Wooley, A.T. (2003) *Nano Letters*, **3**, 359–363.

224 Merting, M., Colombi Ciacchi, L., Seidel, R., Pompe, W. and De Vita, A. (2002) *Nano Letters*, **2**, 841–844.

225 Richter, J., Mertig, M., Pompe, W., Mönch, I. and Schackert, H.K. (2001) *Applied Physics Letters*, **78**, 536–538.

226 Weizmann, Y., Patolsky, F., Popov, I. and Willner, I. (2004) *Nano Letters*, **4**, 787–792.

227 Keren, K., Berman, R.S., Buchstab, E., Sivan, U. and Braun, E. (2003) *Science*, **302**, 1380–1382.

228 Keren, K., Krueger, M., Gilad, R., Ben-Yoseph, G., Sivan, U. and Braun, E. (2002) *Science*, **297**, 72–75.

229 Yan, H., Park, S.H., Finkelstein, G., Reif, J.H. and LaBean, T.H. (2003) *Science*, **301**, 1882–1884.

230 Reches, M. and Gazit, E. (2003) *Science*, **300**, 625–627.

231 Carny, O., Shalev, D.E. and Gazit, E. (2006) *Nano Letters*, **6**, 1594–1597.

232 Vale, R.D. (2003) *Journal of Cell Biology*, **163**, 445–450.

233 dos Remedios, C.G. and Moens, P.D.J. (1995) *Biochimica et Biophysica Acta*, **1228**, 99–124.

234 Patolsky, F., Weizmann, Y. and Willner, I. (2004) *Nature Materials*, **3**, 692–695.

235 Murphy, C.J., San, T.K., Gole, A.M., Orendorff, C.J., Gao, J.X., Gou, L., Hunyadi, S.E. and Li, T. (2005) *The Journal of Physical Chemistry B*, **109**, 13857–13870.

236 El-Sayed, M.A. (2004) *Accounts of Chemical Research*, **37**, 326–333.

237 Mokari, T., Rothenberg, E., Popov, I., Costi, R. and Banin, U. (2004) *Science*, **304**, 1787–1790.

238 Moghimi, S.M., Hunter, A.C. and Murray, J.C. (2005) *FASEB Journal*, **19**, 311–330.

239 Samia, A.C.S., Chen, X. and Burda, C. (2003) *Journal of the American Chemical Society*, **125**, 15736–15737.

240 Lockman, P.R., Oyewumi, M.O., Koziara, J.M., Roder, K.E., Mumper, R.J. and Allen, D.D. (2003) *Journal of Controlled Release*, **93**, 271–282.

241 Allen, T.M. and Cullis, P.R. (2004) *Science*, **303**, 1818–1822.

242 Miller, A.D. (2004) *ChemBioChem*, **5**, 53–54.

7
Philosophy of Nanotechnoscience

Alfred Nordmann

7.1
Introduction: Philosophy of Science and of Technoscience

One way or another, the philosophy of science always informs and reflects the development of science and technology. It appears in the midst of disputes over theories and methods, in the reflective thought of scientists and since the late nineteenth century also in the analyses of so-called philosophers of science. Four philosophical questions, in particular, are answered implicitly or contested explicitly by any scientific endeavor:

- How is a particular science to be defined and what are the objects and problems in its domain of interest?
- What is the methodologically proper or specifically scientific way of approaching these objects and problems?
- What kind of knowledge is produced and communicated, how does it attain objectivity, if not certainty, and how does it balance the competing demands of universal generality and local specificity?
- What is its place in relation to other sciences, where do its instruments and methods, its concepts and theories come from and should its findings be explained on a deeper level by more fundamental investigations?

When researchers publish their results, when they review and critique their peers, argue for research funds or train graduate students, they offer examples of what they consider good scientific practice and thereby adopt a stance on all four questions. When, for example, there is a call for more basic research on some scientific question, one can look at the argument that is advanced and discover how it is informed by a particular conception of science and the relation of science and technology. Frequently it involves the idea that basic science identifies rather general laws of causal relations. These laws can then be applied in a variety of contexts and the deliberate control of causes and effects can give rise to new technical devices. If one encounters such an argument for basic science, one might ask, of course, whether this picture of

Nanotechnology. Volume 1: Principles and Fundamentals. Edited by Günter Schmid

ISBN: 978-3-527-31732-5

basic versus applied science is accurate. While it may hold here and there, especially in theoretical physics, it is perhaps altogether inadequate for chemistry. And thus one may find that the implicit assumptions agree less with the practice and history of science and more with a particular self-understanding of science. According to this self-understanding, basic science disinterestedly studies the world as it is, whereas the engineering sciences apply this knowledge to change the world in accordance with human purposes.

Science and scientific practice are always changing as new instruments are invented, new problems arise, new disciplines emerge. Also, the somewhat idealized self-understandings of scientists can change. The relation of science and technology provides a case point. Is molecular electronics a basic science? Is nanotechnology applied nanoscience? Are the optical properties of carbon nanotubes part of the world as it is or do they appear only in the midst of a large-scale engineering pursuit that is changing the world according to human purposes? There are no easy or straightforward answers to these questions and this is perhaps due to the fact that the traditional ways of distinguishing science and technology, and basic and applied research, do not work any longer. As many authors are suggesting, we should speak of "technoscience" [1, 2] which is defined primarily by the interdependence of theoretical observation and technical intervention [3].[1)] Accordingly, the designation "nanotechnoscience" is more than shorthand for "nanoscience and nanotechnologies" but signifies a mode of research other than traditional science and engineering. Peter Galison, for example, notes that "[n]anoscientists aim to build – not to demonstrate existence. They are after an engineering way of being in science" [5]. Others appeal to the idea of a "general purpose technology" and thus suggest that nanotechnoscience is fundamental research to enable a new technological development at large. Richard Jones sharpens this when he succinctly labels at least some nanotechnoscientific research as "basic gizmology."[2)]

Often, nanoscience is defined as an investigation of scale-dependently discontinuous properties or phenomena [6]. This definition of nanoscience produces in its wake an ill-defined conception of nanotechnologies – these encompass all possible technical uses of these properties and phenomena. In its 2004 report on nanoscience and nanotechnologies, the Royal Society and Royal Academy of Engineering defines these terms as follows:

1) This is in reference to Ian Hacking's distinc-tion of "representing" and "intervening" [4]: In technoscientific research, the business of theoretical representation cannot be dissociated, even in principle, from the material conditions of knowledge production and thus from the interventions that are required to make and stabilize the phenomena. In other words, technoscience knows only one way of gaining new knowledge and that is by first making a new world. If the business of science is the theoretical representation of an eternal and immutably given nature and if the business of technology is to control the world, to intervene and change the "natural" course of events, "technoscience" is a hybrid where theoretical representation becomes entangled with technical intervention.

2) Jones used this term in conversation (and on his web site, www.softmachines.org) and referred, for example, to Nadrian Seeman's systematic exploration of DNA as a building block or component of future technical systems.

> "*Nanoscience is the study of phenomena and manipulation of materials at atomic, molecular and macromolecular scales, where properties differ significantly from those at a larger scale. Nanotechnologies are the design, characterisation, production and application of structures, devices and systems by controlling shape and size at the nanometre scale" [7].*

The notion of "nanotechnoscience" does not contradict such definitions but assumes a different perspective – it looks from within the organization of research where fundamental capabilities are typically acquired in the context of funded projects with a more or less concretely imagined technical goal. This is what Galison means by an engineering way of being in science. Even though a great deal of scientific knowledge and experience goes into the acquisition of such capabilities and the investigation of novel phenomena, it is not quite "science" because the point of this investigation is not normally to question received conceptions and to establish new truths, nor is it to produce and test hypotheses or to develop theories that close important gaps in our understanding of the world. And even though nanoscale research practice involves a good bit of tinkering and pursues technological challenges and promises, it is also not "engineering" because most researchers are not in the business of building devices for more or less immediate use. At best, they lay the groundwork for concrete engineering projects in the future.

For this "engineering way of being in science", a philosophy of technoscience is needed that asks for nanotechnological, biomedical or semiconductor research the four questions that were identified above: what is the role of theory and theory-development in nanoscale research and what kinds of theories are needed for nanotechnological development?; what are the preferred methods and tools and the associated modes of reasoning in nanoscientific research?; what is nanotechnoscience and how are its objects constituted?; and what kind of knowledge do technoscientific researchers typically produce and communicate? The four main sections of this chapter will address these questions – and in all four cases, strictly philosophical considerations will shade into societal dimensions and questions of value. That this is so is due to the fact that there may have been "pure science" but that there is no such thing as "pure technoscience." Indeed, one way of characterizing technoscience is by noting that academic laboratory research is no longer answerable just to standards of peer researchers but has entered the "ethical space" of engineering with its accountability also to patrons and clients, to developers and users [8, 9].

7.2
From "Closed Theories" to Limits of Understanding and Control

7.2.1
Closed Relative to the Nanoscale

In the late 1940s, the physicist Werner Heisenberg introduced the notion of "closed theories". In particular, he referred to four closed theories: "Newtonian mechanics,

Maxwell's theory with the special theory of relativity, thermodynamics and statistical mechanics, non-relativistic quantum mechanics with atomic physics and chemistry". These theories he considered to be closed in four complementary respects:

1. Their historical development has come to an end, they are finished or have reached their final form.
2. They constitute a hermetically closed domain in that the theory defines conditions of applicability such that the theory will be true wherever its concepts can be applied.
3. They are immune to criticism; problems that arise in contexts of application are deflected to auxiliary theories and hypotheses or to the specifics of the set-up, the instrumentation, and so on.
4. They are forever valid: wherever and whenever experience can be described with the concepts of such a theory, the laws postulated by this theory will be proven correct [10].[3)]

All this holds for nanotechnoscience: It draws on an available repertoire of theories that are closed or considered closed in respect to the nanoscale, but it is concerned neither with the critique or further elaboration of these theories, nor with the construction of theories of its own.[4)] This is not to say, however, that closed theories are simply "applied" in nanotechnoscience.

When Heisenberg refers to the hermetically closed character of closed theories (in condition 2 above), he states merely that the theory will be true where its concepts *can* be applied and leaves quite open how big or small the domain of its actual applicability is. Indeed, he suggests that this domain is so small that a "closed theory does not contain any absolutely certain statement about the world of experience" [10]. Even for a closed theory, then, it remains to be determined how and to what extent its concepts can be applied to the world of experience.[5)] Thus, there is no pre-existing domain of phenomena to which a closed theory is applied. Instead, it is "a question of success",

3) Heisenberg's notion of closed theories influenced Thomas Kuhn's conception of a paradigm [11]. It also informed the so-called finalization thesis, one of the first systematic accounts of technoscience [12]. Heisenberg also emphasized a fifth and especially contentious aspect of closed theories: an expansion of their domain of application will not introduce a change to the theory. This aspect and Heisenberg's particular list of closed theories plays no part in the following discussion.

4) In the case of nanotechnoscience, this repertoire includes far more than the four theories singled out by Heisenberg. It is a bold claim, to be sure, that nanotechnoscience is not concerned with the construction of theories of its own. One counterexample might be the discovery and subsequent theoretical work on the giant magnetoresistance effect [13]. Also, there are certain isolated voices who call for the development of theory specifically suited to the complexities of the nanocosm [14, 15]. These voices are isolated, indeed, and the consensus appears to be that the development of nanotechnologies can do without such theories – which might be hard to come by anyway [16].

5) Here, Heisenberg might have been inspired by Heinrich Hertz, who formulated the *Principles of Mechanics* as a closed theory [17]. He defined as mechanical problems all those phenomena of motion that can be accounted for by his fundamental law, albeit with the help of additional assumptions. Phenomena that cannot be accounted for in such a way, are not mechanical problems and simply outside the domain of mechanics (for example, the problems of life).

that is, of calibration, tuning or mutual adjustment to what extent phenomena of experience can be assimilated to the theory such that its concepts can be applied to them.

7.2.2 Applying Theory to the Nanoscale: Fitting Versus Stretching

This notion of "application" has been the topic of many recent discussions on modeling[6] – but it does not capture the case of nanotechnoscience. For in this case, researchers are not trying to bring nanoscale phenomena into the domain of quantum chemistry or fluid dynamics or the like. They are not using models to extend the domain of application of a closed theory or general law. They are not engaged in fitting the theory to reality and *vice versa*. Instead, they take nanoscale phenomena as parts of a highly complex mesocosm between classical and quantum regimes. They have no theories that are especially suited to account for this complexity, no theories, for example, of structure–property relations at the nanoscale.[7] Nanoscale researchers understand, in particular, that the various closed theories have been formulated for far better-behaved phenomena in far more easily controlled laboratory settings. Rather than claim that the complex phenomena of the nanoscale can be described such that the concepts of the closed theory now apply to them, they draw on closed theories eclectically, stretching them beyond their intended scope of application to do some partial explanatory work at the nanoscale.[8] A certain measurement of a current through an organic–inorganic molecular complex, for example, might be reconstructed quantum-chemically or in the classical terms of electrical engineering – and yet, the two accounts do not

6) See the work, in particular, of Nancy Cartwright, Margaret Morrison and Mary Morgan [18, 19].

7) Note that the term "complexity" is used here in a deliberately nontechnical manner. It does not refer to phenomena that fit the constraints of nonlinear complex dynamics, "complexity theory" or the like. The complexity at the nanoscale is one of great messiness, too many relevant variables and properties and multiple complicated interactions. This becomes apparent especially in contrast to the comparatively neat world of the laboratory phenomena that underwrite classical and quantum physics. In its complexity, the "real-world situation" of the nanoworld is precariously situated between classical and quantum regimes.

8) Here is another way to characterize this contrast between "applying" theory by fitting and by stretching: In the standard case of fitting theory to reality and vice versa, the problem concerns ways to compensate for the idealizations or abstractions that are involved in formulating a theory and constructing a model. However, classical theories do not abstract from nanoscale properties and processes, nor do they refer to idealizations of nanoscale phenomena. In this case, the challenge is that of crossing from the intended domain of a classical theory into quite another domain. – like all attempts to distinguish systematically the new nanotechnoscience from old-fashioned "science and engineering," this one is vulnerable to the critique that the two notions of "application" (bringing phenomena into the domain of application, stretching the domain of application to areas for which the theory has not been made) are not categorically distinct but differ only by degree. I thank Eric Winsberg for suggesting this line of thought.

compete against each other for offering a better or best explanation [20]. Armed with theories that are closed relative to the nanoscale, researchers are well equipped to handle phenomena in need of explanation, but they are also aware that they bring crude instruments that are not made specifically for the task and that these instruments therefore have to work in concert. Indeed, nanoscale research is characterized by a tacit consensus according to which the following three propositions hold true simultaneously:

1. There is a fundamental difference between quantum and classical regimes such that classical theories cannot describe quantum phenomena and such that quantum theories are inappropriate for describing classical phenomena.
2. The nanoscale holds intellectual and technical interest because it is an "exotic territory" [14] where classical properties such as color and conductivity emerge when one moves up from quantum levels and where phenomena such as quantized conductance emerge as one moves down to the quantum regime.
3. Nanoscale researchers can eclectically draw upon a large toolkit of theories from the quantum and classical regimes to construct explanations of novel properties, behaviors or processes.

Taken together, these three statements express a characteristic tension concerning nanotechnology, namely that it is thought to be strange, novel and surprising on the one hand, familiar and manageable on the other. More significantly for the present purposes, however, they express an analogous tension regarding available theories: they are thought to be inadequate on the one hand but quite sufficient on the other. The profound difference between classical and quantum regimes highlights what makes the nanocosm special and interesting – but this difference melts down to a matter of expediency and taste when it comes to choosing tools from classical or quantum physics. Put yet another way: what makes nanoscale phenomena scientifically interesting is that they cannot be adequately described from either perspective, but what makes nanotechnologies possible is that the two perspectives make do when it comes to accounting for these phenomena.

Available theories need to be stretched in order to manage the tension between these three propositions. How this stretching actually takes place in research practice needs to be shown with the help of detailed case studies. One might look, for example, at the way in which theory is occasionally "stuck in" to satisfy an extraneous explanatory demand.[9] A more prominent case is the construction of simulation

9) See, for example, a publication in Science on observed effects (large on–off ratios, negative differential resistance) in a molecular device. Asked by peer reviewers to offer an explanation of the observed effect, the authors suggest a somewhat arbitrary but plausible candidate mechanism and call for theories and future experimental work to "elucidate the transport mechanisms" [21]. This discussion was introduced reluctantly since it is clearly unnecessary for the point they wish to make (namely that they can consistently pass a current where no one had done so before) and because it is obviously easy to come up with a sufficiently credible explanation from the toolkit of available theory. The authors implicitly acknowledge that another explanation could easily substitute for theirs.

models where integrations of different levels of theoretical description are tuned to the actual behavior of a nanoscale system or process [22, 23]. This implies also that the very meaning of theories is stretched, especially where they account for the causal structure behind the observed phenomena: as these theories are applied in situations that are taken to be far more complex than the one for which the theories were developed, the causal story offered by them takes a backseat to the contributions they can make towards a description of the phenomena. In other words, algorithms descriptive of a certain dynamics become detached from the causal explanation they originally helped to provide, since it is the initial or structural conditions precisely that are not thought to hold continuously from macro to nano to quantum regimes.[10)]

There is quite another symptom of the ways in which theories and concepts are stretched as they are applied to the nanoscale. The nanoworld is taken to be complex, self-organizing, full of surprises – a world characterized by chemical and biological activity. The aspirations of nanotechnologies therefore emphasize the construction of active rather than merely passive devices.[11)] The so-called first generation of nanotechnological achievements was limited to the generation of new materials (passive structures), the second generation is supposed to incorporate molecular activity into nanotechnical systems.[12)] However, from the point of view of theories that are closed relative to the nanoscale, one cannot "see" any of that novel activity and liveliness but only what has become stabilized in the formulations and formalisms of those theories. Several descriptive or programmatic terms for nanoscale phenomena therefore strain to reach beyond their actual meanings. A prime example of this is the term "selective surface" which attributes agency to something that remains quite passive: Cells may attach to a given surface differentially, but the surface is not therefore doing anything to favor or disfavor certain cells; the selection is entirely on the side of the engineer who selects that surface in order to achieve some functionality. The same holds for

10) At the panel "Ontologies of Technoscience" at the October 2006 Bielefeld Conference ("Science in the Context of Application"), Bernadette Bensaude-Vincent showed that in the development of materials science and nanotechnologies the focus on structure–function relations (Crick's dogma that scientific understanding requires that function is referred to underlying structure) gives way to analyses of dynamic patterns in the observed functions and properties. Davis Baird offered as an example of this a particular nanotechnological detection device that physically instantiates factor analysis and therefore statistically infers underlying causes from observed properties (in other words, it does not perform a physical or chemical analysis to identify the presence of what is measured). Nicole Karafyllis finally suggested that (nano)technologies are now entering into a novel relation to biology as they design function not through construction (e.g. from structural principles) but by way of growth (e.g. by way of harnessing of self-organization).

11) Regarding the prestige of the device *vis-à-vis* the material, see [23a] on Herbert Gleiter as a pioneer of nanotechnology.

12) The notion of first/second generation passive/active devices was established and promoted especially by Mihail Roco of the National Nanotechnology Initiative. This paper is agnostic as to whether the second generation will ever be attained.

"self-cleaning surfaces", "smart materials", "autonomous (self-propelled) movement", the different conceptions of "self-assembly" or "soft machines".[13] All these terms have a specific meaning and at the same time refer to something more visionary, more genuinely "nano" that goes beyond their origin in theories that come from outside the nanoworld.[14]

7.2.3 Mute Complexity

So far, the notion of stretching what we know in one regime to phenomena in another one has been taken descriptively to characterize nanoscale research. Here, however, arises an occasion for critical questioning by scientists, citizens, concerned policy makers. To the extent that one cannot see the specific complexity from the point of view of theories closed relative to the nanoscale, we may find that the difficulties of understanding and controlling nanoscale phenomena are not adequately expressed. By stretching closed theories one recovers partial explanations of phenomena and thereby partial stories only of success. In other words, the assurance that much is amenable to explanation from the large toolkit of available theories finds ample expression, but there is no theoretical framework for the actual struggle of taming and controlling nanoscale phenomena – this part of the story remains untold, locked up in the laboratory.[15] Put bluntly, one might be doing years and years of interesting research only to discover that most of the phenomena one is tinkering with, that one is stabilizing and probing in the laboratory, will never be robust enough to serve as components in nanotechnological devices. There is no language, in other words, to identify specific limits of knowledge and control.

Having arrived at this point in a rather roundabout manner, one might ask whether the limited ways to speak of limits of understanding and control can be shown more straightforwardly. A telling illustration or example is provided by

13) Since the publication of Richard Jones's book on Soft Machines, that concept has been the subject of an emerging discussion [25, 26]. It concerns the question of whether the term "machine" retains any meaning in the notion of a "soft machine" when this is thought of as a non-mechanical, biological machine (while it clearly does retain meaning if thought of as a "concrete" machine in the sense of Simondon).

14) This is especially true, perhaps, for the concept "self-assembly", which has been cautiously delimited, for example, by George Whitesides [27], but which keeps escaping the box and harks backwards and forwards to far more ambitious notions of order out of chaos, spontaneous configurations at higher levels, etc.

15) To be sure, it is a commonplace that laboratory practice is more complex than the stories told in scientific papers. Traditional scientific research often seeks to isolate particular causal relations by shielding them against interferences from the complex macroscopic world of the laboratory. Whether it is easy or difficult to isolate these relations, whether they are stable or evanescent, is of little importance for the scientific stories to be told. The situation changes in respect to nanotechnoscience: its mission is to ground future technologies under conditions of complexity. In this situation, it is more troubling that scientific publications tell stories only of success.

nanotoxicology. It is finding out the hard way that physico-chemical characterization does not go very far, and that even the best methods for evaluating chemical substances do not REACh all the way to the nanoscale [28].[16] In other words, the methods of chemical toxicology go only so far and tell only a small part of the toxicological story – though regarding chemical composition, at least, there are general principles, even laws that can be drawn upon. With regard to the surface characteristics and shape of particles of a certain size, one has to rely mostly on anecdotes from very different contexts, such as the story of asbestos. For lack of better approaches, therefore, one begins from the vantage point of chemical toxicology and confidently stretches available theories and methods as far as they will go – while the complexities of hazard identification, let alone risk assessment (one partially characterized nanoparticle or nanosystem at a time?!) tend to be muted.[17]

There is yet another, again more general, way to make this point. Theories that are closed relative to the nanoscale can only introduce nonspecific constraints. The prospects and aspirations of nanotechnologies are only negatively defined: Everything is thought to be possible at the nanoscale that is not ruled out by those closed theories or the known laws of nature. This, however, forces upon us a notion of technical possibility that is hardly more substantial than that of logical possibility. Clearly, the mere fact that something does not contradict known laws is not sufficient to establish that it can be realized technically under the complex conditions of the nanoregime. Yet once again, there is no theoretical framework or language available to make a distinction here and to acknowledge the specificities and difficulties of the nanoworld – since all we have are theories that were developed elsewhere and that are now stretched to accommodate phenomena from the nanosphere.[18] However, failure to develop an understanding also of limits of understanding and control at the nanoscale has tremendous cost as it misdirects expectations, public debate and possibly also research funding.

16) The pun is intended: REACh refers to the new style of regulating chemical substances in and by the EU. It is widely acknowledged that it does not apply where properties depend not only on chemical composition but also on surface characteristics, size, shape, perhaps also engineered functionality and the specificities of their environments. (Along similar lines, Joachim Schummer [29] has argued that REACh does not even reach the products of conventional synthetic chemistry.)

17) Sabine Maasen and Monika Kurath have shown that this difficulty for chemical toxicology creates interesting new opportunities for nanotoxicology [31]. For another illustration of the predicament, one might recall that carbon nanotubes were "discovered" in the 1980s, that for a good number of years they have been being commercially manufactured and that researchers are still complaining that no two batches are alike.

18) I have been urging that more attention should be paid to limits of understanding and control at the nanoscale. If I am right in this section, I have been asking for something that cannot be done (as of now) in a straightforward way.

7.3
From Successful Methods to the Power of Images

7.3.1
(Techno)scientific Methodology: Quantitative Versus Qualitative

As was shown above, Heisenberg considered "a question of success" the extent to which phenomena of experience can be fitted to closed theories [10]. This suggests the question what "success" amounts to in nanoscale research, that is, what it takes to satisfy oneself that one has reached a sufficiently good understanding or control of the phenomena under investigation.

For Heisenberg and any philosopher of science who is oriented towards theoretical physics, this question boils down to the predictive success of a quantitative science. Here, "quantitative" means more than the employment of numbers and even of precision measurements. The characteristics of quantitative approaches include the following. First, predicted numerical values are compared with values obtained by measurement. The reasonably close agreement between two numbers thus serves to establish the agreement of theory and reality. Second, this quantitative agreement emphatically makes do without any appeal to a likeness or similarity between theoretical models and the real-world systems they are said to represent. Quantitative science rests content if it reliably leads from initial conditions to accurate predictions, it does not require that all the details of its conceptual apparatus (every term in its algorithms) has a counterpart in reality. Both characteristics of quantitative science are familiar especially from twentieth century theoretical physics – but do they serve to characterize also nanotechnoscience [31]?

In the light of the extremely heterogeneous research practices under the general heading of "nanoscience and nanotechnologies" there may not be a general answer to this question. Yet it is fair to say that much nanotechnoscientific research is qualitative. Its epistemic success consists in constructions of likeness.[19)]

The shift sounds innocent enough but may have significant consequences: the agreement of predicted and measured quantities is being displaced by an agreement of calculated and experimental images. The latter qualitative agreement consists

19) Here, a case study of Jan Hendrik Schön might show that he was caught between quantitative and qualitative methodologies. He was "caught cheating", after all, when it was discovered that for different experiments he included an exactly identical plot of current flow. This diagram is supposed to be generated from a series of measurements but the characteristic shape of the curve is also a qualitative short-hand expression for "current is flowing." In a culture of research that is moving increasingly to produce effects, Schön may well have "written" this diagram as it is generally "read" – without regard to the particular values but as a symbol for a certain type of event. Overall, Schön's case is less innocent and more complicated than this [32]. But perhaps in other regards, too, it is symptomatic of the ambivalence that results from the transdisciplinary qualitative orientation of nanotechnoscience even as nanoscale research continues to be informed mostly by rigorously quantitative disciplinary traditions.

primarily in the absence, even deliberate suppression of visual clues by which to hold calculated and experimental images apart. Indeed, the (nano)technoscientific researcher frequently compares two displays or computer screens. One display offers a visual interpretation of the data that were obtained through a series of measurements (e.g. by an electron or scanning probe microscope), the other presents a dynamic simulation of the process he might have been observing – and for this simulation to be readable as such, the simulation software produces a visual output that looks like the output for an electron or scanning probe microscope. Agreement and disagreement between the two images then allows the researchers to draw inferences about probable causal processes and to what extent they have understood them. Here, the likeness of the images appears to warrant the inference from the mechanism modeled in the simulation to the mechanism that is probably responsible for the data that were obtained experimentally. Accordingly (and this cannot be done here), one would need to show how nanoscale researchers construct mutually reinforcing likenesses, how they calibrate not only simulations to observations and visual representations to physical systems but also their own work to that of others, current findings to long-term visions. This kind of study would show that unifying theories play little role in this, unless the common availability of a large tool-kit of theories can be said to unify the research community. Instead of theories, it is instruments (STM, AFM, etc.), their associated software, techniques and exemplary artefacts (buckyballs, carbon nanotubes, gold nanoshells, molecular wires) that provide relevant common referents [33–35].

7.3.2 "Ontological Indifference": Representation Versus Substitution

This is also not the place to subject this qualitative methodology to a sustained critique. Such a critique is easy, in fact, from the point of view of rigorous and methodologically self-aware quantitative science [31]. Far more interesting is the question of why, despite this critique, a qualitative approach appears to be good enough for the purposes of nanoscale research. As Peter Galison has pointed out, these purposes are not to represent the nanoscale accurately and, in particular, not to decide what exists and what does not exist, what is more fundamental and what is derivative. He refers to this as the "ontological indifference" of nanotechnoscience [5]. Why is it, then, that nanotechnological research can afford this indifference? For example, molecular electronics researchers may invoke more or less simplistic pictures of electron transport but they do not need to establish the existence of electrons. Indeed, electrons are so familiar to them that they might think of them as ordinary macroscopic things that pass through a molecule as if it were another material thing with a tunnel going through it [20]. Some physicists and most philosophers of physics strongly object to such blatant disregard for the strangely immaterial and probabilistic character of the quantum world that is the home of electrons, orbitals, standing electron waves [36, 37]. And indeed, to achieve a practical understanding of electron transport, it may be necessary to entertain more subtle accounts. However, it is the privilege of ontologically indifferent technoscience that it

can always develop more complicated accounts as the need arises. For the time being, it can see how far it gets with rather more simplistic pictures.[20)]

Ontological indifference amounts to a disinterest in questions of representation and an interest, instead, in substitution.[21)] Instead of using sparse modeling tools to represent only the salient causal features of real systems, nanoresearchers produce in the laboratory and in their models a rich, indeed oversaturated substitute reality such that they begin by applying alternative techniques of data reduction not to "nature out there" but to some domesticated chunk of reality in the laboratory. These data reduction and modeling techniques, in turn, are informed by algorithms which are concentrated forms of previously studied real systems, they are tried and true components of substitute realities that manage to emulate real physical systems [38].[22)] In other words, there is so much reality in the simulations or constructed experimental systems before them, that nanotechnology researchers can take them for reality itself [39]. They study these substitute systems and, of course, have with these systems faint prototypes for technical devices or applications. While the public is still awaiting significant nanotechnological products to come out of the laboratories, the researchers in the laboratories are already using nanotechnological tools to detach and manipulate more or less self-sufficient nanotechnological systems which "only" require further development before they

20) A particularly interesting and challenging example of this is Don Eigler's famous picture of a quantum corral that confines a standing electron wave. The picture's seemingly photographic realism suggests that the quantum corral is just as thing-like as a macroscopic pond. It brazenly bypasses all discussions regarding the interpretation of quantum mechanics and thus displays its ontological indifference. Nevertheless, it is an icon of nanotechnoscience, testimony to new capabilities of manipulation and visualization and a down-payment of sorts on the promise that technical control does not stop at the threshold to quantum effects.

21) Compare Peter Galison's suggestion above that the relevant contrast is that between demonstrating existence and building things. Yet, as will be shown in Section 7.5 below, "building" is too narrow and too "technical" a notion. It does not do justice to the intellectual engagement, even passion for the challenges encountered at the nanoscale.

22) Rom Harré contrasts scientific instruments that serve as probes into causal processes and modeling apparatus (including simulations) that domesticates or produces phenomena. It is this modeling apparatus that underwrites epistemic success in constructions of likeness: Instruments typically obtain measurements that can be traced back down a causal chain to some physical state, property or process. As such, the instruments are detached from nature – measurements tell us something about the world. Physical models, in contrast, are part of nature and exhibit phenomena such that the relevant causal relations obtain within the apparatus and the larger apparatus–world complex. Whether it domesticates a known phenomenon such as the rainbow or elicits an entity or process that does not occur "naturally", it does not allow for straightforward causal inference to the world within which the apparatus is nested [38]. As the metaphor of domestication and Harré's conception of an apparatus–world complex suggest, causal inference from the apparatus to the world may be required only for special theoretical purposes that are characterized by a specific concern for reality (for example, when something goes wrong and one wants to explore the reasons for this). At the same time, the very fact that the apparatus is nested in the world delivers an (unarticulated) continuity of principles and powers and the affordance of ontological indifference.

can exist as useful devices outside the laboratory, devices that not only substitute for but improve upon something in nature.

7.3.3 Images as the Beginning and End of Nanotechnologies

Again, it may have appeared like a cumbersome path that led from qualitative methodology and its constructions of likeness to the notion that models of nanoscale phenomena do not represent but substitute chunks of reality and that they thereby involve the kind of constructive work that is required also for the development of nanotechnological systems and devices. For a more immediate illustration of this point, we need to consider only the role of visualization technologies in the history of nanotechnological research.[23)] Many would maintain, after all, that it all began for real when Don Eigler and Erhard Schweizer created an image with the help of 35 xenon atoms. By arranging the atoms to spell "IBM" they did not represent a given reality but created an image that replaces a random array of atoms by a technically ordered proto-nanosystem. Since then, the ability to create images and to spell words has served as a vanguard in attempts to assert technical control in the nano-regime – the progress of nanotechnological research cannot be dissociated from the development of imaging techniques that are often at the same time techniques for intervention. Indeed, Eigler and Schweizer's image has been considered proof of concept for moving atoms at will. It is on exhibit in the STM web gallery of IBMs Almaden laboratory and is there appropriately entitled "The Beginning" – a beginning that anticipates the end or final purpose of nanotechnologies, namely to directly and arbitrarily inscribe human intentions on the atomic or molecular scale.

Images from the nanocosm are at this point (early 2008) still the most impressive as well as popular nanotechnological products. By shifting from quantitative coordinations of numerical values to the construction of qualitative likeness, from the conventional representation of reality to the symbolic substitution of one reality by another, nanotechnoscience has become beholden to the power of images. Art historians and theorists like William Mitchell and Hans Belting, in particular, have emphasized the difference between conventional signs that serve the purpose of representation and pictures or images that embody visions and desires, that cannot be controlled in that they are not mere vehicles of information but produce an excess of meaning that is not contained in a conventional message [40, 41].

The power of images poses some of the most serious problems of and for nanoscience and nanotechnologies. This is readily apparent already for "The Beginning". As mentioned above, it is taken to signify that for the first time in history humans have manipulated atoms at will and thus as proof of concept for the most daring nanotechnological visions and by the most controversial nanotechnological visionaries such as Eric Drexler. This was not, of course, what Eigler and

23) It is no accident that this is perhaps the best-studied and most deeply explored aspect of nanotechnologies.

Schweizer wanted to say. Their image is testimony also to the difficulty, perhaps the limits of control of individual atoms. But the power of their image overwhelms any such testimony.

Here arises a problem similar to the one encountered in Section 7.2.3. The specificity, complexity and difficulty of work at the nanoscale do not have a language and do not find expression. The theories imported from other size regimes can only carve out an unbounded space of unlimited potential, novelty, possibility. And the pictures from the nanocosm show us a world that has already been accommodated to our visual expectations and technical practice.[24] Ontologically indifferent, nanotechnoscience may work with simplistic conceptions of electron transport and it produces simplistic pictures of atoms, molecules, standing electron waves which contradict textbook knowledge of these things. For example, it is commonly maintained that nanosized things consist only of surface and have no bulk. This is what makes them intellectually and technically interesting. But pictures of the nanocosm invariably show objects with very familiar bulk-surface proportions, a world that looks perfectly suited for conventional technical constructions. And thus, again, we might be facing the predicament of not being told or shown what the limits of nanotechnical constructions and control might be.

The power of images also holds another problem, however. In the opposition of conventional sign and embodied image the totemistic, fetishistic, magical character of pictures comes to the fore. To the extent that the image invokes a presence and substitutes for an absence, its kinship to voodoo-dolls, for example, becomes apparent. This is not the place to explore the analogy between simulations and voodoo-dolls [31], but it should be pointed out that nanotechnologies in a variety of ways cultivate a magical relation to technology – and their imagery reinforces this. Indeed, in the history of humankind we might have begun with an enchanted and uncanny nature that needed to be soothed with prayer to the spirits that dwelled in rocks and trees. Science and technology began as we wondered at nature, became aware of our limits of understanding and yet tamed and rationalized nature in a piece-meal fashion. Technology represents the extent to which we managed to defeat a spirited, enchanted world and subjected it to our control. We technologized nature. Now, however, visitors to science museums are invited to marvel at nanotechnologies, to imagine technological agency well beyond human thresholds of perception, experience and imagination and to pin societal hopes for technological innovation not on intellectual understanding but on a substitutive emulation that harnesses the self-organizing powers of nature. We thus naturalize technology, replace rational control over brute environments by a magical dependency on smart environments and we may end up rendering technology just as uncanny as nature used to be with its earthquakes, diseases and thunderstorms [42, 43].[25]

24) Compare footnote 20 above.

25) This is a strong indictment not of particular nanotechnologies but of certain ways of propagating our nanotechnological future. Considered another way, it is simply an engineering challenge to design nanotechnology for the human scale.

7.4
From Definitions to Visions

7.4.1
Wieldy and Unwieldy Conceptions

The first two sections gave rise to the same complaint. After surveying the role of theories and methodologies for the construction of technical systems that can substitute for reality, it was noted that this tells us nothing about the specificity, complexity and difficulty of control at the nanoscale. The nanocosm appears merely as that place from where nanotechnological innovations emanate and so far it appears that it can be described only in vaguely promising terms: the domain of interest to nanoscience and nanotechnologies is an exotic territory that comprises all that lies in the borderland of quantum and classical regimes, all that is unpredictable (but explicable) by available theories and all that is scale-dependently discontinuous, complex, full of novelty and surprise.[26)]

However, as one attempts a positive definition of nanotechnoscience and its domain of phenomena or applications, one quickly learns how much is at stake. In particular, definitions of "nanotechnology" suggest the unity of a program so heterogeneous and diverse that we cannot intellectually handle or manage the concept any more. By systematically overtaxing the understanding, such definitions leave a credulous public and policy makers in awe and unable to engage with "nanotechnology" in a meaningful manner. The search for a conceptually manageable definition is thus guided by an interest in specificity but also by a political value – it is to facilitate informed engagement on clearly delimited issues. In purely public contexts, therefore, it is best not to speak of nanotechnology in the singular at all but only of specific nanotechnologies or nanotechnological research programs [44]. In the present context, however, an effort is made to circumscribe the scope or domain of nanotechnoscience, that is, to consider the range of phenomena that are encountered by nanoscience and nanotechnologies. This proves to be a formidable challenge.

7.4.2
Unlimited Potential

There is an easy way to turn the negative description of the domain into a positive one. One might say that nanoscience and nanotechnologies are concerned with everything molecular or, slightly more precisely, with the investigation and manipulation of molecular architecture (as well as the properties or functionalities that depend on molecular architecture).

26) Tellingly, the most sophisticated definition of nanoscience is quite deliberate in saying nothing about the "nanocosm" at all. Indeed, this definition is not limited to nanoscale phenomena or effects but intends a more general nanoscience of scale-dependently discontinuous behaviors at all scales: nanoscience is everywhere where one encounters a specific kind of novelty or surprise [6].

Everything that consists of atoms is thus an object of study and a possible design target of nanoscale research. This posits a homogeneous and unbounded space of possibility, giving rise, for example, to the notion of an all-powerful nanotechnology as a combinatorial exercise that produces the "little BANG" [45] – since bits, atoms, neurons, genes all consist of atoms, since all of them are molecular, they all look alike to nanoscientists and engineers who can recombine them at will. And thus comes with the notion of an unlimited space of combinatorial possibilities the transgressive character of nanotechnoscience: categorial distinctions of living and inanimate, organic and inorganic, biological and technical things, of nature and culture appear to become meaningless. Although hardly any scientist believes literally in the infinite plasticity of everything molecular, the molecular point of view proves transgressive in many nanotechnological research programs. It is particularly apparent where biological cells are redescribed as factories with molecular nanomachinery. Aside from challenging cultural sensibilities and systematic attempts to capture the special character of living beings and processes, nanotechnoscience here appears naively reductionist. In particular, it appears to claim that context holds no sway or, in other words, that there is no top–down causation such that properties and functionalities of the physical environment partially determine the properties and behaviors of the component molecules.[27)]

This sparsely positive and therefore unbounded view of nanoscale objects and their combinatorial possibilities thus fuels also the notion of unlimited technical potential along with visions of a nanotechnological transgression of traditional boundaries. Accordingly, this conception of the domain of nanoscience and nanotechnologies suffers from the problem of unwieldiness – it can play no role in political discourse other than to appeal to very general predispositions of technophobes and technophiles [46].

Three further problems, at least, come with the conception of the domain as "everything molecular out there." And as before, internally scientific problems are intertwined with matters of public concern. There is first the (by now familiar) "scientific" and "societal" problem that there is no cognizance of limits of understanding and control – as evidenced by a seemingly naive reductionism. There is second the (by now also familiar) problem that technoscientific achievements and conceptions have a surplus of meaning which far exceeds what the research community can take responsibility for – the power of images is dwarfed by the power of visions (positive or negative) that come with the notion of unlimited potential. And there is finally the problem of the relation of technology and nature.

27) I cannot pass judgement on these claims. However, even Richard Jones's Soft Machines [32] with its vivid appreciation of the complexities of "biological nanotechnology" does not reflect the findings of developmental biologists regarding environmental stimuli to gene expression. Recent work on adult stem cells appears to reveal that they can be reverted to earlier states but that they nevertheless "remember" what they were. Such findings complicate immensely the apparently unbounded promise that nanotechnology can solve all problems at the level of molecules.

Martin Heidegger, one of the sharpest critics of modern technology, chastised it for treating all of nature as a mere resource that is "standing in reserve" to be harnessed by science and industry [47]. The power of his argument derives precisely from the fact that he saw all of modern technoscience as one: it is a scaffolding or harness (the German word is *Gestell*) that recruits humans and nature into a universal scheme of production. Rather than accept as a gift what nature, poetry or craft brings forth, it demands the deliverance of what it has learned rationally to expect from the study of nature as a calculable system of forces. Conceived as a unified enterprise with an unbounded domain of "everything molecular", nanotechnology fits the bill of such an all-encompassing modern technology. It does so because it employs what one might call a thin conception of nature. According to this conception, nature is circumscribed by the physical laws of nature. All that accords with these laws is natural. Thus, nanotechnology can quickly and easily claim for itself that it always emulates nature, that it manufactures things nanotechnologically just as nature does when it creates living organisms. This conception, however, is too "thin" or superficial to be credible and it suffers from the defect that the conditions of (human) life on earth have no particular valence in it: from the point of view of physics and the eternal laws of nature, life on Earth is contingent and not at all necessary. The laws predate and will outlive the human species. In contrast, a substantial, richly detailed or "thick" conception of nature takes as a norm the special evolved conditions that sustain life on Earth. Here, any biomimetic research that emulates nature will be characterized by care and respect as it seeks to maintain these special conditions. This involves an appreciation of how these conditions have evolved historically. On this conception, context holds sway and a molecule that occurs in a technical system will not be the same as one in a biological system, even if it had the same chemical composition.

It is an open question and challenge to nanoscience and nanotechnologies, however, whether it can embrace such a thick or substantial conception of nature.

7.4.3
A Formidable Challenge

It was not very difficult to identify at least four major problems with the commonly held view that the domain of nanotechnological research encompasses "everything molecular" It proves quite difficult, in contrast, to avoid those problems. In particular, it appears to defy common sense and the insights of the physical sciences to argue that molecules should have a history or that they should be characterized by the specific environments in which they appear. Is it not the very accomplishment of physical chemistry ever since Lavoisier that it divested substances of their local origins by considering them only in terms of their composition, in terms of analysis and synthesis? [48]. And should one not view nanoscience and nanotechnologies as an extension of traditional physics, physical chemistry and molecular biology as they tackle new levels

of complexity? All this appears evident enough, but yet there are grounds on which to tackle the formidable challenge and to differentiate the domain of nanoscientific objects.[28)]

As noted above, bulk chemical substances are registered and assessed on the grounds of a physico–chemical characterization. Once a substance has been approved, it can be used in a variety of contexts of production and consumption. On this traditional model, there appears no need to consider its variability of interactions in different biochemical environments (but see [29, 30]). Although the toolkit of nanotoxicology is still being developed, there is a movement afoot according to which a carbon nanotube is perhaps not a carbon nanotube. What it is depends on its specific context of use: dispersed in water or bound in a surface, coated or uncoated, functionalized or not – all this is toxicologically relevant. Moreover, a comprehensive physico-chemical characterization that includes surface properties, size and shape would require a highly complex taxonomy with too many species of nanoparticles, creating absurdly unmanageable tasks of identification perhaps one particle at a time. Instead, the characterization of nanoparticles might proceed by way of the level of standardization that is actually reached in production and that is required for integration in a particular product – with a smaller or larger degree of variability, error tolerance, sensitivity to environmental conditions, as the case may be for a specific product in its context of use. Nanotoxicology would thus be concerned with product safety rather than the safety of component substances. On this account, the particles would indeed be defined by their history and situation in the world and thus thickly by their place within and their impact upon nature as the specific evolved conditions of human life on Earth.[29)]

There is another, more principled, argument for a thickly differentiated account of the objects that make up the domain of nanoscience and nanotechnologies. The unbounded domain of "everything molecular" includes not only the objects and properties that we now have access to and that we can now measure and control. It also includes those objects and properties that one may gain access to in the future. This way of thinking is indifferent to the problem of actual technical access also in that it does not consider how observational instruments and techniques structure, shape and perhaps alter the objects in the domain. On this account, the domain appears open and unlimited because

28) One might argue that a definition requiring scale-dependent discontinuities already does offer such a differentiation [6]. This is not the case, however. It is a beginning at best. As shown above, this definition excludes certain phenomena and processes from nanoscience and thus claims specificity, but it leaves nanotechnology entirely undetermined: nanoscience tends to certain novel or surprising properties and processes, nanotechnology is whatever one can make of these properties and processes. More significantly, however, the appearance of scale-dependent novel properties can be claimed rather generically. Not every property at the nanoscale is discontinuous in respect of scale. However, for every substance one can claim that it may or will have some such properties simply by virtue of the proportion of atoms in the boundary layers.

29) Nanotoxicology in particular, and nanotechnological research in general, might thus become a "social science of nature" [49].

it refers to an imaginary (future) state of total information and the nonintrusive and presence of observers in the nanoworld. In contrast to this account, the domain could be delimited more concretely and its visionary surplus could be contained more effectively if it did not include all nanoscale objects "out there" but considers how these objects are constituted, how they become accessible to nanoscale research. Accordingly, the domain of objects and processes would consist of just those phenomena and effects that are revealed by scanning tunneling microscopy and other specifically nanotechnological procedures [50, 51].

However, more so than the current attempt to formulate a philosophy of nanotechnoscience, this proposal by Peter Janich ascribes to nanotechnoscience a methodological unity or basis in common practice. He suggests a philosophical program of systematizing the operations by which nanoscale objects become amenable to measurement and observation. Such a systematic reconstruction of the domain of objects of nanotechnological research might begin by looking at length measurement or scanning probe microscopy. However, research practice is not actually unified in this manner. Even scanning probe microscopy – to many a hallmark or point of origin for nanotechnologies – plays a minor role in the work of many nanoscale researchers [52]. Also, the above-mentioned struggles to attain standard measures or to characterize nanomaterials testify to the unruliness of the objects of research. They are not constituted through methodical procedures that individuate objects and make them comparable throughout the scientific community. Instead, it appears that they are constituted through complicated interactions that are difficult to reproduce and that rely on proximate likeness.

Since Janich's approach faces considerable odds, all one can do perhaps is to generalize the previous lesson from nanotoxicology: The objects of nanoscale research are constituted through their specific histories – histories that concern their origin (in a tissue sample, in the soil, in a chemically produced batch), that include nanotechnological interventions as well as their location finally in a technical system. This would promote, of course, the fragmentation of "nanotechnology" into as many "nanotechnologies" as there are nanotechnological devices or applications. A nightmare vision for some, others may consider this an intellectual requirement. If this is so, it becomes impossible to uphold the idea of carbon nanotubes as all-purpose technical components. If they contribute to the performance of some product, then they are individuated or characterized as being carbon-nanotubes-in-that-product and they are as safe or unsafe as that product is. By the same token, they are no longer conceived as molecular objects that are combinable in principle with just about any other. The open space of unlimited potential differentiates into a manifold of specific technological trajectories.

The formidable challenge has not been met by this proposal. It does help dramatize, however, the inherent tension in the commonly held view of nanotechnological objects, as well as the difficulties (once again) of prediction and control at the nanoscale.

7.5
From Epistemic Certainty to Systemic Robustness

7.5.1
What Do Nanoscientists Know?

The previous sections considered research practices of nanotechnosciences – how theories are stretched to the complexities at the nanoscale, how a qualitative methodology serves the construction of likeness and inferences from that likeness, how the research objects are individuated and encountered. All these practices contribute to the generation of knowledge but it remains to be explored in which sense this is "objective knowledge." As in traditional science, the findings of nanotechnoscientific research are published in scientific journals, so the questions is, more concretely, what kind of knowledge is expressed or communicated in a nanoscientific journal article? To answer this question properly, contrasts need to be established and particular publications compared. Here, a summary must suffice.

A typical research article in classical science states a hypothesis, offers an account of the methods, looks at the evidence produced and assesses the hypothesis in the light of the evidence. It participates in a public process of evaluating propositions, of finding certain statements true or false and of seeking certainty even where it is impossible to attain. In contrast, a technoscientific research article provides testimony to an acquired capability. It offers a sign or proof of what has been accomplished in the laboratory and tells a story of what has been done. The telling of the story does not actually teach the capability but it offers a challenge to the reader that they might develop this capability themselves. As opposed to epistemic knowledge (concerned with truth or falsity of propositions), nanoscale research produces skill knowledge. This is not an individualized skill, however, or tacit knowledge. Acquired capabilities can be objective and public, specifically scientific and communicable. They grasp causal relations and establish habits of action. They are assessed or validated not by the application of criteria or norms but by being properly entrenched in a culture of practice. One cannot judge their truth or falsity (skills are not true or false) but one can judge the robustness of demonstrability: If one has acquired a capability, one can more or less consistently do something in the context of an "apparatus–world complex" [38]. As opposed to the truth or falsity, certainty or uncertainty of hypotheses, the hallmarks of technoscientific knowledge are robustness, reliability, resilience of technical systems or systematic action.

7.5.2
The Knowledge Society

This account of skill knowledge presses the question of where the "science" is in "technoscience". The answer to this question can be found in the first section of this chapter: it is in the (closed) theories that are brought as tools to the achievement of partial control and partial understanding. Nanotechnoscience seeks not to improve theory or to change our understanding of the world but primarily to manage

complexity and novelty. As such, nanotechnoscience is just technical tinkering, just product development, just an attempt to design solutions to societal problems or to shape and reshape the world. However, the conceptual and physical tools it tinkers with do not come from ordinary experience, from common sense and a craft tradition but concentrate within them the labors of science. So, the "science" of "nanotechnoscience" is what goes into it. What comes out is skill knowledge and this knowledge does not rely on a corresponding scientific understanding. As long as one can produce an effect in a reasonably robust manner, it does not really matter whether scientific understanding catches up. Indeed, the complexities may be such that it cannot fully catch up.[30)]

The standard example of technology being ahead of science is the steam engine, which was developed without a proper understanding of the relation between heat and work [53]. This understanding came much later and, indeed, was prompted by the efficient performance of the steam engine. The steam engine itself was therefore not applied science but the result of technical tinkering. It was made of valves, pumps, gears, and so on, of which there was good nonscientific craft-knowledge – and it worked just fine before the advent of thermodynamics. In a sense, it did not need to be understood.

As opposed to the steam engine, nanotechnological devices (whatever they may be), genetically modified organisms, drug delivery systems are offsprings of the knowledge society. They are not made of valves and pumps but assembled from highly "scientized" components such as algorithms, capabilities acquired by scientifically trained researchers who are using measuring and monitoring devices that have plenty of knowledge built in [39]. The science that goes into the components is well understood, not so the interactions of all the components and their sensitivities in the context of the overall technical system. Still, like the steam-engine, it may work just fine without being fully understood. And although one cannot attain positive knowledge from which to derive or predict its performance, we may learn to assess its robustness.

7.5.3 Social Robustness

The shift from hypotheses that take the form of sentences to actions within technocultural systems, from epistemic questions of certainty to systemic probes of robustness has implications also for the "risk society" that looks to government mostly for protection from risk [54].[31)]

30) This diagnosis is not entirely novel or surprising. Technology, writes Heidegger, is always ahead of science and, in a deep sense, science is only applied technology [47]. By this he means not only that laboratory science requires instruments and experimental apparatus for stabilizing the phenomena. He means more generally that a technological attitude informs the scientific way of summoning phenomena to predictably appear once certain initial conditions are met.

31) The precautionary principle refers to the certainty and uncertainty of knowledge regarding risks. Where technology assessment shifts from truth of sentences about risk to the robustness or resilience of emerging technical systems and their interaction with other technical systems. In this case, a different kind of prudential approach is required – for example, Dupuy and Grinbaum's "ongoing normative assessment" [55].

Expectations of certainty and assurances of safety will not be met by nanotechnologies. Other technologies already fail to meet them. Certainty about the safety of a new drug, for example, is produced by the traditional method of a clinical trial that establishes or refutes some proposition about the drug's efficacy and severity of side-effects. A far more complex and integrated mechanism is required where such certainty is unattainable and where robustness needs to be demonstrated. Here, several activities have to work in tandem, ranging from traditional toxicology, occupational health and epidemiology all the way to the deliberate adoption of an unknown risk for the sake of a significant desired benefit. If this integration works, social robustness will be built into the technical system along with the robustness of acquired skills, tried and true algorithms, measuring and monitoring apparatus. The fact that nanoscale researchers demonstrate acquired capabilities and that they thus produce "mere" skill knowledge creates a demand for skill knowledge also in a social arena where nanotechnological innovations are challenged, justified and appropriated.

7.6 What Basic Science Does Nanotechnology Need?

The preceding sections provided a survey of nanotechnoscience in terms of disciplinary questions (a complex field partially disclosed by stretching closed theories), of methodology (constructions and qualitative judgments of likeness), of ontology (a thin conception of nature as unlimited potential) and of epistemology (acquisition and demonstration of capabilities). This does not exhaust a philosophical characterization of the field which would have to include, for example, a sustained investigation of nanotechnology as a conquest of space or a kind of territorial expansion.[32] Also, nothing has been said so far about nanotechnology as an enabling technology that might enable, in particular, a convergence with bio- and information technologies. Finally, it might be important to consider nanotechnoscience as an element or symptom of a larger cultural transition from scientific to technoscientific research.

This survey is limited in other ways. It glossed over the heterogeneity of research questions and research traditions. And it focused exclusively on the way in which nanotechnological research has developed thus far. There is nothing in the preceding account to preclude a profound reorientation of nanoscience and nanotechnologies. Indeed, one reorientation might consist in the whole enterprise breaking apart and continuing in rather more traditional disciplinary settings – with "nano" ceasing to be a funding umbrella but becoming a prefix that designates a certain approach. Thus, under the sectoral funding umbrellas "food and agriculture", "energy", "health", "manufacturing" or "environment" researchers with the "nano" prefix would investigate how problems and solutions can be viewed at the molecular level. Their work

32) One implication of this is that nanotechnology should not be judged as the promise of a future but, instead, as a collective experiment in and with the present [56].

would then have to be integrated into more comprehensive approaches to the problem at hand.

Alternatively, nanotechnological researchers may pursue and promote disciplinary consolidation and unification.[33] In that case, they might be asking the question, "what kind of basic science does nanotechnology need?". From quantum mechanics, hydrodynamics, and so on, derive the (closed) theories that serve as the toolkit on which nanoscale research is drawing. While these are basic sciences, of course, they are not therefore the basis of nanoscience. What, then, is the basic scientific research that needs to be done in order to ground nanotechnologies properly or to establish nanoscience as a field in its own right? There have been no attempts so far to address this question in a systematic way.[34] And obviously, one should not expect any consensus regarding the following list of proposed basic research for nanotechnology.

In terms of empirical grounding or a theoretical paradigm, some call for general theories of (supra-)molecular structure–property relations, others imagine that there will be a future science of molecular and nanotechnical self-organization.[35] Following the suggestion of Peter Janich (see above, Section 7.4.3), one might identify and systematize how nanoscale phenomena are constituted through techniques of observation and measurement – this might render theories of instrumentation basic to nanoscience.[36]

Another kind of basic research entirely would come from so-called *Bildwissenschaft* (image or picture science) that could provide a foundation for image-production and visualization practice in nanotechnoscience. Such investigations might contribute visual clues for distinguishing illustrations from animations, from simulations, from visualizations of microscopically obtained data. They might also turn to image–text relations or develop conventions for reducing the photographic intimations of realism while enhancing informational content.[37]

33) The field of "nanomedicine" appears to be moving in that direction by distinguishing its research questions and paradigms from "medical nanotechnologies." It is not at all clear yet whether nanomedicine will emerge from this with a disciplinary identity of its own, including perhaps a unique body of theory.

34) To be sure, there are piecemeal approaches. One might say, for example, that a theory of electron transport is emerging as a necessary prerequisite for molecular electronics (but see [57]). Also, the giant magnetoresistance effect might be considered a novel nanotechnological phenomenon that prompted "basic" theory development [13].

35) See, for example, [15]. In [14] Michael Roukes calls for the identification of the special laws that govern the nanoscale. To be sure, there is profound skepticism in the scientific community (a) that there can be laws of structure-property relations at the nanoscale and (b) that they are needed in order to pursue nanotechnological research. On this latter view, the account provided in the first four sections of this paper provides sufficient "grounding" of nanotechnology.

36) See, for example, [58] on modeling of measurements at the nanoscale. Can this kind of theory development and modification serve to constitute a nanoscale research community – or does it belong to a special tribe of instrument developers that merely enters into a trade with other nanotechnology researchers? [59].

37) Compare the suggestion by Thomas Staley (at the conference on Imaging Nanospace) that visualizations of data could be constructed like maps with graphic elements even text imposed upon the quasi-photographic image [60]. This might break the spell of the powerful image (see above, Section 7.3.3) and return ownership of the image to the scientific community.

Finally, one might ask whether nanotechnoscience can and should be construed as a "social science of nature".[38] As an enabling, general-purpose or key technology it leaves undetermined what kinds of applications will be enabled by it. This sets it apart from cancer research, the Manhattan project, the arms race, space exploration, artificial intelligence research, and so on. As long as nanotechnoscience has no societal mandate other than to promote innovation, broadly conceived, it remains essentially incomplete, requiring social imagination and public policy to create an intelligent demand for the capabilities it can supply. As research is organized to converge upon particular societal goals [61], nanoscience and nanotechnology might be completed by incorporating social scientists, anthropologists and philosophers in its ambitions to design or shape a world atom by atom.

Nanotechnologies are frequently touted for their transformative potential, for bringing about the next scientific or industrial revolution. This chapter did not survey a revolutionary development, but pragmatic and problematic integrations of pre-existing scientific knowledge with the novel discoveries at the nanoscale. If one expects science to be critical of received theories and to produce a better understanding of the world, if one expects technology to enhance transparency and control by disenchanting and rationalizing nature, these pragmatic integrations appear regressive rather than revolutionary. If one abandous these expectations and makes the shift from epistemic certainty to systemic robustness, these pragmatic integrations hold the promise of producing socially robust technologies. In the meantime, there is no incentive for researchers and hardly any movement on the side of institutions to consider seriously the question of a disciplinary reorientation and consolidation of the nanosciences and nanotechnologies. A nanotechnological revolution has not happened yet: we may be waiting for it in vain and this is probably a good thing.[39]

38) See footnotes 3 and 29. The term Soziale Naturwissenschaft was coined in the context of the finalization thesis and could be designated more literally as a social natural science – science of a nature that is socially shaped through applied science, technology and human action. It is thus not social science but an integrated approach that acknowledges the social character of the world. Here, this proposal is taken up in two ways. Materials (as opposed to matter) and molecules defined by their history and situation are social entities, as such objects of this socials science of nature. Second, nanotechnoscience is a program for shaping and reshaping, for designing and redesigning, for reforming the world. To the extent that this is also a social reform it is systematically incomplete without societal agenda setting: What are the projects, the problems to be solved, the targets and design norms of nanotechnoscience?

39) Some are waiting, of course, not for a radically new and progressive way of doing science but for the far-off scientific breakthroughs that inspire speculations about human enhancement and mind–machine interfaces. However, in the context not of pure philosophy but of supposed implications of current research, "revolutionary" human enhancement is a non-issue that can only distract from more urgent questions [62].

References

1 Latour, B. (1987) *Science in Action*, Harvard University Press, Cambridge, MA.

2 Haraway, D. (1997) *Modest_Witness@Second_Millenium*, Routledge, New York.

3 Nordmann, A. (2004) Was ist TechnoWissenschaft? – Zum Wandel der Wissenschaftskultur am Beispiel von Nanoforschung und Bionik, in *Bionik: Aktuelle Forschungsergebnisse in Natur-, Ingenieur- und Geisteswissenschaften* (eds T. Rossmann and C. Tropea), Springer, Berlin, pp. 209–218.

4 Hacking, I. (1983) *Representing and Intervening*, Cambridge University Press, New York.

5 Galison, P. (2006) The pyramid and the ring, presented at the conference of the Gesellschaft für analytische Philosophie (GAP), Berlin.

6 Brune, H., Ernst, H., Grunwald, A., Grünwald, W., Hofmann, H., Krug, H., Janich, P., Mayor, M., Rathgeber, W., Schmid, G., Simon, U., Vogel, V. and Wyrwa, D. (2006) *Nanotechnology: Assesssments and Perspectives*, Springer, Berlin.

7 Nanoscience and Nanotechnologies: Opportunities and Uncertainties, Royal Society and Royal Academy of Engineering, London. 2004.

8 Echeverría, J. (2003) *La Revolución Tecnoscientífica*, Fondo de Cultura Económica de España, Madrid.

9 Johnson, A.(3 March 2005) Ethics and the epistemology of engineering: the case of nanotechnology, presented at the conference Nanotechnology: Ethical and Legal Issues, Columbia, SC.

10 Heisenberg, W. (1974) The notion of a "closed theory" in modern science, in *Across The Frontiers* (ed. W. Heisenberg), Harper and Row, New York, pp. 39–46.

11 Bokulich, A. (2006) Heisenberg meets Kuhn: closed theories and paradigms. *Philosophy of Science*, **73**, 90–107.

12 Schäfer, W. (ed.) (1983) *Finalization in Science*, Reidel, Dordrecht.

13 Wilholt, T. (2006) Design rules: industrial research and epistemic merit. *Philosophy of Science*, **73**, 66–89.

14 Roukes, M. (2001) Plenty of room indeed. *Scientific American*, **285** (3), 48–57.

15 Eberhart, M. (2002) Quantum mechanics and molecular design in the twenty first century. *Foundations of Chemistry*, **4**, 201–211.

16 Maynard, A.D. *et al*. (2006) Safe handling of nanotechnology. *Nature*, **444**, 267–269.

17 Hertz, H. (1956) *The Principles of Mechanics*, Dover, New York.

18 Cartwright, N. (1999) *The Dappled World: a Study of the Boundaries of Science*, Cambridge University Press, Cambridge.

19 Morgan, M. and Morrison, M. (eds) (1999) *Models as Mediators*, Cambridge University Press, Cambridge.

20 Nordmann, A. (2004) Molecular disjunctions: staking claims at the nanoscale, in *Discovering the Nanoscale* (eds D., Baird, A. Nordmann and J., Schummer), IOS Press, Amsterdam, pp. 51–62.

21 Chen, J., Reed, M.A., Rawlett, A.M. and Tour, J.M. (1999) Large on–off rations and negative differential resistance in a molecular electronic device. *Science*, **286**, 1550–1552.

22 Winsberg, E. (2006) Handshaking your way to the top: simulation at the nanoscale, in *Simulation: Pragmatic Constructions of Reality* (eds J. Lenhard, G. Küppers and T. Shinn), Springer Dordrecht (*Sociology of the Sciences Yearbook*, Vol. 25), 139–154.

23 Batterman, R.W. (2006) Hydrodynamics versus Molecular Dynamics: Intertheory Relations in Condensed Matter Physics, in *Philosophy of Science*, **73**, 888–904.

23a Nordmann, A. (2006) Unsichtbare Ursprünge: Herbert Gleiter und der Beitrag der Materialwissenschaft, in *Nanotechnologien im Kontext: Philosophische, ethische und gesellschaftliche Perspektiven* (eds A. Nordmann, J.

Schummer and A. Schwarzl), Akademische Verlagsgesellschaft, Berlin, 81–96.

24 Bensaude-Vincent, B. (2001) The Construction of a Discipline: Material Sciences in the United States. *Historical Studies in the Physical and Biological Sciences*, **31**, 223–248.

25 Bensaude-Vincent, B. and Guchet, X. (2007) Nanomachine: one word for three different paradigms. *Techné*, **11** (1), 71–89.

26 Bensaude-Vincent, B., Two cultures of nanotechnology? in *Nanotechnology Challenges: Implications for Philosophy, Ethics and Society* (eds J. Schummer and D. Baird), World Scientific Publishing, Singapore, 7–28.

27 Whitesides, G.M. and Grzybowski, B. (2002) Self-Assembly at all scales. *Science*, **295**, 2418–2421.

28 Scientific Committee on Emerging and Newly Identified Health Risks (SCENIHR) (2005) Opinion on the Appropriateness of Existing Methodologies to Assess the Potential Risks Associated with Engineered and Adventitious Products of Nanotechnologies, European Commission, Health and Consumer Protection Directorate-General, Brussels.

29 Schummer, J. (2001) Ethics of chemical synthesis. *International Journal for the Philosophy of Chemistry*, **7**, 103–124.

30 Kurath, M. and Maasen, S. (2006) Toxicology as a nanoscience? – disciplinary identities reconsidered. *Particle and Fibre Toxicology*, **3**, 6.

31 Nordmann, A. (2006) Collapse of distance: epistemic strategies of science and technoscience. *Danish Yearbook of Philosophy*, **41**, 7–34.

32 Jones, R. (2004) *Soft Machines*, Oxford University Press, Oxford.

33 Mody, C. (2004) How probe microscopists became nanotechnologists, in *Discovering the Nanoscale* (eds D., Baird, A. Nordmann and J. Schummer), IOS Press, Amsterdam, pp. 119–133.

34 Hennig, J. (2006) Lokale Bilder in globalen Kontroversen: Die heterogenen Bildwelten der Rastertunnelmikroskopie, in *The Picture's Image. Wissenschaftliche Visualisierungen als Komposit* (eds I. Hinterwaldner and M., Buschhaus), Fink, Munich, pp. 243–260.

35 Johnson, A. (2008) Modeling molecules: from computational chemistry to computational nanotechnology. Perspectives on Science, submitted.

36 Scerri, E. (2000) Have orbitals really been observed? *Journal of Chemical Education*, **77**, 1492–1494.

37 Vermaas, P.E. (2004) Nanoscale technology: a two-sided challenge for interpretations of quantum mechanics, in *Discovering the Nanoscale* (eds D. Baird, A. Nordmann and J. Schummer), IOS Press, Amsterdam, pp. 77–91.

38 Harré, R. (2003) The materiality of instruments in a metaphysics for experiments, in *The Philosophy of Scientific Experimentation* (ed. H. Radder), University of Pittsburgh Press, Pittsburgh, PA, pp. 19–38.

39 Winsberg, E. (2007) A tale of two methods, unpublished, http://www.cas.usf.edu/∼ewinsb/papers.html.

40 Mitchell, W. (2005) *What Do Pictures Want? The Lives and Loves of Images*, University of Chicago Press, Chicago.

41 Belting, H. (2001) *Bildanthropologie. Entwürfe für eine Bildwissenschaft*, Fink, Munich.

42 Nordmann, A. (2006) Noumenal technology: reflections on the incredible tininess of nano, in *Nanotechnology Challenges: Implications for Philosophy, Ethics and Society* (eds J. Schummer and D. Baird), World Scientific, Singapore, pp. 49–72. originally published as Nordmann, A. (2005) *Techné*, **8** (3), 3–23.

43 Nordmann, A. Technology naturalized: a challenge to design for the human scale, in *Philosophy and Design: from Engineering to Architecture* (eds P.E. Vermaas, P. Kroes, A. Light and S.A. Moore), Springer, Dordrecht, 173–184.

44 Nordmann, A. (2007) Entflechtung – Ansätze zum ethisch-gesellschaftlichen

Umgang mit der Nanotechnologie, in *Nano – Chancen und Risiken aktueller Technologien* (eds A. Gazsó, S. Greßler and F., Schiemer), Springer, Berlin, pp. 215–229.

45 ETC Group (2003) The strategy for converging technologies. *ETC Communiqué*, **78**, 1–8.

46 Gaskell, G., Ten Eyck, T., Jackson, J. and Veltri, G. (2004) Public Attitudes to nanotechnology in Europe and the United States. *Nature Materials*, **3**, 496.

47 Heidegger, M. (1977) *The Question Concerning Technology and Other Essays*, Harper and Row, New York.

48 Bensaude-Vincent, B. (1992) The balance: between chemistry and politics. *Eighteenth Century*, **33**, 217–237.

49 Böhme, G. and Schäfer, W. (1983) Towards a social science of nature, in *Finalization in Science* (ed. W. Schäfer), Reidel, Dordrecht, pp. 251–269.

50 Janich, P. (2006) Wissenschaftstheorie der Nanotechnologie, in *Nanotechnologien im Kontext: philosophische, ethische und gesellschaftliche Perspektiven* (eds A. Nordmann, J. Schummer and A. Schwarz), Akademische Verlagsgesellschaft, Berlin, pp. 1–32.

51 Grunwald, A. (2006) Nanotechnologie als Chiffre der Zukunft, in *Nanotechnologien im Kontext: philosophische, ethische und gesellschaftliche Perspektiven* (eds A. Nordmann, J. Schummer and A. Schwarz), Akademische Verlagsgesellschaft, Berlin, pp. 49–80.

52 Baird, D. and Shew, A. (2004) Probing the history of scanning tunneling microscopy, in *Discovering the Nanoscale* (eds D. Baird, A. Nordmann and J. Schummer), IOS Press, Amsterdam, pp. 145–156.

53 Baird, D. (2004) *Thing Knowledge: a Philosophy of Scientific Instruments*, University of California Press, Berkeley, CA.

54 Beck, U. (1986) *Risikogesellschaft. Auf dem Weg in eine andere Moderne*, Suhrkamp, Frankfurt. Beck, U. (1992) *Risk Society: Towards a New Modernity*, Sage, London.

55 Dupuy, A. and Grinbaum, J.-P. (2004) Living with uncertainty: toward the ongoing normative assessment of nanotechnology. *Techné*, **8** (2), 4–25.

56 Nordmann, A. (2007) Design choices in the nanoworld: a space odyssey, in *Nano Researchers Facing Choices, The Dialogue Series* (eds M. Deblonde, L. Goorden, *et al.*), Vol. **10**, Universitair Centrum Sint-Ignatius, Antwerp. pp. 13–30. Nordmann, A. (2007) Gestaltungsspielräume in der Nanowelt: Eine Space-Odyssee, in *Nanotechnologie: Erwartungen, Anwendungen, Auswirkungen* (eds D. Korczak and A. Lerf), Asanger, Kröning, pp. 159–184.

57 Tao, N.J. (2006) Electron transport in molecular junctions. *Nature Nanotechnology*, **1**, 173–181.

58 Clifford, C.A. and Seah, M.P. (2006) Modelling of nanomechanical nanoindentation measurements using an AFM or nanoindenter for compliant layers on stiffer substrates. *Nanotechnology*, **17**, 5283–5292.

59 Galison, P. (1997) *Image and Logic: a Material Culture of Microphysics*, Chicago University Press, Chicago.

60 Staley, T.W., The coding of technical images of nanospace: analogy, disanalogy and the asymmetry of worlds. *Techné*, in press.

61 HLEG (High Level Expert Group "Foresighting the New Technology Wave") (2004) Converging Technologies: Shaping the Future of European Societies, Office for Official Publications of the European Communities, Luxembourg.

62 Nordmann, A. (2007) If and then: a critique of speculative nanoethics. *NanoEthics*, **1** (1), 31–46.

8
Ethics of Nanotechnology. State of the Art and Challenges Ahead

Armin Grunwald

8.1
Introduction and Overview

In view of the revolutionary potentials attributed to nanosciences and nanotechnology with respect to nearly all fields of society and individual life [1, 2], it is not surprising that "nano" has attracted great interest in the media and in the public. Parallel to high expectations, for example in the fields of health, growth and sustainable development, there are concerns about risks and side-effects. Analyzing, deliberating and assessing expectable impacts of nanotechnology on future society are regarded as necessary parts of present and further development. There have already been commissions and expert groups dealing with ethical, legal and social implications of nanotechnology (ELSI) [3, 4]. An ethical reflection on nanotechnology emerged and has already led to new terms such as "nano-ethics" and the recent foundation of a new journal *Nano-Ethics*. The quest for ethics in and for nanotechnology currently belongs to public debate in addition to scientific self-reflection. The ethical aspects of nanotechnology discussed in the (so far few) treatises available show broad evidence of this relatively new field of science and technology ethics.

Nanotechnology has been attracting increasing awareness in practical philosophy and in professional ethics. However, there has been some time delay compared with the development of nanotechnology itself: "While the number of publications on NT [nanotechnology] *per se* has increased dramatically in recent years, there is very little concomitant increase in publications on the ethical and social implications to be found" [5, p. R10]. Certain terms, such as privacy, man–machine interface, the relationship between technology and humankind or equity are often mentioned.

In the last few years, ethical reflection on nanotechnology developed quickly and identified many ethically relevant issues [6, 7]. However, well-justified criteria for determining why certain topics, such as nanoparticles or crossing the border between technology and living systems, should be ethically relevant are mostly not given and there is no consensus yet. In particular, the *novelty* of the ethical questions touched by developments emerging from nanotechnology compared with ethical issues in

Nanotechnology. Volume 1: Principles and Fundamentals. Edited by Günter Schmid

ISBN: 978-3-527-31732-5

well-known fields of technology is often not clear. Furthermore, although ethical aspects of nanotechnology have been identified, their analysis and the elaboration of proposals for how to deal with them in society is still at the beginning. Ethics of nanotechnology is, therefore, still an emerging field, in spite of the thematic broadness and the scientific awareness of the ongoing discussion.

In view of this situation, the purpose of this chapter consists primarily in studying current and foreseeable developments in nanotechnology from the viewpoint of philosophical ethics. Which developments are ethically relevant? Are there ethical questions which have already been tackled by current or recent discussions in the ethics of technology or in bioethics? Could the analysis of ethical aspects of nanotechnology benefit from other ethical discussions? Are there developments which pose completely new ethical questions? To this end, it is necessary to clarify the understanding of the notion of "ethics" to be used in order to look for criteria for deciding when something is ethically relevant in a transparent way (Section 8.2). These criteria are then applied to the field of nanotechnology and the ethical challenges are "mapped out" and described briefly, in order to give a broad overview (Section 8.3). This overview is complemented by two in-depth case studies of ethical aspects in nanotechnology: the challenge of dealing with possible risks of nanoparticles and the role of the precautionary principle (Section 8.4) and the human enhancement case, which is directly related to nanotechnology via the debate on "converging technologies" (Section 8.5). Dealing constructively and in a rational way with these ethical challenges requires specific conceptual and methodical developments. In particular, some effort has to be invested into handling the dimension of the future in normative as well as in epistemological regard in a non-partisan way (Section 8.6).

8.2 The Understanding of Ethics[1)]

In modern discussion, the distinction between factual morals on the one hand and ethics as the reflective discipline in cases of moral conflicts or ambiguities on the other has widely been accepted [10]. This distinction takes into account the plurality of morals in modern society. As long as established traditional moral convictions (e.g. religious ones) are uncontroversial and valid among all relevant actors and as long as they are sufficient to deal with the respective situation and do not leave open relevant questions, ethical reflection is not in place. Morals are, in fact, the action-guiding maxims and rules of an individual, of a group or of society as a whole. Ethical analysis, on the other hand, takes these morals as its subjects to reflect on. Ethics is concerned with the justification of moral rules of action, which can lay claim to validity above and beyond the respective, merely particular morals [8]. In

1) This chapter summarises general work of the author in the field of ethics of technology [8, 9] in order to introduce the basic notions to be used in the following in a transparent way. See also [1, Section 6.2].

particular, ethics serves the resolution of conflict situations which result out of the actions or plans of actors based on divergent moral conceptions by argumentative deliberation only, which is grounded in philosophical ideas such as the Categorical Imperative by Immanuel Kant, the Golden Rule or the Pursuit of Happiness (Utilitarianism).

Normative aspects of science and technology lead, in a morally pluralistic society, unavoidably to societal debates at the least and often also to conflicts over technology. We can witness recent examples in the fields of nuclear power and radioactive waste disposal, stem cell research, genetically modified organisms and reproductive cloning. As a rule, what is held to be desirable, tolerable or acceptable is controversial in society. Open questions and conflicts of this type are the point of departure for the *ethics of technology* [10, 11]. Technology conflicts are, as a rule, not only conflicts over technological means (e.g. in questions of efficiency), but also include diverging ideas over visions of the future, of concepts of humanity and on views of society. Technology conflicts are often controversies about *futures*: present images of the future – which are considerably influenced by our illustrations of the scientific and technological advance – are highly contested [12]. The role of the ethics of technology consists of the analysis of the normative structure of technology conflicts and of the search for rational, argumentative and discursive methods of resolving them. In this "continental" understanding, ethics is part of the philosophical profession. In ethical reflection in the various areas of application, however, there are close interfaces to and inevitable necessities for interdisciplinary cooperation with the natural and engineering sciences involved as well as with the humanities. Even transdisciplinary work might be included in cases of requests for broad participation, for example in the framework of participatory technology assessment [13].

Technology is not nature and does not originate of itself, but is consciously produced to certain ends and purposes – namely, to bring something about which would not happen of itself. Technology is therefore always embedded in societal goals, problem diagnoses and action strategies. In this sense, there is no "pure" technology, that is a technology completely independent of this societal dimension. Therefore, research on and development of new technologies always refer to normative criteria of decision-making including expectations, goals to be reached and values involved [14].

But even if technology is basically beset with values, this does not imply that every decision in research and technology development must be scrutinized in ethical regard. Most of the technically relevant decisions can, instead, be classified as a "standard case" in moral respect in the following sense [6, 9]: they do not subject the normative aspects of the basis for the decision (criteria, rules or regulations, goals) to specific reflection, but assume them to be given for the respective situation and accept the frame of reference they create. In such cases, no explicit ethical reflection is, as a rule, necessary, even if normative elements self-evidently play a vital role in these decisions – the normative decision criteria are clear, acknowledged and unequivocal. It is then out of the question that this could be a case of conflict with moral convictions or a situation of normative ambiguity – the information on the normative framework can be integrated into the decision by those affected and by

those deciding on the basis of axiological information, without analyzing it or deliberating on it. The (national and international) legal regulations, the rules of the relevant institutions (e.g. corporate guidelines), where applicable, the code of ethics of the professional group concerned, as well as general societal usage, are elements of this normative framework [14]. The requirements to be fulfilled in the affected normative framework in order that the respective moral situation can be assumed to be a "standard case" can be operationalized according to the following criteria [6, 11]:

- *Pragmatic completeness:* the normative framework has to treat the decision to be made fully with regard to normative aspects.
- *Local consistency:* there must be a "sufficient" measure of consistency between the normative framework's elements.
- *Unambiguity:* among the relevant actors, there must be a sufficiently consensual interpretation of the normative framework.
- *Acceptance:* the normative framework must be accepted by those affected as the basis for the decision.
- *Compliance:* the normative framework also has to be complied with in the field concerned.

Standard situations in moral respect in this sense are governing decision-making in many fields (e.g. in many cases of laboratory work, in public administration or in private businesses). Technical innovations and scientific progress, however, can challenge such situations by presenting new questions or by shaking views previously held to be valid. This is then the entry point for ethical reflection in questions of science and engineering, for the explicit confirmation, modification or augmentation of the normative framework [15] but also for influencing the direction of further development. This conceptual framework provides a point of departure to identify ethically relevant aspects in new technological developments. Whether there are new challenges for ethics in nanotechnology and which ethical questions will be addressed will be investigated against this background.

8.3 Ethical Aspects of Nanotechnology – an Overview

Ethical challenges of nanotechnology are, following the preceding section, challenges of the existing normative frameworks including the social values represented by them by emerging nanotechnological innovations. The task of identifying ethical challenges emerging from nanotechnology, therefore, can be transformed into the search for affected normative frameworks and social values. Because of the novelty of nanotechnology, which still consists mostly of nanoscience [1, 2], corresponding technology development will frequently take place in form of "radical design processes" rather than as "normal design processes" which may be characterized by more incremental approaches (following [14]). Therefore, we can expect that ethical challenges will be relevant in many cases [15]. Establishing a "map" of the

"landscape" of ethical questions caused by nanotech innovation means answering the following questions:

- Which are the ethical aspects of nanotechnology and related innovations in the sense defined above?
- Which of the identified ethical aspects of nanotechnology are *specific* for nanotechnology and *novel* to ethics?
- Where are relations to recent or ongoing ethical debates in other technology fields, if any?

The resulting "map" of ethical aspects described in the following is organized with reference to existing ethical debates (e.g., debates on privacy, equity or human nature). This classification has the advantage that it automatically allows one to refer to normative frameworks and therefore enables us to investigate whether (a) existing frameworks are sufficient to deal with the value problems involved and (b) if not, whether there are new challenges to the frameworks which are specifically caused by nanotechnology. In this way, it is possible to arrive at a structured and well-founded picture of ethical aspects in nanotechnology.[2)]

8.3.1
Equity: Just Distribution of Opportunities and Risks

A first type of ethical aspects of nanotechnology might result from considerations of equity and distributional justice. Ethical questions concern the *distribution* of the benefits of nanotechnology among different groups of the population or among different regions of the world, as well as the spatial and temporal distribution of the risks of nanotechnology [16]: "Nanotech offers potential benefits in areas such as biomedicine, clean energy production, safer and cleaner transport and environmental remediation: all areas where it would be of help in developing countries. But it is at present mostly a very high-tech and cost-intensive science and a lot of the current research is focused on areas of information technology where one can imagine the result being a widening of the gulf between the haves and the have-nots" [5].

Problems of distributive justice are inherent to many fields of technical innovation. Because scientific and technical progress requires considerable investment, it usually takes place where the greatest economic and human resources are already available. Technical progress does, by its nature, tend to increase existing inequalities of distribution. For example, nanotechnology-based medicine will, in all probability, be rather expensive. Questions of equity and of access to (possible) medical treatments could become urgent in at least two respects: *within* industrialized societies, existing inequalities in access to medical care could be exacerbated by a highly technicized medicine making use of nanotechnology and – with regard to *less developed societies* – because likewise, already existing and particularly dramatic inequalities between industrialized and developing nations could be further increased. Apprehensions

2) This analysis builds on earlier work in the field [1, 6] and goes beyond the state reached. In particular, the categorical classification in fields of ethical interest has been improved.

with regard to both of these types of a potential "nano-divide" (after the well-known "digital divide") are based on the assumption that nanotechnology can lead not only to new and greater options for individual self-determination (e.g. in the field of medicine), but also to considerable improvement of the competitiveness of national economies. Current discussions on distributive justice on both national and international levels (in the context of sustainability as well) are therefore likely to gain increased relevance with regard to nanotechnology.

A specific future field of debate with respect to equity will be the "human enhancement" issue ([17]; see Section 8.5 of this chapter). If technologies of improving human performance were to be available then the question arises of who will have access to those technologies, especially who will be able to pay for them and what will happen to persons and groups excluded from the benefits. There could develop a separation of the population into "enhanced" and "normal" people where a situation is imaginable with "normal" to be used as a pejorative attribute [18] and a coercion towards enhancement might occur: "Merely competing against enhanced co-workers exerts an incentive to use neuro-cognitive enhancement and it is harder to identify any existing legal framework for protecting people against such incentives to compete" [19, p. 423]. Special problems can be expected for disabled persons [20].

Equity aspects, however, are not really new ethical aspects caused by nanotechnology, but are rather intensifications of problems of distribution already existing and highly relevant. Problems of equity belong indispensably to modern technology in general and are leading to ongoing and persistent debates in many fields. The digital divide [21] is, perhaps, the best-known example. But also in military respects or with regard to access to medical high-tech solutions these inequalities exist and are debated in many areas. There is no new ethical question behind them but there might be new and dramatic cases emerging driven by nanotechnological R&D. This point has already arrived at the international level of the nanotech debate [22].

8.3.2 Environmental Issues[3)]

Technology is of major importance for the sustainability of humankind's development. On the one hand, technology determines to a large extent the demand for raw materials and energy, needs for transport and infrastructure, mass flows of materials, emissions and amount and composition of waste. Technology is, on the other hand, also a key factor of the innovation system and influences prosperity, consumption patterns, lifestyles, social relations and cultural developments. Therefore, the development, production, use and disposal of technical products and systems have impacts on the ecological, economic and social dimensions of sustainable development. In most cases, these impacts are ambivalent with regard to sustainable development [23]: there are both positive contributions and negative consequences. The overall concept of

3) A detailed analysis of nanotech potential with regard to sustainable development can be found in [26] and the respective Special Issue of the *Journal of Cleaner Production*.

sustainable development requires production and consumption patterns to be shaped in a manner such that the needs of the current generations can be satisfied in a way that they do not limit or threaten opportunities of future generations for satisfying their needs [24, 25]. Of special relevance in this respect are (following [26]):

- The *limited availability of many natural resources* such as clean water, fossil fuels and specific minerals highlights the importance of the efficiency of their use, of recycling and of substituting non-renewable resources by renewable ones.
- The *limited carrying capacity of the environment* (atmosphere, groundwater and surface water, oceans and pedosphere, ecosystems) emphasizes the necessity for limiting or reducing emissions and for regenerating damaged environments.
- The postulate of *intergenerational equity* as a core part of the idea of sustainable development requires consideration of the distribution of risks and benefits of new technologies among the population and in the global dimension ([22]; see also the preceding section of this paper).
- The sustainability issue of *participation* leads to consequences for the processes of opinion-forming and decision-making in shaping technology and its interfaces with the public (e.g., by means of participatory technology assessment [13]).

Many scientists and engineers claim that nanotechnology promises less material and energy consumption and less waste and pollution from production. Nanotechnology is also expected to enable new technological approaches that reduce the environmental footprints of existing technologies in industrialized countries or to allow developing countries to harness nanotechnology to address some of their most pressing needs [22]. Nanoscience and nanotechnology may be a critical enabling component of sustainable development when they are used wisely and when the social context of their application is considered [26]. There are a lot of high expectations concerning positive contributions of nanotechnology to sustainable development.

However, all the potential positive contributions to sustainable development may come at a price. The ambivalence of technology with respect to sustainable development also applies to nanotechnology [27]. The production, use and disposal of products containing nanomaterials may lead to their appearance in air, water, soil or even organisms [1, Chapter 5, 28]. Nanoparticles could eventually be transported as aerosols over great distances and be distributed diffusely. Despite many research initiatives throughout the world, only little is known about the potential environmental and health impacts of nanomaterials. This situation applies also and above all for substances which do not occur in the natural environment, such as fullerenes or nanotubes. The challenge of acting under circumstances with high uncertainties but with the nanoproducts already at the marketplace is the heart of the ethical challenges by nanoparticles (because of the high relevance and because nanoparticles and their possible risks are under intensive public observation today [29, 30], this topic will be dealt with in-depth (see Section 8.4).

Questions of eco- or human toxicity of nanoparticles, on nanomaterial flow, on the behavior of nanoparticles in spreading throughout various environmental media, on

their rate of degradation or agglomeration and their consequences for the various conceivable targets are, however, not ethical questions (see Section 8.2). In these cases, empirical–scientific disciplines, such as human toxicology, eco-toxicology or environmental chemistry, are competent. They are to provide the knowledge basis for practical consequences for working with nanoparticles and for disseminating products based on them. However, as the debate on environmental standards of chemicals or radiation has shown [31, 32], the results of empirical research do not determine how society should react. Safety and environmental standards – in our case for dealing with nanoparticles – are to be based on sound knowledge but cannot logically be derived from that knowledge. In addition, normative standards, for example concerning the intended level of protection, the level of public risk acceptance and other societal and value-laden issues enter the field. Because of this situation, it is not surprising that frequently conflicts about the *acceptability* of risks occur [33, 34] – and this is obviously a non-standard situation in moral respect (see Section 8.2). Therefore, the field of determining the acceptability and the tolerability of risks of nanoparticles is an ethically relevant issue.

In particular, there are a lot of sub-questions where ethical investigation and debate are asked for in the field of nanoparticles. Such questions are [1, Section 6.2]:

- What follows from our present lack of knowledge about the possible side-effects of nanoparticles? This is a challenge to acting rationally under the condition of high uncertainty – a common problem in practical ethics.
- Is the precautionary principle [35] relevant in view of a lack of knowledge and what would follow from applying this principle [28, 31, 32] (see Section 8.4)?
- Which role do the – doubtlessly considerable – *opportunities* of nanoparticle-based products play in considerations of this sort? According to which criteria may benefits and hazards be weighed against each other, especially in cases when the benefits are (relatively) concrete, but the hazards are hypothetical?

A further question would be by which means such balancing could be performed in an inter-subjectively valid and commonly acceptable way. The quantification of risks and benefits by using utility values might be very difficult or even impossible in cases of high uncertainty about the probability, the kind and the extent of a possible damage as well as in cases of ethical problematic procedures of quantification (such as expressing the value of human life in monetarian units).

Are comparisons of the possible risks of nanoparticles with other types of risk possible, in order to learn from them? Can criteria for assessing the nanoparticle risks be gained from experience in developing new chemicals or medicines? Are we allowed to look at risks of our daily life in order to determine what nanoparticle risk should be acceptable in general [31]? Which normative premises enter into such comparisons and by which argumentative means could they be morally justified?

The questions of the acceptability and comparability of risks, the advisability of weighing up risks against opportunities and the rationality of action under uncertainty are, without doubt, of great importance in nanotechnology. A new field of

application is developing here for the ethics of science and technology. The type of questions posed, however, is well known from established discussions on risks (e.g. risks by exposure to radiation or by new chemicals). Really novel ethical questions are not to be expected in spite of the high practical relevance of the field.

From an ethical point of view this situation is well known: there are, on the one hand, positive expectations with regard to sustainable development which legitimate a moral postulate to explore further and to exhaust those potentials. On the other hand, there are risks and uncertainties. This situation is the basic motivation of technology assessment (TA) as an operationalization of ethical reflections on technology [36]. The basic challenge with strong ethical support is *shaping* the further development of nanotechnology in the direction of sustainable development [25, 37].

8.3.3 Privacy and Control

Another field regularly mentioned among the ethical aspects of nanotechnology is the threat to privacy through new monitoring and control technologies. Nanotechnology offers a range of possibilities for gathering, storing and distributing personal data to an increasing extent. In the course of miniaturization, a development of sensor and memory technology is conceivable which, unnoticed by its "victim", drastically increases the possibilities for acquiring data. Furthermore, miniaturization and networking of observation systems (e.g., in the framework of "pervasive" or "ubiquitous" computing) could considerably impede present control methods and data protection regulations or even render them obsolete [38]. Passive observation of people could, in the distant future, be complemented by actively manipulating them – for instance, if it would be possible to gain direct technical access to their nervous system or brain [19, 39].

Within the private sphere, health is a particularly sensitive area. The development of small analyzers – the "lab on a chip" – can make it possible to compile comprehensive personal diagnoses and prognoses on the basis of personal health data. This technology can facilitate not only medical diagnoses, but can also make fast and economical comprehensive screening possible. Everyone could let him- or herself be tested, for example, for genetic dispositions for certain disorders – or could be urged by his/her employer or insurance company to do so. In this manner, individual persons could find themselves put under social pressure. Without sufficient protection of their private sphere, people are rendered manipulable, their autonomy and freedom of action are called into question. Stringent standards for data protection and for the protection of privacy therefore have to be set.

Questions of privacy, of monitoring and controlling people are doubtlessly ethically relevant. But all of these questions of monitoring and of data and privacy protection are not posed exclusively by nanotechnology. Even without nanotechnology, observation technologies have reached a remarkable stage of development which poses questions on the preservation of the private sphere. Even today, so-called smart tags, based on RFID technology (radiofrequency identification), are being employed for access control, such as ticketing, for example, in public transportation and in

logistics. These objects have at present a size of several tenths of a millimeter in each dimension, so that they are practically unnoticeable to the naked eye. Further miniaturization will permit further reductions in size and the addition of more functions – without nanotechnology being needed – but nanotechnology will promote and accelerate these developments.

The ethically relevant questions on a right to know or not to know, on a personal right to certain data, on a right to privacy, as well as the discussions on data protection and on possible undesirable inherent social dynamisms and, in consequence, of a drastic proliferation of genetic and other tests, have been a central point in bio- and medical-ethical discussions for some time. Nanotechnological innovations can accelerate or facilitate the realization of certain technical possibilities and therefore increase the urgency of the problematics of the consequences; in this area, however, they do not give rise to qualitatively new ethical questions.

8.3.4
Military Use of Nanotechnology

Nanotechnology can improve not only multiple peaceful uses but also military and future arms systems. The foundation of the so-called Institute for Soldier Nanotechnologies (http://web.mit.edu/isn/) to enhance soldier survivability makes it clear that nanotechnology has applications in the military field. It is predicted that nanotechnology will bring revolutionary changes in this areas as well [40–42]. Nevertheless, progress in a military technology will not only improve survival and healing, it always implies its use to enhance the efficacy of weapons, surveillance systems and other military equipment. As nanotechnology will provide materials and products that are stronger, lighter, smaller and more sensitive, there may be projectiles with greater velocity and smaller precision-guidance systems. Moreover, nanotechnology will influence the processing in energy generation and storage, displays and sensors, logistics and information systems, all being important elements of warfare. A particular point of interest is the use of BCI (brain–computer interaction) for navigation support for jet pilots, which is being researched by many projects funded by the US defense agency DARPA [43]. In several countries the Departments or Ministries of Defense have arranged nanotechnological programs. No one wants to be left behind [42].

The ethical concerns related to these developments mentioned in the available literature may be classified into the following points [41, 42]:

- An arms race similar to that of nuclear weapons cannot be excluded.
- Present asymmetric power relation could be increased or intensified, for instance to the disadvantage of developing countries [7].
- Some nanobased weapons might be much smaller and, perhaps, cheaper that traditional ones. This could increase the risks of proliferation and of terrorist usage of those weapons.

As long as the military in general is regarded as ethically allowed – and this is the case in most concepts of ethics as far as the military is used for legitimate purposes

such as self-defense of democratic states or legitimate interventions into totalitarian states or for humanitarian reasons – there is little reason to investigate ethical aspects of the use of new technologies in the military areas as a specific topic. At any time, technology has been an important factor in military affairs and had a decisive influence on power relations. Nations with highly developed industrial capacities will be able to exploit the military possibilities of scientific and technological advances to a greater extent than less developed nations.

However, the developments must be observed carefully from the standpoint of peace-keeping and arms control. Possibly areas of international agreements must be reconsidered [41], above all the arms control agreements (e.g. Biological Weapons Convention; limits on conventional forces by new weapons types outside of treaty definitions) or the international laws of warfare (e.g. through the introduction of autonomous fighting systems not reliably discriminating between combatants and non-combatants). There is a certain risk that military "facts" could be created before an open debate about such developments has been launched. In this way, awareness in ethical regard is required in this field but novel ethical questions are not in view at present.

8.3.5 Health

Miniaturization is an essential means of progress in many medical areas. Smaller samples for *in vitro* analysis allow less invasive and less traumatic methods of extraction. Better interfaces and biocompatible materials provide new occasions for implants and restoring damaged organic facilities. The field of medical application seems to be the largest area of future nanotech applications [2]. Nanomedicine is defined as "(1) the comprehensive monitoring, control, construction, repair, defense and improvement of all human biological systems; working from the molecular level, using engineered nanodevices and nanostructures; (2) the science and technology of diagnosing, treating and preventing disease and traumatic injury, of relieving pain and of preserving and improving human health, using molecular tools and molecular knowledge of the human body; (3) the employment of molecular machine systems to address medical problems, using molecular knowledge to maintain and improve human health at the molecular scale" [44, p. 418].

With the help of nanotechnology-based diagnostic instruments, diseases or predispositions for diseases could possibly be discovered earlier than at present [45]. Through the development of "lab-on-a-chip"-technology, the emerging tendency towards personalized medicine would be further promoted. In therapy, there is the prospect, with the help of nanotechnology, of developing targeted treatments free of side-effects. The broad use of nanoparticle dosage systems could lead to progress in medicinal treatment. Through nanotechnological methods, the biocompatibility of artificial implants can be improved. Drug delivery systems could considerably enhance the efficiency of medication and minimize side-effects: "Although many of the ideas developed in nanomedicine might seem to be in the realm of science fiction, only a few more steps are needed to make them come true, so the 'time-to-market' of

these technologies will not be as long as it seems today. Nanotechnology will soon allow many diseases to be monitored, diagnosed and treated in a minimally invasive way and it thus holds great promise of improving health and prolonging life. Whereas molecular or personalized medicine will bring better diagnosis and prevention of disease, nanomedicine might very well be the next breakthrough in the treatment of disease" [46, p. 1012).

Addressing symptoms more efficiently or detecting early onsets of diseases is without doubt recommended by ethics. These potentials are so remarkable that ethical reflection almost seems to be superfluous – if one looks solely at the potentials. A comprehensive analysis, however, has to include – as noted above – also possible side-effects, especially risks [48–50]. New types of responsibility and new tasks for weighing up pros and cons might occur. For example, new forms of drug delivery based on nanotechnology (using fullerenes, nanostructured membranes, gold nanoshells, dendrimers [1, Section 3.3]) could also have consequences which are not expected and might not be welcome. Careful observations of the advance of knowledge and early investigations of possible side-effects have to be conducted. HTA (Health Technology Assessment) offers several established approaches for early warning. The ethically relevant issues are [7]:

- the gulf between diagnostic and therapeutic possibilities and the problem of undesirable information;
- data protection and the protection of privacy (see Section 8.3.3), especially the danger of genetic discrimination;
- preventive medicine and screening programs;
- increases in costs through nanomedicine and problems of access and equity (see Section 8.3.1);
- deferment of illness during the lifetime of humans;
- changes in the understanding of illness and health [52].

However, there is probably no field of science in which dealing with risks is so well established as in medicine and pharmaceutics. Advances in medicine (diagnosis and therapy) are evidently related to risks and there are a lot of established mechanisms such as approval procedures for dealing with them. There is nothing new about this situation. Therefore, using nanotechnology for medical purposes is a standard situation in moral respects (following the notion introduced in Section 8.2). Against this background, it seems to be improbable that direct applications of nanotechnology for medical purposes might lead to completely new ethical questions [47, 51]. The ethical topics to be aware of are not specific for the use of nanotechnology but are also valid for a lot of other advances in medical science and practice.

The boundaries of such a standard situation in moral respects would, however, be transgressed in some more visionary scenarios as in the vision of longevity or the abolition of aging. Nanotechnology could, in connection with biotechnology and, perhaps, neurophysiology, build the technological basis for realizing such visions. In this respect, the idea has been proposed that nanomachines in the human body could permanently monitor all biological functions and could, in case of dysfunction, damage or violation, intervene and re-establish the "correct" status. In this way, an

optimal health status could be sustained permanently [53] which could considerably enlarge the human lifespan. Such methods, however, would require dramatic technological progress [2, Section 7.2.3]. According to the state of present knowledge, neither the prediction of the time needed for such developments nor an assessment of their feasibility at all can seriously be given.

A new area that is both practically and ethically interesting and much closer to realization consists of creating direct connections between technical systems and the human nervous system [39, 54, 55]. There is intensive current work on connecting the world of molecular biology with that of technology. An interesting field of development is nanoelectronic neuro-implants (neurobionics), which compensate for damage to sensory organs or to the nervous system [43]. Micro-implants could restore the functions of hearing and eyesight. Even today, simple cochlear or retina implants, for example, can be realized. With progress in nano-informatics, these implants could approach the smallness and capabilities of natural systems. Because of the undoubtedly positive goals of healing and restoring damaged capabilities, ethical reflection could, in this case, concentrate above all on the definition and prevention of misuse. Technical access to the nervous system, because of the possibilities for manipulation and control which it opens up, is a particularly sensitive issue. A more complex ethical issue would be neuro-cognitive *enhancement* of functions of the brain [19] (see Section 8.5).

Summing up, in the field of medical applications of nanotechnology there are, considered from the standpoint of today's knowledge, no ethical concerns which are specifically related to the use of nanotechnology. There are a lot of positive potentials which probably also will bear risks – these risks, however, might be dealt with by "standard" measures of risk analysis and management established in medical practice [48]. However, things might change in the more distant future (e.g., if there were to be a shift from the classical medical viewpoint to the perspective of "enhancing" human performance; see Section 8.5).

8.3.6
Artificial Life

Basic life processes take place on a nanoscale, because life's essential building blocks (such as proteins) have precisely this size. By means of nanobiotechnology, biological processes will, due to frequently expressed expectations, be made technologically controllable. Molecular "factories" (mitochondria) and "transport systems", which play an essential role in cellular metabolism, can be models for controllable bio-nanomachines. Nanotechnology on this level could permit the "engineering" of cells and allow a "synthetic biology" constructing living systems "from the bottom" or modifying existing living systems (such as viruses) by technical means. An intermeshing of natural biological processes with technical processes seems to be conceivable. The classical barrier between technology and life is increasingly being breached and crossed.

The technical design of life processes on the cellular level, direct links and new interfaces between organisms and technical systems portend a new and highly dynamic scientific and technological field. Diverse opportunities, above all, in the

field of medicine, but also in biotechnology, stimulate research and research funding [1, Section 3.3]. New ethical aspects are certainly to be expected in this field. They will possibly address questions of artificial life and of rearranging existing forms of living systems by technical means, for example the reprogramming of viruses. Their concrete specification, however, will only be possible when research and development can give more precise information on fields of application and products.

Without knowing much about products and systems emerging from the mentioned developments, there is one thing which seems clear already today. We can surely expect discussions about risks because technically modifying or even creating life is morally and with respect a very sensitive field [56]. The corresponding discussions of risks will have structural similarities to the discussion on genetically modified organisms (GMOs) because in both cases the "source code of life" is attached by technical means. It could come to discussions about safety standards for the research concerned, about containment strategies, about "field trials" and release problems. The danger of misuse will be made a topic of debate, such as technically modifying viruses in order to produce new biological weapons that could possibly be used by terrorists. In this area of nanotechnology, opposition, rejection and resistance in society could be feared, comparable to the GMO case. There will be a demand for early dealing with possible ethical and risk problems and for public dialogue and involvement.

In spite of the partly still speculative nature of the subject, ethical reflection of the scientific advance on crossing the border between technology and life does not seem to be premature [49, 50]. There are clear indications that scientific and technical progress will intensify the – at present non-existent – urgency of these questions in the coming years. However, against the questions to be answered in this section, it has to be stated that these ethical questions are not really specific to nanotechnology. The slogan "shaping the world atom by atom" [57] does not make a difference between technology and life and is, therefore, background to crossing the border between technology and life. But the ethical debates to be expected can rely on preceding investigations. Since the 1980s, these subjects have been repeatedly discussed in the debates on GMO, on artificial intelligence and on artificial life. There is a long tradition of ethical thought which has to be taken into account when facing the new challenges in the field.

8.3.7
Human Enhancement[4)]

Within the tradition of technical progress, that has, at all times, transformed conditions and developments – which, until then, had been taken as given, as unalterable fate – into influenceable, manipulable and formable conditions and developments, the human body and its psyche are rapidly moving into the dimension of the formable. The vision of "enhancing human performance" has been conjured

4) Because of the high relevance in current debates and the challenge to traditional thinking involved, this topic has been selected for an in-depth investigation (see Section 8.5).

up, above all, in the field of "converging technologies" [17]. Nanotechnology, in combination with biotechnology and medicine, opens up perspectives for fundamentally altering and rebuilding the human body. At present, research is being done on tissue and organ substitution, which could be realized with the help of nano- and stem cell technologies. Nano-implants would be able to restore human sensory functions or to complement them, but they would also be able to influence the central nervous system (for an overview on types of enhancement, see [20]).

While the examples of medical applications of nanotechnology cited above [45] remain within a certain traditional framework – because the purpose consists of "healing" and "repairing" deviations from an ideal condition of health, which is a classical medical goal –chances (or risks) of a remodeling and "improvement" of the human body are opened up. This could mean extending human physical capabilities, for example to new sensory functions (e.g., broadening the electromagnetic spectrum that the eye is able to perceive). It could, however, also – by means of the direct connection of mechanical systems with the human brain – give rise to completely new interfaces between humans and machines. Even completely technical organs and parts of the body (or even entire bodies) are being discussed, which, in comparison with biological organisms, are supposed to have advantages such as – perhaps – increased stability against external influences [17].

There are initial anthropological questions of our concept of humanity and of the relationship between humanity and technology. With them, however, and at the same time, the question poses itself of how far human beings *can*, *should* or *want to* go in remodeling the human body and to what end(s) this should or could be done. The practical relevance of such ethical questions in view of a possible technical improvement of human beings (with the substantial participation of nanotechnology) may, at first sight, seem limited. Three considerations, however, contest this estimation. First, the vision of the technical enhancement of human beings is actually being seriously advocated. Research projects are being planned in this direction and milestones for reaching this goal are being set up, whereby nanotechnology takes on the role of an "enabling technology". Second, the visions of human enhancement show consequences by merely communicating them – they currently modify the *conditio humana* [58]. Third, technical enhancements are by no means completely new, but are – in part – actually established, as the example of plastic surgery as a technical correction of physical characteristics felt to be imperfections shows, and as is the case in the practice of administering psycho-active substances. It is not difficult to predict that the possibilities and the realization of technical improvements of human beings will increase; demand is conceivable. In view of the moral questions connected with this development and of their conflict potential, ethical reflection is needed in this field already today although the feasibility of the enhancement technologies at all and the time scale of their availability cannot be assessed with any certainty. Because nanotechnology is seen as the "enabling technology" of these developments, the emerging ethical questions can be related to nanotech developments. However, we have to keep in mind that there is a tradition of ethical thought in anthropology, bioethics, medicine ethics and ethics of technology which provides a lot of argumentations and analyses for the field of "human enhancement".

8.4
Nanoparticles and the Precautionary Principle[5)]

Nanoparticles and nanomaterials are among the first outcomes of nanotechnology to enter the marketplace and can already be found in everyday products [59]. From the "imperative of responsibility" [60] or the quest for "projected futures" [61, 62], philosophical conclusions have been drawn for slowing the process of bringing more and more nanoparticles into the environment and the human body without knowing much about possible side-effects. Ethical reflection is asked for concerning the question of how to deal responsibly with this situation involving high uncertainties of knowledge.

8.4.1
The Risk Debate on Nanoparticles

Currently, special attention in public risk debate is being paid to synthetic nanoparticles. A vast potential market for nano-based products is seen in this field. By means of admixtures or specific applications of nanoparticles, for example, new properties of materials can be brought about, for instance, in surface treatment, in cosmetics and in sunscreens. Consumers and citizens could easily come in contact with nanoparticles already today and the probability of directly having contact with synthetic nanoparticles will increase considerably in the next few years because of the expanding market for the respective products [59]. In spite of this situation, there is still little knowledge about possible impacts of nanoparticles on human health and the environment. In order to allow rational risk management strategies, knowledge about ways of spreading, behavior in the atmosphere or in fluids, lifetime of nanoparticles *as* nanoparticles until agglomeration to form other (larger) particles, their behavior in the human body and in the natural environment, and so on, would be needed. Such knowledge, however, is currently not available to an extent which would allow for classical risk management strategies [1, Section 5.2, [27, 28]].

The growing awareness of this risk issue in combination with the fact of having nearly no knowledge available about side-effects of nanotechnology led to severe irritations and to a kind of helplessness in the early stages of that debate. Some statements from that time will illustrate that situation. A re-assurance company stated: "The new element with this kind of loss scenario is that, up to now, losses involving dangerous products were on a relatively manageable scale whereas, taken to extremes, nanotechnology products can even cause ecological damage which is permanent and difficult to contain. What is therefore required for the transportation

5) Parts of this work have been performed within the expert group "Nanotechnology. Assessment and Perspectives" of the European Academy Bad Neuenahr-Ahrweiler [1]. I would like to thank the group members for many fruitful discussions. More specifically, I am deeply indebted to Harald Krug, who introduced the knowledge available in human and eco-toxicology about impacts of nanoparticles into those discussions. An extended version of this chapter is currently being published [66].

of nanotechnology products and processes is an organizational and technical loss prevention programme on a scale appropriate to the hazardous nature of the products" [64, p. 13]. The ETC group postulated a ban on the commercial use of nanoparticles until more knowledge was available: "At this stage, we know practically nothing about the possible cumulative impact of human-made nanoscale particles on human health and the environment. Given the concerns raised over nanoparticle contamination in living organisms, the, ETC group proposes that governments declare an immediate moratorium on commercial production of new nanomaterials and launch a transparent global process for evaluating the socio-economic, health and environmental implications of the technology" [29, p. 72]. A completely different but also far-reaching recommendation aims at "containing" nanotech research: "CRN has identified several sources of risk from MNT (molecular nanotechnology), including arms races, gray goo, societal upheaval, independent development and programmers of nanotech prohibition that would require violation of human rights. It appears that the safest option is the creation of one – and only one – molecular nanotechnology programme and the widespread but restricted use of the resulting manufacturing capability" [65, p. 4]. This containment strategy would imply a secret and strictly controlled nanotech development, which seems to be unrealistic and unsafe as well as undemocratic. These different proposals have enriched (and heated) public and scientific debate on possible nanoparticle risks. Nanotechnology, itself still in an embryo state, experienced itself, at that time, more or less suddenly, as a subject of public risk debate. The actors in the field seemed not to be prepared for this case. Against this background, it is understandable that the first years of the nanotech risk debate may be characterized mainly by mere suspicions, speculations and uncertainties rather than by knowledge-based and rational deliberation.

In analyzing this situation with respect to conclusions for risk management, careful normative reflection is required [35]. More precisely, thinking about the precautionary principle implies the absence of a standard situation in moral, in epistemic and in risk respect ([6]; see also Section 8.2 of this chapter). Ethical reflection is needed to shed light on the normative premises of the options at hand as well as on the criteria of decision-making. Such an ethical "enlightenment" is a necessary precondition of deliberative procedures within which society could identify adequate levels of protection, threshold values or action strategies. Questions of the acceptability and comparability of risks, the advisability of weighing up risks against opportunities and the rationality of action under uncertainty are, without doubt, of great importance in the field of nanoparticles (for the general challenge related to rationality and risk, see [33]). A new field of application is developing here for the ethics of technology, where close cooperation with toxicology, social sciences and jurisprudence is necessary [63].

8.4.2
The Precautionary Principle

Risk management strategies accompanying the implementation of new technologies and the introduction of new materials are standing in a long tradition. In earlier

times, often a "wait-and-see" approach had been taken. New substances have been introduced assuming that either probably no negative developments and impacts at all would occur or that, in case of adverse effects, *ex post* repair and compensation strategies would be appropriate. The asbestos case is one of the best-known experiences where this approach failed and where the failure had dramatic consequences [67].

Such experiences with hazards caused by new materials, by radiation or by new technologies (see [68] for impressive case studies) led to risk regulations in different fields, in order to prevent further negative impacts on health and the environment. Important areas are [1, Section 5.1]:

- regulations for working places with specific risk exposures (nuclear power plants, chemical industry, aircraft, etc.) to protect staff and personnel;
- procedural and substantial regulations for nutrition and food (concerning conservation procedures, maximum allowed concentrations of undesirable chemicals such as hormones, etc.) to protect consumers;
- environmental standards in many areas to sustain environmental quality (concerning ground water quality, maximum allowed rate of specific emissions from fabrication plants, power plants, heating in households, etc.);
- safety standards and liability issues to protect users and consumers (in the field of automobile transport, for power plants, engines, technical products used in households, etc.).

There are established mechanisms of risk analysis, risk assessment and risk management in many areas of science, medicine and technology, for example in dealing with new chemicals or pharmaceuticals. Laws such as the Toxic Substance Control Act in the USA [69] constitute an essential part of the normative framework governing such situations. This "classical" risk regulation is adequate if the level of protection is defined and the risk can be quantified as the product of the probability of the occurrence of the adverse effects multiplied by the assumed extent of the possible damage. In situations of this type, thresholds can be set by law, by self-commitments or following participatory procedures, risks can be either minimized or kept below a certain level and also precautionary measures can be taken to keep particular effects well below particular thresholds by employing the ALARA (as low as reasonably achievable) principle [63]. Insofar as such mechanisms are able to cover challenges at hand to a sufficient extent, there is a standard situation in moral respect (see Section 8.2) regarding the risk issue. As the ongoing debate shows (which will be explained in more detail below), the field of risks of nanoparticles will not be a standard situation in this sense.

As soon as the conditions for the classical risk management approach are no longer fulfilled, uncertainties and ambivalent situations are the consequence. This is the case if, on the one hand, scientific knowledge concerning possible adverse effects is not available at all or is controversial or highly hypothetical or if empirical evidence is still missing. On the other, classical risk management might not be applicable if adverse affects could have catastrophic dimensions with respect to the extent of possible damage, also in case of (nearly) arbitrary small probabilities of their

occurrence. In the field of nuclear power plants, scenarios of this type have been used as counter-arguments against that technology. The catastrophic dimension of possible, at least thinkable, accidents should, in the eyes of opponents, be a decisive argument even in case of an extreme low probability of such events. This type of situation motivated, for example, Hans Jonas [60] to postulate a "heuristics of fear" and the obligation to use the worst scenario as orientation for action.

In the philosophical debate, however, it became clear that Jonas's approach – besides inherent philosophical problems of the naturalistic and teleological approach – might be very appropriate to raise awareness with regard to precautionary situations but completely inadequate to be operationalized by regulation. Jonas's approach missed completely a legitimate procedure for deciding about the applicability and adequacy of precautionary strategies. What can still be learned from Jonas's work is the high relevance of normative reflection in cases where classical risk management would no longer be adequate [35]. Such situations often are welcomed entry points for ideology and interest-driven statements in arbitrary directions. In fact, it is very difficult to identify what a "rational" approach to dealing with non-classical situations could be and in which way it could be proven to be rational [33].

The observation that in many cases severe adverse effects in the course of the introduction of new materials had not been detected in an early stage but rather led to immense damage to human health, the environment and also the economy [68] motivated debates about precautionary regulation measures which could be applied *in advance* of having certain and complete knowledge – because it might then be too late to prevent damage. Wide international agreement on the precautionary principle was reached during the Earth Summit [United Nations Conference on Environment and Development (UNCED)] in Rio de Janeiro in 1992 and became part of Agenda 21: "In order to protect the environment, the precautionary approach should be widely applied by States according to their capabilities. Where there are threats of serious or irreversible damage, lack of full scientific certainty shall not be used as a reason for postponing cost-effective measures to prevent environmental degradation" [70]. The precautionary principle was incorporated in 1992 in the Treaty on the European Union: "Community policy on the environment shall aim at a high level of protection taking into account the diversity of situations in the various regions of the Community. It shall be based on the precautionary principle . . ." (Article 174).

The precautionary principle substantially lowers the (threshold) level for action of governments (see [35] for the following). It considerably alters the situation in comparison with the previous context in which politicians could use (or abuse) a persistent dissent among scientists as a reason (or excuse), simply *not to take action at all*. In this way, political action simply could come much too late. It is, however, a difficult task to establish legitimate decisions about precautionary measures without either to run into the possible high risks of a "wait-and-see" strategy or to overstress precautionary argumentation with the consequence of no longer to be able to act any more or to cause other types of problems (e.g., for the economy) without need. The following characterization of the precautionary principle shows – in spite of the fact that it still does not cover all relevant aspects – the complex inherent structure of the precautionary principle:

> "*Where, following an assessment of available scientific information, there is reasonable concern for the possibility of adverse effects but scientific uncertainty persists, measures based on the precautionary principle may be adopted, pending further scientific information for a more comprehensive risk assessment, without having to wait until the reality and seriousness of those adverse effects become fully apparent*" *(modified from [35])*.

Thinking about applying the precautionary principle generally starts with a *scientific examination*.

An assessment of the state of the knowledge available in science and of the types and extents of uncertainties involved is needed. In assessing the uncertainties involved, *normative* qualifiers come into play [35]. It has to be clarified whether there is *reasonable concern* in this situation of uncertainty. The qualifier "reasonable concern" as employed by the EC guidelines makes no prejudice about the degree of likelihood, but this qualifier relates to a judgment on the quality of the available information [35]. Therefore, the assessment of the knowledge available including its uncertainties is crucial to precautionary reflections.

8.4.3 The Precautionary Principle Applied to Nanoparticles

Following the preceding analysis, the central questions in the current situation concerning the use of nanoparticles and the knowledge about possible impacts are as follows [66]:

1. Is there a precautionary situation at all, characterized by *epistemic uncertainty* or *unquantifiable risk*?
2. Is there "*reasonable concern* for the *possibility* of adverse effects" in the field of nanoparticles to legitimate the application of the precautionary principle?
3. If yes, what would follow from this judgment with respect to adequate precautionary measures?

The first question has to be answered *positive* (e.g. [1, Section 5.2, 27]). There are unknown or uncertain cause-effect relations and still unknown probabilities of risks resulting from the production, use and proliferation of nanoparticles. The scope of possible effects, their degrees and the nature of their "seriousness" (in relation to the chosen level of protection) can currently, even in the best cases, only be estimated in qualitative terms. Therefore, we are witnessing a typical situation of uncertainty where established risk management strategies are not to be applied [3, p. 4f, 27].

The answer to the second question, whether there is "*reasonable concern* for the *possibility* of adverse effects" in the field of synthetic nanoparticles is also to be answered positive. There are first toxicological results from the exposure of rats to high concentrations of specific nanoparticles which showed severe up to lethal consequences. Because the exposure concentrations were extremely high and the transfer of knowledge achieved by the exposure of rats to the situation of humans is

difficult, these results do not allow for the conclusion of evidence of harm to human health – but they support assuming "reasonable concern for the *possibility* of adverse effects" caused by synthetic nanoparticles. There is a strict difference from classical risk management where *evidence of adverse effects* is required, not evidence of their *possibility* (only).

The third question is the most difficult one. In the precautionary situation given by positive answers to the first two questions, the challenge is to identify a "rational" course of action. The suspicion of adverse effects of nanoparticles could serve as a legitimating reason for very strict measures such as a moratorium on nanoparticle use [29, 30]. However, a mere suspicion – this would be sufficient in the above-mentioned Jonas scenario – does not legitimate strict precautionary measures. Instead, this would depend on a scientific assessment of the state of the art and that the quality of the information available [35]. This scientific assessment of the knowledge available for the nanoparticle case has recently been performed [1, Section 5.2] in a comprehensive manner. It resulted that there are indications for nanoparticle risks for health and the environment in some cases but that there is, based on the state of the art, *no reason for serious concern* about adverse effects but well-grounded serious concern about the *possibilities* of such effects. In the same direction: "Taking into account our present-day knowledge, there is, with regard to nano-specific effects (excluding self-organization effects and cumulative effects of mass production), no reason for particularly great concern about global and irreversible effects of the specific technology 'per se', with it being on a par with the justifiable apprehension concerning nuclear technology and genetic engineering" [27, p. 16].

The mere *possibility* of serious harm implied by a wider use of nanoparticles, however, *does not legitimize* using the precautionary principle as an argument for a moratorium or other prohibitive measures. However, because the state of knowledge permanently changes, continuous monitoring and assessment of the knowledge production concerning impacts of nanoparticles on human health and the environment are urgently required. One of the most dramatic lessons which can be learned from the asbestos story is exactly the lesson about the crucial necessity for such a systematic assessment [67].

The next question is what other (and weaker) types of measures are required in the present (precautionary) situation. These could, for example, be self-organization measures in science, such as the application of established codes of conduct or the adaptation of existing regulation schemes to special features of nanoparticles. It seems helpful to reconsider, as an example, the implementation of the precautionary principle in the genetically modified organisms (GMOs) case: "The European Directive 2001/18 (superseding Directive 90/220), concerning the deliberate release of GMOs into the environment, is the first piece of international legislation in which the precautionary principle is translated into a substantial precautionary framework. ... In the framework of this Directive the precautionary principle is translated into a regulatory framework which is based on a so-called *case-by-case* and a *step-by-step* procedure. The case-by-case procedures facilitates a mandatory scientific evaluation of every single release of a GMO. The step by step procedure facilitates a progressive line of development of GMOs by evaluating the

environmental impacts of releases in decreasing steps of physical/biological containment (from greenhouse experiments, to small-scale and large-scale fields tests up to market approval). This procedural implementation of the precautionary principle implies an ongoing scientific evaluation and identification of possible risks" [35]. The application of the precautionary principle would, in this argument, imply "that there is a need for a cautious step-by-step diffusion of risk-related activities or technologies until more knowledge and experience is accumulated" [27, p. 6].

Against this background there is a wide consensus that no really new regulatory measures are needed: ". . . the existing regulative framework is adequate to deal with the introduction of new substances in the nanotechnology sector" [27, p. 27]. There are particular points of criticism concerning existing regulatory systems [69]; these points may be taken into account by modifications and supplements. Treating nanoparticles as new chemicals seems to be the adequate risk management approach [1].

In the framework of the precautionary principle, there are urgent tasks to be done by science [66], especially in order to close the still large knowledge gaps [27]. A second line of activities required consists of systematically collecting all relevant knowledge and of regularly conducting comprehensive evaluations of the respective state of knowledge. A third field of activities should be directed to contributing to public debate and to improving communication at the interface between science and society. The following steps for further action, thereby meeting several of recommendations by other groups [3, 27, 71], have been proposed [1]:

- develop a nomenclature for nanoparticles and assign a new Chemical Abstracts Service (CAS) registry number to engineered nanoparticles;
- group and classify nanomaterials with respect to categories of risk, toxicity and proliferation;
- treat nanoparticles as new substances and develop and approve tools for screening and testing;
- improve the knowledge base in toxicology;
- develop guidelines and standards for good practices in cautiously dealing with nanoparticles;
- avoid or minimize production and unintentional release of waste nanoparticles;
- establish comprehensive evaluation of the state of knowledge and its evaluation with respect to implications for risk management as well as for identifying knowledge gaps which should urgently be closed;
- create institutions to monitor nanotechnologies and the knowledge about possible risks;
- establish a permanent and open dialogue with the public and with industry.

This step-by-step approach conforms to present knowledge and seems to be an adequate way of applying the precautionary principle in a pragmatic way. However, there is no guarantee of preventing all kinds of possible risks by applying these steps. To repeat an important lesson from the asbestos story, the present situation which

may be characterized by "no evidence of harm" must not be reinterpreted as "evidence of no harm" [67].

With respect to the "reasonable concern for the possibility of hazards", specific caution in dealing responsibly with synthetic nanoparticles is required. Such particles should be handled analogously to *new* chemicals even in the case that the chemical composition is well known beyond the nano character. Dealing with new nanoparticles is still based on a case-by-case approach because established nomenclature and classification schemes are not well prepared to be applied to nanoparticles. This situation should be changed as soon as possible [1, 63].

Beyond risk management and regulation, public risk communication has to be observed carefully because irritations in this communication could have a dramatic impact on public acceptance and political judgment. Denying the existence of risks of new technologies often causes mistrust and suspicion instead of creating optimism. To speak frankly about possible risks increases the chance of a trusty relationship between science and society. Building trust in public debate needs an open debate about chances and risks and often requires schemes for comparing different types and amounts of risk to be available [56, 63]. Transparency about the premises of different risk assessment exercises is urgently required. Knowledge about risks includes knowledge about the validity and the limits of that knowledge. Communicative and participative instruments of technology assessment [13, 72] could help improve mutual understanding and public risk assessment.

8.5
Human Enhancement by Converging Technologies

The report "Converging Technologies for Improving Human Performance" [17] caused a worldwide wave of intensive debate about the future of human nature and on the current agenda of science [71]. It was based on earlier debates in the context of biotechnical and genetic improvement strategies [73, 74]. The idea of NBIC (Nano–Bio–Info–Cogno) convergence [1], however, led to a new degree of intensity. Nanotechnology has been seen as the "enabling technology" of such new strategies of enhancement which allows the "converging technologies" to be given a place in a book on nanotechnology. Without any doubt the question for human enhancement will not be covered by a standard situation in moral respect.

In this section, we will use a topic-oriented approach, starting by discussing the "human enhancement" topic first in general and then mentioning the contributions to this old topic of nanotechnologies and the converging technologies (Section 8.5.1). The implications of the human enhancement idea by NBIC technologies for the need for orientations will be discussed (Section 8.5.2), followed by addressing specific ethical questions (Section 8.5.3). As a result, a rather probable scenario of the further development will be given and discussed in an ethical respect (Section 8.5.4).

8.5.1
Human Enhancement: Visions and Expectations

Human enhancement is a very old theme. Mankind's dissatisfaction with itself has been known from ancient times – discontent with mankind's physical endowments, its physical and intellectual capabilities, with its vulnerability to exogenic eventualities such as disease, with the inevitability of aging and, finally, of death, dissatisfaction with its moral capacities or – and this will probably be particularly frequent – with one's physical appearance. Various methods have been developed and established in order to give certain wishes for improvement a helping hand. Today's esthetic surgery, as a kind of business with considerable and further growing returns, is at present the probably most widespread method of human enhancement. Pharmaceuticals enhancing the performance of the mind or preventing tiredness are increasingly used [19]. Extending the physical limits of human capabilities through intensive training in competitive sports can also be understood as enhancement. Making use of *technical means* for improving performance in sport (doping), however, is being practiced as we often can read about in newspapers but is still held to be unsportsmanlike and unethical.

If the types of enhancement just listed apply to individuals (high athletic performance, individual beauty), *collective* human enhancement is, in its turn, also no new topic. Humankind's often deplored defects in terms of morals or civilization led, for example, in the – with regard to morality as well – progress-optimistic European Enlightenment to approaches towards trying to improve humankind *as a whole* through education. Beginning with the individual, above all, in school education, far-reaching processes towards the advancement of human civilization were to be stimulated and supported.

In the twentieth century, the totalitarian regimes in the Soviet Union and in the German Nazi Reich also applied strategies for improvement – due to the respective ideologies, either related to ideas of socio-Darwinist racism and anti-Semitism or to the orthodox communist and anti-bourgeois ideology. These historical developments show that strategies of improvement must be scrutinized very careful with respect to possibly underlying totalitarian ideologies.

In the current discussion of human enhancement, it is neither a question of an improvement through education and culture nor by indoctrination or power, but of *technical* improvement. Initiated by new scientific visions and utopias which are under discussion, completely new possibilities of human development have been proposed. The title of the report of an American research group to the National Science Foundation conveys its program: "Converging Technologies for Improving Human Performance" [17]. Nanotechnology and the Converging (NBIC) Technologies offer, according to this report, far-reaching perspectives for perceiving even the human body and mind as formable, to improve them through precisely targeted technical measures, and, in this manner, also to increase their societal performance. Strategies of enhancement start at the individual level but are aiming, in the last consequence, at the societal level: "Rapid advances in convergent technologies have the potential to enhance both human performance and the nation's productivity. Examples of payoff

will include improving work efficiency and learning, enhancing individual sensory and cognitive capacities, revolutionary changes in healthcare, improving both individual and group efficiency, highly effective communication techniques including brain to brain interaction, perfecting human–machine interfaces including neuromorphic engineering for industrial and personal use, enhancing human capabilities for defense purposes, reaching sustainable development using NBIC tools and ameliorating the physical and cognitive decline that is common to the aging mind" [17, p. 1]. Among these proposed strategies of enhancement are the following:

- *The extension of human sensory faculties:* the capabilities of the human eye can be augmented, for example, with respect to visual acuity ("Eagle Eye") or with regard to a night vision capability by broadening the electromagnetic spectrum visible in the infrared direction; other sensory organs, such as the ear, could likewise be improved, or completely new sensory capabilities, such as the radar sense of bats, could be made accessible to human beings.
- *Expanding memory through technical aids:* it would be conceivable, by means of a chip which could be directly connected to the optic nerve [43], to record all of the visual impressions perceived in real time and to store them externally. In this manner, all of the visual impressions which accumulate in the course of a lifetime could be recalled at any time. In view of our forgetfulness, this could be an attractive idea for many people. Also the human memory could be enhanced by technical means (neuro-cognitive enhancement [19]).
- *Retardation of aging:* according to our present knowledge, aging can, roughly speaking, be interpreted as a form of degradation on the cellular level. If one could succeed in discovering and repairing immediately all forms of such degradation, aging could be greatly delayed or even abolished. For example, the famous nanorobots acting as "submarines" in our bodies could detect and eliminate pathogens: "In the hunt for pathogens, doctors send tiny machines into the furthest recesses of the body. These 'mini-submarines' are so small that they are not visible to the naked eye even as a speck of dust. Rotating cutting devices from the same dwarf world burrow their way through blocked blood vessels to eliminate the causes of heart attacks and strokes" [64, p. 3]. Overcoming aging with the help of nanotechnology would, then, in the sense of medical ethics, be nothing other than fighting epidemics or other diseases.

This thinking aims at broadening human capabilities in comparison with those we traditionally ascribe to a healthy human being. It is obvious that an entire series of ethical or anthropological questions are associated with these visionary expectations (or even just possibilities), which increase the contingency of the *conditio humana* [58, 76]. These questions pertain to the moral *permissibility* or *forbiddenness* of enhancement, to a possible *duty* to enhancement (if such a duty were possible), to the consequences of enhancement with regard to distributive justice (who can afford to have him- or herself enhanced [18]), to the *consequences* for our *concept of humankind* and for the *society of the future*, to the questions of the possible limits of technical enhancement and to legitimizing criteria for drawing such a boundary line. Without any doubt it seems clear that these challenges demarcate a non-standard situation in

moral respect (see Section 8.2) and that, therefore, this is a field where ethical reflection is required in spite of the partially speculative status (see Section 8.3.7).

8.5.2 Occasions of Choice and Need for Orientation

Scientific and technical progress leads to an increase in the options for human action and has therefore at first sight an *emancipatory function*: an augmentation of the possibilities for acting and deciding and a diminution of the conditions which have to be endured as unalterable takes place. Whatever had been inaccessible to human intervention, whatever had to be accepted as non-influenceable nature or as fate becomes an object of technical manipulation or shaping. This is an increase of contingency in the *conditio humana*, a broadening of the spectrum of choices possible among various options [58].

Influencing the faculties of the "healthy" human body in the form of an improvement can be shown to be the *logically consistent continuation of* scientific and technical progress [76]. Up to now the physical capabilities of healthy humans have to be taken as given, as a heritage of the evolution of life. The sensoric capabilities of the eye or the ear, for example, cannot be extended (except by technical means outside the human body, such as microscopes). It would be an act of emancipation from Nature to be able to influence these capabilities intentionally by technical means. New occasions of choice would appear: in which directions would an enhancement be sensible, which additional functions of the human body should be realized, and so on? Also, more individuality could be the result if decisions on specific enhancements could be made at the individual level. In this way, human enhancement seems to contribute further to realizing grand normative ideas of the Era of Enlightenment.

However, there is also another side of the coin. The first price to be paid is the dissolution of established self-evidence about human beings and their "natural" capabilities. The increase of contingency is also an increased need for orientation and decision-making [58]. Second, the outcomes of those decisions about technical improvements can then be attributed to human action and decision – increased accountability and responsibility are the consequence [77]. Third, not only are new options opened but existing ones might be closed. For example, it is rather probable that disabled persons in a society using enhancement technologies to a large extent would have greater problems of conducting their life [20]. Furthermore, problems of access and equity would arise in increasing amount [18]. This ambivalent situation is characteristic of many modern technologies and does not per se legitimize arguments against enhancement but calls for attention and awareness and also for rational debates about dealing with the "dark side" of enhancement technologies.

8.5.3 Human Enhancement – No Simple Answers from Ethics

Increasing humankind's occasions of choice and its emancipation from the givenness of its biological and intellectual constitution also bring about uncertainties, a loss

of security which have been unquestioned up to now and the need for new orientation to answering the above-mentioned general questions, as well as other, more specific questions, for example concerning neuro-implants: "However, ethical dilemmas regarding the enhancement of the human brain and mind are potentially more complex than, for example, genetically enhancing one's growth rate, muscle mass or even appearance because we primarily define and distinguish ourselves as individuals through our behavior and personality" [55]. This diagnosis poses the question of how far humans *may, should, want* to go in the (re-)construction of the human body and mind with the aim of enhancement [18, 74]. In advance of identifying the ethical questions involved, it seems appropriate to highlight an anthropological aspect affected: the difference between healing and enhancement.

Legitimate interventions into the human body and mind are at present carried out with the aims of healing or preventing disease or deficiencies. *Improving* human beings is, as yet, not a legitimate aim of medicine. Although the borderline between healing and enhancing interventions can hardly be drawn unambiguously [74] and although the terms "health" and "illness" are not well clarified [20, 52], there is obviously a categorical difference between the intentional enhancement of human beings and healing disorders [20]. Healing orients itself conceptually on a condition of health held to be ideal. This can either be explicitly defined or merely implicitly understood – in both cases, healing means closing the gap between the actual condition and the assumed ideal condition. What is to be understood under the ideal condition has certainly been defined culturally in different manners in the course of history. In each individual case, however, this is, at least in context, obvious enough. The ophthalmologist who subjects his patient to an eye-test has a conception of what the human eye should be able to perceive. He or she will propose technical improvements of the current state (e.g. a pair of spectacles) only for deviations from this conception and only from a certain degree of deviation on. The purpose of such measures is restoring the normal state, which succeed may more or less well. Traditional medical practice is probably unimaginable without the manner of thinking that a normal or ideal state serves in the background as a normative criterion for defining deviation. Medical treatment does not extend beyond this normal or ideal state. Just this – for medical practice, essential – way of thinking would, in view of the possible technical improvement of human beings, probably become meaningless. A situation would develop where normative frameworks for technical interventions into the human body and mind would be needed but will not be available because this is a completely new situation in moral respects. This poses the question of how far humans *may, should, want* to go in the (re-)construction of the human body and mind with the aim of improving them.

Ethical issues raised by this situation have been characterized with regard to neuro-cognitive enhancement [19]. This classification can be extended to other forms of enhancement also:

- *Safety:* risks of enhancement might consist of unintended side-effects for the brain of the person to be enhanced (e.g., by unsuccessful technical interventions or in the long term).

- *Coercion:* the use of enhancement technologies by parts of the population might exert force on others also to enhance themselves. Otherwise disadvantages in professional or private life might be the consequence.
- *Distributive justice:* access to enhancement technologies might be distributed over the population or between different regions of the world in a very inequitable way, for example because of high costs of enhancement ([18]; for the topic of equity in general, see Section 8.3.2).
- *Personhood and intangible values:* the way in which we see ourselves as humans and our imagination of a "good life" could be changed.

First we have to draw our attention to the fact that the spontaneous rejection with which the concept of human enhancement is often confronted in the population is, in itself, no ethical argument *per se*. The fact that we are not accustomed to dealing with the enhancement issue and the cultural alien-ness of the idea of technically enhanced human beings are social facts and are quite understandable – but they have only limited argumentative force as such. Spontaneous rejection might occur only because of feeling uncomfortable and unfamiliar in thinking about technical enhancement and could be changed by getting more familiar with it. Therefore, feeling and intuition are factual but still have to be scrutinized for whether there are ethical arguments hidden behind them at all and how strong these arguments would be.

The often-mentioned assertion that a human being's "naturalness" would be endangered or even be eliminated by technical improvement is also no strong argument *per se*. Humankind's naturalness or culturality are competing and partially linked patterns of interpretation of the human condition. Using humankind's naturalness as an argument in the sense that we should not technically improve the evolutionarily acquired faculties of sight, hearing, thinking, and so on, just because they are naturally developed and evolutionarily adapted, would be a naive naturalistic fallacy: out of the fact that we find ourselves to be human beings, for instance, with eyes which function only within a certain segment of the electromagnetic spectrum, nothing follows – normatively –directly at all. Limiting human capabilities to the naturally given properties would reduce humans to museum pieces and would blind out the cultural aspects of being humans, to which also belongs transcending the status quo, that is thinking beyond what is given, as has been the thesis of many writers on philosophical anthropology such as Arnold Gehlen.

From these considerations it cannot be concluded that human enhancement is permitted or even imperative. It merely follows that one should not make it too easy on oneself with an ethical repudiation. Strong imperative arguments are, in fact, not in sight [18]. However, argumentatively, the repudiation front also is not very strong. It points to a great extent to the *consequences* of enhancement – consequences which, like the fears of an increasing separation of society [18], are, to a great extent, hypothetical and which can therefore provide only very general and provisional orientation. In the final analysis, the ethical debate seems to narrow itself down to single-case argumentation: which concrete improvement is meant, which goals and

purposes are connected with it, which side-effects and risks are to be apprehended and the question of weighing up these aspects against the background of ethical theories, such as Kantianism or utilitarianism. *Universally* applicable verdicts such as a strong imperative duty or a clear rejection of any improvement whatsoever seem at present to be scarcely justifiable. What follows from this situation for the future is the responsibility to reflect on the *criteria* for the desirability or acceptability of concrete possibilities for enhancement. A lot of work is still in front of ethics [75, 78].

Recently, it has been proposed to structure the field of possible ethical standpoints and perspectives with regard to enhancement technologies ([79, p. 9], applied to cognitive enhancement, but the scheme seems to be more generally usable) in the following, instructive way:

- *Laissez-faire* – emphasizes freedom of individuals to seek and employ enhancement technologies based on their own judgment.
- *Managed technological optimism* – believes that although these technologies promise great benefits, such benefits cannot emerge without an active government role.
- *Managed technological skepticism* – views that the quality of life arises more out of society's institutions than its technologies.
- *Human essentialism* – starts with the notion of human essence (whether God-given or evolutionary in origin) that should not be modified.

The decision to take one of these perspectives and to work with it in conceptualizing future developments in this field depend on the assessment of the strength of ethical arguments pro and con and also on images of "human nature".

8.5.4 Enhancement Technologies – A Marketplace Scenario Ahead?

What would follow from the description of the current debate given above? In order to give an answer we should take a brief look at the field of human cloning, which is not a technical enhancement of humans but would be a deep technical intervention into human life. Human cloning is regarded against the background of many ethical positions, for example Kantian ethics [74] as a form of instrumentalization of humans. The genetic disposition of an individual would be intentionally fixed by cloning. The persons affected would not be able to give their agreement to be cloned in advance. Cloning means an intentional and external determination of the genome of a later individual. An "informed consent", the information of the affected person about the cloning process and its impacts as well as the agreement of that person, would not be possible because the cloning has to be done at the early embryo phase. Therefore, human (reproductive) cloning and research to this end were banned in many countries soon after the clone sheep "Dolly" had been presented.

The ethical debate about human enhancement is, so far, completely different from the debate about human cloning. A restrictive regulation for human enhancement technologies and the research which could enable it, similar to the ban on human cloning technologies, has not been postulated yet. What makes the difference in debating about regulation while the intuitive and spontaneous public reactions are

similar in both fields? The thesis is that the situation in ethical respect itself is completely different.

In the debate on human cloning, there is a strong ethical argument which motivated severe concern. The postulate of human dignity implies, in many interpretations, the idea that humans must not be instrumentalized without their consent. But cloning means determining *intentionally* the genome of an individual without any chance of consent in advance. In contradiction to the usual fertilization, which includes a high degree of statistical influence, cloning would double or multiply a well-known genome. This would imply a determination of the developing persons without any chance of reversibility [74, 80, 81]. This point seems to be the ethical reason behind the strong regulatory measures which have been taken extremely quickly.

In the field of human enhancement, however, things are completely different in this respect. Human enhancement in the fields mostly mentioned [1] would be applied to adults. Additional or improved sensory capabilities, for example, would be implemented in the same way as other medical procedures today: the candidates would be informed about the operational approach, about costs, about the process of adapting the new features, as well as about possible risks and side-effects. After having received such information, the enhancement candidates would either leave the "enhancement station" (in order not to use the term "hospital") or would sign a letter of agreement and, thereby, would give their "informed consent".

Ethical argumentation, therefore, will be performed by using different types of arguments. The arguments given so far are mostly concerned with possible *impacts* of technical enhancement. Expectable problems of distributive justice and equity are often mentioned [18, 50, 75]. Such argumentations, however, work with highly uncertain knowledge about future impacts of enhancement technologies. As can be learned from the history of technology assessment, it is very difficult to assess the consequences of completely new technologies where no or only little experience would be available as a basis for projections. In such cases, normative biases, pure imagination and ideology are difficult to separate from what could be derived from scientific knowledge. Often, an argument can be countered by a contradicting argument – for example, in the case of coercion mentioned above: "The straightforward legislative approach of outlawing or restricting the use of neurocognitive enhancement in the workplace or school is itself also coercive. It denies people the freedom to practice a safe means of self-improvement, just to eliminate any negative consequences of the (freely taken) choice not to enhance" [19, p. 423]. Another example would be the problem of equity: "Unequal access is generally not grounds for prohibiting neurocognitive enhancement, any more than it is grounds for prohibiting other types of enhancement, such as private tutoring or cosmetic surgery . . . . In comparison with other forms of enhancement that contribute to gaps in socioeconomic achievement, from good nutrition to high-quality schools, neuro-cognitive enhancement could prove easier to distribute equitably" [19, p. 423]. Or take arguments which make use of a possible change of the personhood of people by enhancement: "And if we are not the same person on Ritalin as off, neither we are the same person after a glass of wine as before or on vacation as before an exam"

19, p. 424]. Therefore, ethical analysis building on such highly uncertain assumptions about possible impacts could only constitute weak ethical arguments, pro enhancement as well as contra.

In this situation, it is rather probable that the introduction of enhancement technologies could happen according to the *economic market model*. Enhancement technologies would be researched, developed and offered by science and transferred to the marketplace. History shows (see Section 8.5.1) that a demand for such enhancement technologies is imaginable, as is currently the case in the field of esthetic surgery. Public interest would be restricted to possible situations of market failure. Such situations would be of the types discussed above: problems of equity, prevention of risks, questions of liability or avoidance of misuse. This scenario could be perceived as defensive with respect to current ethical standards, as cynical or as poor in ethical respects. Many people might feel uneasy about it. However, at the moment, such a scenario would fit the state of ethical analysis of technical enhancement of the human body and mind.

Such a scenario, however, has not to be equivalent to a *laissez-faire* model of the use of enhancement technologies [79; see the quotation at the end of Section 5.3]. Also, a marketplace scenario would have to be embedded in a societal environment, consisting of normative frameworks [15], ethical standards and regulation. To clarify the ethical issues involved in enhancement technologies and their relations to social values and to regulation is, to a large extent, still a task ahead.

8.6
Conceptual and Methodical Challenges

It is a characteristic trait of modern societies that they draw the orientation needed for opinion formation and decision-making increasingly from debates about future developments and less and less from existing traditions and values [77, 82]. Modern secular and scienticized society generally orients itself, instead of on the past, more on wishes and hopes, but also on fears with regard to the future. The notion of a "risk society" [77] and the global movement towards sustainable development [25] are examples of this approach. Ethical reflection, therefore, has to be related to communication about those future prospects. The necessity of providing orientation in questions such as human enhancement leads to the methodical challenge of applying the "moral point of view" on non-existing but only *projected* ideas with their own uncertainties and ambiguities.

8.6.1
Ethical Assessments of Uncertain Futures

Ethical analysis on nanotech issues is to a large extent confronted with the necessity to deal with, in part far-reaching, future projections [58, 76]. Ethical inquiry in fields such as artificial life or human enhancement takes elements of future communication like visions as its subject. Providing orientation then means to analyze, assess

and judge those far-reaching future prospects on the basis of today's knowledge and seen from the today's moral point of view.

Frequently, the "futures" used in the debate on nanotech and society differ to a large extent and, sometimes, waver between expectations of paradise and fears of apocalyptic catastrophes. Expectations which regard nanotech as the anticipated solution of all of humanity's problems standing in Drexler's [53] technology-optimistic tradition contrast radically with fears expressed in the technology-skeptical tradition of Joy's [83] line of argumentation where self-reproducing nanorobots are no longer simply a vision which is supposed to contribute to the solution of humanity's gravest problems [53], but are communicated in public partially as a nightmare. The visionary pathos in many technical utopias is extremely vulnerable to the simple question of whether everything could not just be completely different – and it is as good as certain that this question will also be asked in an open society. But as soon as it is posed, the hoped effect of futuristic visions evaporates and can even turn into its opposite [1, Section 5.3].

The general problem is that the spectrum of proposed future projections often seems to be rather arbitrary, for example between expectations of paradise and apocalypse, warning against catastrophes in completely diverging directions. One example will be given: the uncertainty of our knowledge about future developments of nanotechnology and its consequences in connection with the immense imagined potential for damage, of possibly catastrophic effects, are taken as an occasion for categorizing even the precautionary principle (see Section 8.4) as insufficient for handling these far-reaching future questions [61, 62]. Instead, the author's view of society's future with nanotechnology leaves open solely the *existential renunciation* of nanotechnology as the only solution, going beyond even Hans Jonas' *Imperative of Responsibility* [60], inasmuch as they formulate a "duty to expect the catastrophe" in order to prevent the catastrophe (an analysis of the inherent contradictory structure of this argument is included in [76]). It seems interesting that the argumentative opponents, the protagonists of human enhancement by converging technologies, also warn against catastrophes, but only in the opposite sense: "If we fail to chart the direction of change boldly, we may become the victims of unpredictable catastrophe" [17, p. 3]. If, however, the ultimate catastrophe is cited in both directions as a threat, this leads to an arbitrariness of the conclusions. Ethics, therefore, is confronted not only with problems of judging states or developments by applying well-justified moral criteria but also with the necessity to assess the "futures" used in these debate with respect to their "rationality".

This situation leads to new methodical challenges for ethical inquiry. Especially the uncertainty of the knowledge available standing behind the future prospects makes it difficult to assess whether a specific development at the human–machine interface or concerning the technical improvement of the human body should be regarded as science fiction (SF), as technical potential in a far future, as a somewhat realistic scenario or as probably becoming part of reality in the near future. Ethical analysis has to take into account this uncertainty and has to be combined with an "epistemology" of future projections [76] and with new types of a future knowledge assessment [84]. In the following, we will propose an "Ethical

Vision Assessment" and an "Ethical Foresight Approach" which would be parts of a concomitant ethical reflection parallel to the ongoing scientific advance, in close relation to that advance but also in a distance which allows independent assessments and judgments.

8.6.2 Ethical Vision Assessment

Far-reaching visions have been put forward in the nanotech debate since its very beginning. A new paradise has been announced since the days of Eric Drexler's book *Engines of Creation* [53]. This line of story-telling about nanotech has been continued in the debate on human enhancement by converging technologies: "People will possess entirely new capabilities for relations with each other, with machines and with the institutions of civilization. ... Perhaps wholly new ethical principles will govern in areas of radical technological advance, such as the routine acceptance of brain implants, political rights for robots and the ambiguity of death in an era when people upload aspects of their personalities to the Solar System Wide Web" [17, p. 19]. It might seem that such speculations should not or could not be subject to ethical inquiry at all.

Because of the "power of visions" in the societal debates – for examples for research funding, motivating young scientists and for public acceptance – it is of great importance to scrutinize such visions carefully instead of denoting them as obscure and fantastic. In spite of the "futuristic" and speculative nature of those visions (see [85] for the notion of "futuristic visions" and [76] and [1] for first steps toward an epistemological analysis of future knowledge), an early ethical dealing with nanotech visions would be an important and highly relevant task in order to allow more transparent debate, especially about science's agenda [1, 71]. There is a need to make those visions more specific, to extract their epistemic and normative key assumptions and ideas and to relate them to other key issues in public debate. Visions in nanotechnology are very different in nature and often refer to more general ideas, such as [49]

- the notion of free and curiosity-driven research as justification and paradigm in itself;
- the Enlightenment ideology of human-centered and emancipatory progress;
- a transhumanist dogma driving forward human evolution by physically changing humans and transforming them into technology;
- the aims and emphases of different religious belief systems including different ideas about the "quality of life";
- the promotion of global development, health and sustainable development, partially following goals of environmentalism, movements for social justice or particular campaign groups in the framework of the civil society;
- a neo-liberal winner-takes-all capitalist system;
- compassion, motivated by the desire to alleviate human suffering;
- a "medical success" culture driven by the felt obligation or goal to leave no disease or condition without a technical solution and to correct every physical disadvantage.

Scrutiny of nanotech visions could be done in the framework of an "ethical vision assessment" [1, Section 8.6]. Vision assessment can be analytically divided in several steps which are not sharply separated and not linearly ordered but which serve different sub-objectives and involve different methods [85]:

1. With respect to analysis (*Vision Analysis*), it would be a question of disclosing the *cognitive and normative* contents of the visions and of judging epistemologically the extent of their reality and practicability, self-evidently on the basis of current knowledge. Then, an important aspect, the prerequisites for the visions' realizability and the time spans involved would have to be investigated. In both analytical steps, observing the language used on the one hand and the question of the antecedents of the predictive statements play a special role: " ... the nanoethics researcher must be attentive to the twists and turns of language which can be symptoms bringing light to the most hidden layers of the scientific or technological imagination" [61]. Further, the visions' *normative* contents have to be reconstructed analytically: the visions of a future society or of the development of human beings, as well as possible diagnoses of current problems, to the solution of which the visionary innovations are supposed to contribute. For a "rational" discussion, the transparent disclosure of the stocks of knowledge, uncertainties and values involved is necessary, above all, with regard to the relationship of fact to fiction [86]. The contribution of such reflective analyses could consist in this respect in the "clarification" of the pertinent communication: the partners in communication should know explicitly what they are talking about as a prerequisite for more rational communication. It is a matter of society's "self-enlightenment" and of supporting the appropriate learning processes.

2. Vision Assessment would, further, include evaluative elements (*Vision Evaluation*). These are questions of how the cognitive aspects are to be categorized, how they can be judged according to the degree of realization or realizability, according to plausibility and evidence and which status the normative aspects have, for example relative to established systems of values or to ethical standards. The purpose is the transparent disclosure of the relationship between knowledge and values, between knowledge and the lack of it and the evaluation of these relationships and their implications. On the one hand, one can draw upon the established evaluation methods of technology assessment, which often include a participative component [13, 72]. On the other, in the field of human enhancement there are some far-reaching questions in a normative respect which stand to discussion and which require ethical and philosophical reflection (see Section 8.5 and [18, 73, 74]).

3. Finally, it is a matter of deciding and acting (*Vision Management*). The question is how the public, the media, politics and science can be advised with regard to a "rational" use of visions. The question of alternatives, either already existing or to be developed, to the visions already in circulation, stands here in the center of interest, in accordance with the basic position of technology assessment, of always thinking in terms of alternatives and options. In this manner, visions

based on technology can be compared with one another or with non-technological visions.

In particular, it would be the assignment of Vision Assessment in an ethical respect to confront the various and, in part, completely divergent normative aspects of the visions of the future directly with one another. This can, on the one hand, be done by ethical analysis and desk research; on the other, however, the representatives of the various positions should discuss their differing judgments in workshops directly with and against one another, in order to lay open their respective premises and assumptions (e.g., using technology assessment procedures [72]).

8.6.3 Ethical Reflection in Technology Foresight

In the last consequence, it would be necessary to transform the future prospects of nanotech visionarists [17] into more negotiable and communicable, transparent scenarios of the future. This could be realized by using the toolbox of (technology) foresight [87]. According to the above-mentioned ideas and challenges, ethical reflection of far-ranging future visions should be combined with research on and procedures for making the various future prospects more transparent. A close relation between an epistemology of future knowledge [84] ethical deliberation emerges. This situation motivates having a brief look into the field of "technology foresight" which has been developed in the last 10 years, especially at the European level [87]. It seems to be possible to benefit from experiences which have been made there in dealing with different kinds of future knowledge and assessments concerning the agenda-setting processes. In particular, it adds to the previously mentioned dimensions of epistemology and ethics possibilities for public debate and involving actors other than researchers and philosophers.

Foresight is the process of looking systematically into the longer-term future of science, technology, the economy and society with the aim of identifying the areas of strategic research (agenda setting) and the emerging generic technologies likely to yield the greatest economic and social benefits at low risk [87]. As the term implies, these approaches involve thinking about emerging opportunities and challenges, trends and breaks from trends, future risks and ways of dealing with them, and so on. Foresight involves bringing together key actors of change and sources of knowledge, in order to develop *strategic visions* and to establish networks of knowledgeable actors who can respond better to policy challenges in awareness of each others' knowledge resources, values, interests and strategic orientations. The contexts in which foresight can be employed are equally wide-ranging: much work to date has focused on national competitiveness and especially the prioritization and development of strategic goals for areas of research in science and technology. Foresight thus occupies the space in which planning, future studies, technology assessment, strategic deliberation and policy development overlap. It is not a matter of academic- or consultancy-based forecasts of the future (though it has to take these into account as necessary knowledge grounds).

As can easily be seen, there are some similarities to the situation described above in the field of nanotech: far-reaching visions and expectations, uncertainty of knowledge, diverse positions of different actors and the aim at influencing agenda setting. Obviously, there are also differences: usually, in foresight processes ethical reflection is not part of the game whereas regional cooperation, creation of new networks, mobilization and contributing to economic welfare by exploiting chances of new technologies and of regional resources are major issues. The reason for bringing the more philosophical issue of assessing and reflecting nanotech visions with respect to epistemological and ethical questions with established foresight methodologies lies in the common challenge of being confronted with the necessity to deal "rationally" with highly uncertain future knowledge and the inseparably interwoven normative elements of expectations, desires and fears often involved. Giving advice to the scientific agenda, for example via contributing to the processes of defining issues and priorities of the public funding of science and technology, following an open and democratic debate, is, therefore, difficult to achieve.

In various foresight exercises, it has been a common experience that in order to arrive at a workable view on the future in the respective field it is crucial to focus on a concrete and tangible topic. Compared with this experience, it seems impossible to apply a foresight exercise to such a broad and grand topic like nanotechnology. On the contrary, more specific subtopics should be addressed such as the use of nanoparticles in cosmetics, developments towards artificial life or opportunities or threats concerning the equity of access to nanotech benefits. Then it is imaginable to set up a foresight process which would

- include research and reflection parts as the vision assessment activities described above;
- involve other societal groups (stakeholders, customers, policymakers, regulators, business people, etc.);
- provide a balanced and ethically as well as epistemologically reflected view on the respective part of the nanotech field which then could be used as a valuable input for debates about science's agenda in this field.

The task of such a foresight exercise would consist, first, of a "rationalization" of diffuse future prospects in a double manner: rationalization concerning the epistemic contents as well as concerning the normative aspects involved. Second, the resulting images of the future should be communicated to and debated with a broader audience in society. In this way, there could be a success in transforming the rather diffuse visionary prospects of nanotechnology into more specific scenarios in specific areas, for instance of the human–machine interface.

8.6.4 Concomitant Ethical Reflection on Nanotechnology

Since the very beginning of ethical reflection in science and technology, there has been an ongoing discussion about an adequate relation in time between scientific–technological advances and ethics. Ethics often seems to pant helplessly behind

technical progress and to fall short of the occasionally great expectations [88]. The rapid pace of innovation in technicization has the effect that ethical deliberations often come too late: when all of the relevant decisions have already been made, when it is long since too late for shaping technology. Technological and scientific progress sets facts which, normatively, can no longer be revised [74]: "It is a familiar cliché that ethics does not keep pace with technology" [38]. This "ethics last" model means that first there have to be concrete technological developments, products and systems which then could be reflected by ethics. Ethics in this perspective, could, at best, act as a repair service for problems which are already on the table.

In contrast, the "ethics first" model postulates comprehensive ethical reflection on possible impacts already *in advance* of technological developments. Ethics actually can provide orientation in the early phases of innovation, for example because future projections and visions emerging on the grounds of scientific and technical advances may be subject to ethical inquiry. Because there are early ideas available about the scientific and technical knowledge and capabilities as well as about their societal impacts – risks as well as chances – long before market entry, it is possible to reflect and discuss their normative implications. For example, Jonas worked on ethical aspects of human cloning long before cloning technology was available even in the field of animals. Obviously, ethical reflection in this model has to deal with the situation that the knowledge about technology and its consequences is uncertain and preliminary.

This does not necessarily mean that ethical deliberations have to be made for absolutely every scientific or technical idea. The problems of a timely occupation with new technologies appear most vividly in the diverse questions raised by the visions of salvation and horror as regards nanotechnology and human enhancement. What sense is there in concerning oneself hypothetically with the ethical aspects of an extreme lengthening of the human life span or with self-replicating nanorobots [38]? The "ethics first" perspective is exaggerated in these cases to such an extent that any relevance will be lost. Most scientists are of the opinion that these are speculations which stem much rather from the realm of science fiction than from problem analysis which is to be taken seriously: "If discussions about the ethics and dangers of nanotechnology become fixated on a worry that exists only in science fiction, we have a serious problem" [89]. We should not forget that ethical reflection binds resources and there should therefore be certain evidence for the "validity" of these visions, if resources are to be invested in them which could then be lacking elsewhere. Therefore, methods of assessing visions of human enhancement are required which allow for an epistemological investigation of the visions under consideration (see the sections above).

Ethical judgment in very early stages of development could provide orientation for shaping the *process* of scientific advance and technological development (e.g., with regard to questions of equity or of risks of misuse). In the course of the continuing concretization of the possibilities for application of nanotechnologies, it is then possible continuously to concretize the – at first abstract – estimations and orientations on the basis of newly acquired knowledge and finally to carry out an ethically reflected technology assessment. In this way, ethical analysis is an ongoing process, accompanying the scientific and technological advance.

Due to nanotechnology's and the converging technologies' early stage of development, we have here a rare case of an advantageous opportunity: there is the chance and also the time for concomitant reflection, as well as the opportunity to integrate the results of reflection into scientific agenda and technology design and thereby to contribute to the further development of science and technology [38]. In view of the visionary nature of many the prospects in nanotechnology and of long and longer spans of time within which the realization of certain milestones can be expected, there is, in all probability, enough time to analyze the questions posed. In general, it applies in this case that this reflective discussion should take place already in the early phases of development, because then the greatest possibilities for influencing the process of scientific development are given. The chances are good that, in the field of nanotechnology, ethical reflection and the societal discussion do not come too late, but can accompany scientific–technical progress critically and can, in particular, help to influence science's agenda by ethically reflected advice.

References

1 Schmid, G., Brune, H., Ernst, H., Grünwald, W., Grunwald, A., Hofmann, H., Janich, P., Krug, H., Mayor, M., Rathgeber, W., Simon, B., Vogel, V. and Wyrwa, D. (2006) *Nanotechnology – Assessment and Perspectives*, Springer, Berlin.

2 Paschen, H., Coenen, C., Fleischer, T., Grünwald, R., Oertel, D. and Revermann, C. (2004) *Nanotechnologie. Forschung und Anwendungen*, Springer, Berlin.

3 Royal Society (2004) *Nanoscience and Nanotechnologies: Opportunities and Uncertainties*, Royal Accademy, London.

4 Nanoforum (2004) *Nanotechnology. Benefits, Risks, Ethical, Legal and Social Aspects of Nanotechnology*, http://www.nanoforum.org, accessed 2 October 2006.

5 Mnyusiwalla, A., Daar, A.S. and Singer, P.A. (2003) Mind the gap. Science and ethics in nanotechnology. *Nanotechnology*, **14**, R9–R13.

6 Grunwald, A. (2005) Nanotechnology – a new field of ethical inquiry? *Science and Engineering Ethics*, **11**, 187–201.

7 Ach, J.S. and Jömann, N. (2006) Size matters. Ethical and social challenges of nanobiotechnology – an overview. LIT, Münster.

8 Grunwald, A. (1999) Ethische Grenzen der Technik? Reflexionen zum Verhältnis von Ethik und Praxis, in *Ethik in der Technikgestaltung. Praktische Relevanz und Legitimation* (eds A. Grunwald and S. Saupe), Springer, Berlin, pp. 221–252.

9 Grunwald, A. (2003) Methodical reconstruction of ethical advises, in *Expertise and Its Interfaces* (eds G. Bechmann and I. Hronszky), Edition Sigma, Berlin, pp. 103–124.

10 Gethmann, C.F. and Sander, T. (1999) Rechtfertigungsdiskurse, in *Ethik in der Technikgestaltung. Praktische Relevanz und Legitimation* (eds A. Grunwald and S. Saupe), Springer, Berlin, pp. 117–151.

11 Grunwald, A. (2000) Against overestimating the role of ethics in technology. *Science and Engineering Ethics*, **6**, 181–196.

12 Brown, N., Rappert, B. and Webster, A. (eds) (2000) *Contested Futures. A Sociology of Prospective Techno-science*, Ashgate, Burlington.

13 Joss, S. and Belucci, S. (eds), (2002) *Participatory Technology Assessment – European Perspectives*, Centre of the Study of Democracy, London.

14 van Gorp, A. (2005) Ethical Issues in Engineering Design: Safety and Sustainability, *Simon Stevin Series in the Philosophy of Technology*, Delft.

15 van Gorp, A. and Grunwald, A. (2008) Ethical responsibilities of engineers in design processes, risks, regulative frameworks and societal division of labour, in preparation.

16 Gethmann, C.F. (1994) Die Ethik technischen Handelns im Rahmen der Technikfolgenbeurteilung, in *Technikbeurteilung in der Raumfahrt – Anforderungen, Methoden, Wirkungen* (eds A. Grunwald and H. Sax), Edition Sigma, Berlin, pp. 146–159.

17 Roco, M.C. and Bainbridge, W.S. (eds) (2002) *Converging Technologies for Improving Human Performance*, National Science Foundation, Arlington, VA.

18 Siep, L. (2005) *Die biotechnische Neuerfindung des Menschen, presented at the XX. Deutscher Kongress für Philosophie*, Berlin.

19 Farah, M.J., Illes, J., Cook-Deegan, R., Gardner, H., Kandel, E., King, P., Parens, E., Sahakian, B. and Wople, P.R. (2004) Neurocognitive enhancement: what can we do and what should we do? *Nature Reviews. Neuroscience*, **5**, 421–425.

20 Wolbring, G. (2005) The Triangle of Enhancement Medicine, Disabled People and the Concept of Health: A New Challenge for HTA, Health Research and Health Policy. Research Paper, Alberta; http://www.cspo.org/ourlibrary/documents/HTA.pdf, accessed 27 July 2007.

21 Krings, B.-J. and Riehm, U. (2006) Die Nutzung und Nichtnutzung des Internets. Eine kritische Reflexion der Diskussion zum "Digital Divide", in *Netzbasierte Kommunikation, Identität und Gemeinschaft. Net-based Communication, Identity and Community* (eds U. Nicanor and A. Metzner-Szigeth), Berlin, pp. 233–251.

22 Salamanca-Buentello, F., Persad, D.L., Court, E.B., Martin, D.K., Daar, A.S. and Singer, P. (2005) Nanotechnology the Developing World. *PLoS Medicine*, **2**, e97.

23 Fleischer, T. and Grunwald, A. (2002) Technikgestaltung für mehr Nachhaltigkeit – Anforderungen an die Technikfolgenabschätzung, in *Technikgestaltung für eine nachhaltige Entwicklung* (ed. A. Grunwald), Edition Sigma, Berlin, pp. 95–146.

24 World Commission on Environment and Development (1987) Our Common Future, World Commission on Environment and Development, Oxford.

25 Grunwald, A. and Kopfmüller, J. (2006) *Nachhaltigkeit*, Campus, Frankfurt, New York.

26 Fleischer, T. and Grunwald, A. (2008) Making nanotechnology developments sustainable. A role for technology assessment? *Journal of Cleaner Production* (forthcoming).

27 Haum, R., Petschow, U., Steinfeldt, M. and von Gleich, A. (2004) *Nanotechnology and Regulation within the Framework of the Precautionary Principle*, Institut für ökologische Wirtschaftsforschung, Berlin.

28 Colvin, V. (2003) *Responsible Nanotechnology: Looking Beyond the Good News*, Center for Biological and Environmental Nanotechnology at Rice University, http://www.eurekalert.org, accessed 2 October 2006.

29 ETC Group, *The Big Down. Atomtech: Technologies Converging at the Nanoscale*, (2003) http://www.etcgroup.org, accessed 2 October 2006.

30 Friends of the Earth, *Nanomaterials, Sunscreens and Cosmetics. Small Ingredients*, (2006) http://www.foe.org/camps/comm/nanotech/nanocosmetics.pdf, accessed 19 November 2006.

31 Gethmann, C.F. and Mittelstrass, J. (1992) Umweltstandards. *GAIA*, **1**, 16–25.

32 Gethmann, C.F., Pinkau, K., Renn, O., Decker, K., Levi, H.W., Mittelstrass, J., Peyerimhoff, S., Putlitz, G. zu, Randelzhofer, A., Streffer, C. and Weinert, F.E. (1998) *Environmental Standards*.

Scientific Foundations and Rational Procedures of Regulation with Emphasis on Radiological Risk Management, Boston.

33 Shrader-Frechette, K.S. (1991) *Risk and Rationality – Philosophical Foundations for Populist Reforms*, University of California Press, Berkeley, CA.

34 Grunwald, A. (2005) Zur Rolle von Akzeptanz und Akzeptabilität von Technik bei der Bewältigung von Technikkonflikten. *Technikfolgenabschätzung Theorie Praxis*, **14**, 54–60.

35 von Schomberg, R. (2005), The precautionary principle and its normative challenges, in *The Precautionary Principle and Public Policy Decision Making* (eds E. Fisher, J. Jones and R. von Schomberg), Edward Elgar, Cheltenham, UK, Northampton, pp. 141–165.

36 Grunwald, A. (1999) Technology assessment or ethics of technology? Reflections on technology development between social sciences and philosophy. *Ethical Perspectives*, **6**, 170–182.

37 Fleischer, T. (2003) Technikgestaltung für mehr Nachhaltigkeit: Nanotechnologie, in *Nachhaltigkeitsprobleme in Deutschland. Analyse und Lösungsstrategien* (eds R. Coenen and A. Grunwald), Edition Sigma, Berlin, pp. 356–373.

38 Moor, J. and Weckert, J. (2003) *Nanoethics: Assessing the Nanoscale from an Ethical Point of View*, Technical University of Darmstadt, Darmstadt.

39 Abbott, A. (2006) In search of the sixth sense. *Nature*, **442**, 125–127.

40 Altmann, J. and Gubrud, A.A. (2002) Risks from military uses of nanotechnology – the need for technology assessment and preventive control, in *Nanotechnology – Revolutionary Opportunities and Societal Implications* (eds M. Roco and R. Tomellini), European Commission, Luxembourg.

41 Altmann, J. (2004) *Military Nanotechnology. Potential Applications and Preventive Arms Control*, Routledge, New York.

42 Haper, T. (2002) *Nanotechnology Arms Race: Why Nobody Wants To Be Left Behind* http://www.nanotechweb.org/articles/column/1/1/11/1, accessed 2 December 2006.

43 Stieglitz, T. (2006) Neuro-technical interfaces to the central nervous system. *Poiesis Praxis*, **4**, 95–110.

44 Freitas, J.A. (1999) *Nanomedicine. Volume I: Basic Capabilities*, Landes Biosciences, Georgetown, TX.

45 European Technology Platform (2006) Nanomedicine – Nanotechnology for Health, European Commission, Luxembourg.

46 Kralj, M. and Pavelic, K. (2003) Medicine on a small scale. How molecular medicine can benefit from self-assembled and nanostructured materials. *EMBO Reports*, **4**, 1008–1012.

47 Farkas, R. and Monfeld, C. (2004) Ergebnisse der Technologievorschau Nanotechnologie pro Gesundheit 2003. *Technikfolgenabschätzung Theorie Praxis*, **13**, 42–51.

48 Baumgartner, C. (2004) Ethische Aspekte nanotechnologischer Forschung und Entwicklung in der Medizin. *Parlament*, **B23–24**, 39–46.

49 Bruce, D. (2006) Ethical and social issues in nanobiotechnologies. *EMBO Reports*, **7**, 754–758.

50 Ach, J. and Siep, L. (eds) (2006) *Nano-bio-ethics. Ethical Dimensions of Nanobiotechnology*, Berlin.

51 MacDonald, C. (2004) Nanotech is novel; the ethical issues are not. *Scientist*, **18**, 3.

52 Gethmann, C.F. (2004) Zur Amphibolie des Krankheitsbegriffs, in *Wissen und Verantwortung. Band 2: Studien zur medizinischen Ethik* (eds A. Gethmann-Siefert and K. Gahl), Alber, Freiburg.

53 Drexler, K.E. (1986) *Engines of Creation – The Coming Era of Nanotechnology*, Anchor Books, Oxford.

54 Scott, S.H. (2006) Converting thoughts into action. *Nature*, **442**, 141–142.

55 Turner, D.C. and Sahakian, B.J. (2006) Ethical questions in functional neuroimaging and cognitive enhancement. *Poiesis Praxis*, **4**, 81–94.

56 Grunwald, A. (2004) The case of nanobiotechnology. Towards a prospective risk assessment. *EMBO Reports*, **5**, 32–36.

57 National Nanotechnology Initiative (1999) Shaping the World Atom by Atom, Washington, DC.

58 Grunwald, A. (2007) Converging technologies: visions, increased contingencies of the conditio humana and search for orientation, *Futures*, **39**, pp 380–392.

59 Luther, W. (ed.) (2004) *Industrial Application of Nanomaterials – Chances and Risks*, http://www.techportal.de, accessed 2 October 2006.

60 Jonas, H. (1979) Das Prinzip Verantwortung, Suhrkamp Frankfurt/Main, English edn.: *The Imperative of Responsibility*, London 1984.

61 Dupuy, J.-P. (2005) The Philosophical Foundations of Nanoethics. Arguments for a Method, Lecture at the University of South Carolina, 3rd March.

62 Dupuy, J.-P. and Grinbaum, A. (2004) Living with uncertainty: toward the ongoing normative assessment of nanotechnology. *Techné*, **8**, 4–25.

63 Renn, O. and Roco, M. (2006) Nanotechnology and the need for risk governance. *Journal of Nanoparticle Research*, **8** (2), pp. 153–191.

64 Münchener Rückversicherung, (2002) Nanotechnology – What is in Store for Us? Münchener Rückversicherungsgesellschaft, http://www.munichre.com/publications/302-03534_en.pdf, accessed 12 November 2006.

65 Phoenix, C. and Treder, M. (2003) *Applying the Precautionary Principle to Nanotechnology*, http://www.crnano.org/Precautionary.pdf, accessed 2 October 2006.

66 Grunwald, A. (2008) Nanotechnology and the precautionary principle, in Nanotechnology and Nanoethics: Framing the Field (ed. F. Jotterand), Berlin, in press.

67 Gee, D. and Greenberg, M. (2002) Asbestos: from 'magic' to malevolent mineral, in *The Precautionary Principle in the 20th Century. Late Lessons from Early Warnings* (eds P. Harremoes, D. Gee, M. MacGarvin, A. Stirling, J. Keys, B. Wynne and S. Guedes Vaz), Sage, London, pp. 49–63.

68 Harremoes, P., Gee, D., MacGarvin, M., Stirling, A., Keys, J., Wynne, B. and Guedes Vaz, S. (eds) (2002) *The Precautionary Principle in the 20th Century. Late Lessons from Early Warnings*, Sage, London.

69 Wardak, A. (2003) Nanotechnology & Regulation, a Case Study Using the Toxic Substance Control Act (TSCA), Woodrow Wilson International Center, Foresight and Governance Project, Paper 2003–2006.

70 United Nations (2002) Report of the United Nations Conference on Environment and Development, A/CONf.151/26 (Vols. I–III), United Nations, New York.

71 Nordmann, A. (2004) *Converging Technologies – Shaping the Future of European Societies*, European Commission, Brussels.

72 Decker, M. and Ladikas, M. (eds) (2004) *Bridges Between Science, Society and Policy, Technology Assessment – Methods and Impacts*, Springer, Berlin.

73 Fukuyama, F. (2002) Our Posthuman Future: Consequences of the Biotechnology Revolution, Farrar, Strauss and Giraux.

74 Habermas, J. (2001) *Die Zukunft der menschlichen Natur*, Suhrkamp, Frankfurt.

75 Khushf, G. (2004) Systems theory and the ethics of human enhancement: a framework for NBIC convergence. *Annals of the New York Academy of Sciences*, **1013**, 124–149.

76 Grunwald, A. (2006) Nanotechnologie als Chiffre der Zukunft, in *Nanotechnologien im Kontext* (eds A. Nordmann, J. Schummer and A. Schwarz), Akademie Verlag, Berlin, pp. 49–80.

77 Beck, U. (1992) *Risk Society, Towards a New Modernity*, Sage, London.

78 Khushf, G. (2004) The ethics of nanotechnology – visions and values for a new generation of science and engineering, in *Emerging Technologies and Ethical Issues in Engineering* (ed National Academy of Engineering), Washington, DC, pp. 29–55.

79 Sarewitz, D. and Karas, T.H. (2006) Policy Implications of Technologies for Cognitive Enhancement, Arizona State University, Consortium for Science, Policy and Outcomes, 3–5 May.

80 Hermerén, G.(1 March 2004) Nano ethics primer, presented at the conference "Mapping Out Nano Risks", Brussels.

81 Habermas, J. (2002) Replik auf Einwände, *Deutsche Zeitschrift für Philosophie*. **50**, pp. 214–226.

82 Stehr, N. (2004) *The Governance of Knowledge*, Sage, London.

83 Joy, B. (2000) Why the future does not need us. *Wired Magazine*, 2nd April, pp. 238–263.

84 Pereira, A.G.,von Schomberg, R. andFuntowicz, S. (2007) Foresight knowledge assessment. *International Journal of Foresight and Innovation Policy*, **4**, pp. 44–59.

85 Grunwald, A. (2004) Vision assessment as a new element of the technology futures analysis toolbox. Proceedings of the EU–US Scientific Seminar: New Technology Foresight, Forecasting and Assessment Methods, Seville. http://www.jrc.es/projects/fta/index.htm, accessed 2 December 2006.

86 Schmidt, J. (2003) Zwischen Fakten und Fiktionen: NanoTechnoScience als Anfrage an prospektive Wissenschaftsbewertung und Technikfolgenabschätzung, in *Zukunftsorientierte Wissenschaft* (eds W. Bender and J. Schmidt), LIT Münster, pp. 207–220.

87 FOREN (2001) A Practical Guide to Regional Foresight, http://foren.jrc.es, accessed 2 December 2006.

88 Ropohl, G. (1995) Die Dynamik der Technik und die Trägheit der Vernunft, in *Neue Realitäten – Herausforderung der Philosophie* (eds H. Lenk and H. Poser), Akademie Verlag, Berlin, pp. 221–237.

89 Ball, P. (2003) *Nanoethics and the Purpose of New Technologies*, Royal Society for Arts, London, http://www.whitebottom.com/philipball/docs/Nanoethics.doc, accessed 2 December 2006.

9
Outlook and Consequences

Günter Schmid

"Nanotechnology could become the most influential force to take hold of the technology industry since the rise of the Internet. Nanotechnology could increase the speed of memory chips, remove pollution particles in water and air and find cancer cells quicker. Nanotechnology could prove beyond our control and spell the end of our very existence as human beings. Nanotechnology could alleviate world hunger, clean the environment, cure cancer, guarantee biblical life spans or concoct super weapons of untold horror .... Nanotechnology could spur economic development through spin-offs of the research. Nanotechnology could harm the opportunities of the poor in developing countries .... Nanotechnology could change the world from the bottom up. Nanotechnology could become an instrument of terrorism. Nanotechnology could lead to the next industrial revolution .... Nanotechnology could change everything."

This only incompletely cited collection of meaningful as well as senseless predictions is listed in a brochure on "The Ethics and Politics of Nanotechnology" by UNESCO published in 2006 [1]. It impressively demonstrates how the understanding of nanotechnology depends from the observer's individual opinion, the spreading medium or from the different political trends. Depending on the respective author's personal attitude, hopes are raised or catastrophes are predicted.

In spite of the variations of the definition of nanotechnology (see Chapters 1 and 2), from a scientific point of view, strictly observed in this and the other books in this series, we must refrain from unserious promises, but also from scenarios drawing the decline of mankind. Both belong to the field of science fiction and are not based on scientific findings. Indeed, there are enough scientifically substantiated facts making nanotechnology one of the most influential technologies we have ever had, in agreement with the very first of the UNESCO headlines cited above. This can be followed from the few examples given in Sections 2.2.1–2.2.7 in Chapter 2, but even more from the following volumes that deal with information technology, medicine, energy and instrumentation on the nanoscale, making observations in the nanoworld possible. Without entering science fiction fields, our present knowledge in nanoscience allows the prognosis that nano-based data storage systems will offer

Nanotechnology. Volume 1: Principles and Fundamentals. Edited by Günter Schmid

ISBN: 978-3-527-31732-5

capacities to store the whole of the world's literature on a single chip and to construct notebooks with the capacity of a computer center of today. Low-priced solar cells, moldable accumulators of high capacities, highly efficient fuel cells and hydrogen storage systems will become available. Last but not least, progress to be expected in medical diagnosis and therapy will revolutionize health care, beginning with enduringly working drug delivery systems, diagnostic potentials, improved by orders of magnitude, up to individual cancer therapies, based on the personal genome of a patient.

These few examples indeed demonstrate the innovative power of nanoscience and -technology. On the other hand, we should not close our eyes to possible dangers linked with the expansion of nanotechnology. From experiences in the past, for instance in the cases of nuclear energy and genetically modified organisms, we should avoid similar mistakes in the case of nanotechnology. Especially scientists are asked to contribute to an objective discussion in public, free of ideologies and pre-opinions.

An indispensable condition to discuss chances and risks of nanotechnology, free from any prejudice, is a minimum knowledge about nanotechnology by the public. This is not easy to achieve, due to the very complex nature of nanotechnology extending from physics to the life sciences. Therefore, education processes have to start as early as possible in primary and secondary schools, in colleges and universities. This is especially necessary in order to protect people from wrong prophets, in a positive as well as in a negative sense. There is no doubt that nanotechnology will change our lives, but it depends on us what this change will look like.

References

1 The Ethics and Politics of Nanotechnology United Nations Educational, Scientific and Cultural Organization (2006), Paris.

1
Pollution Prevention and Treatment Using Nanotechnology

Bernd Nowack

1.1
Introduction

Environmental nanotechnology is considered to play a key role in the shaping of current environmental engineering and science. Looking at the nanoscale has stimulated the development and use of novel and cost-effective technologies for remediation, pollution detection, catalysis and others [1]. However, there is also a wide debate about the safety of nanoparticles and their potential impact on environment and biota [2, 3], not only among scientists but also the public [4, 5]. Especially the new field of nanotoxicology has received a lot of attention in recent years [6, 7]. Nanotechnology and the environment – is it therefore a Janus-faced relationship? There is the huge hope that nanotechnological applications and products will lead to a cleaner and healthier environment [8]. Maintaining and re-improving the quality of water, air and soil, so that the Earth will be able to support human and other life sustainably, are one of the great challenges of our time. The scarcity of water, in terms of both quantity and quality, poses a significant threat to the well-being of people, especially in developing countries. Great hope is placed on the role that nanotechnology can play in providing clean water to these countries in an efficient and cheap way [9]. On the other hand, the discussion about the potential adverse effects of nanoparticles has increased steadily in recent years and is a top priority in agencies all over the world [10, 11]. Figure 1.1 shows the hits for a search for "risk" related to nanotechnology in the Web of Science. Publications that deal in one way or other with "risk" have skyrocketed in the last few years since 2002.

The same properties that can be deleterious for the environment can be advantageous for technical applications and are exploited for treatment and remediation. Figure 1.2 shows a few examples of this Janus face of nanotechnology: engineered particles with high mobility are needed for efficient groundwater remediation, but at the same time this property will render a particle more difficult to remove during water treatment. The toxicity of some nanoparticles can be used for water disinfection where killing of microorganisms is intended, whereas the same property is unwanted

Nanotechnology. Volume 2: Environmental Aspects. Edited by Harald Krug

ISBN: 978-3-527-31735-6

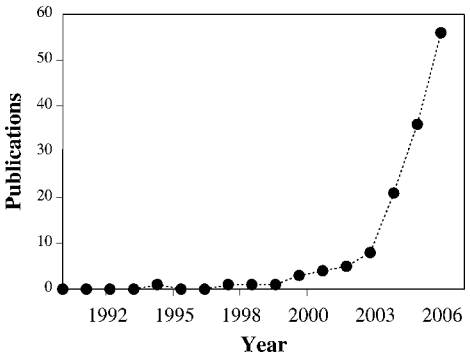


Figure 1.1 Hits in the Web of Science for the search terms "(nanotechnol* OR nanopart* OR nanotub*) AND risk" for the years 1990–2006.

when nanoparticles eventually enter the environment. The catalytic activity of a nanoparticle can be advantageous when used for the degradation of pollutants, but can induce a toxic response when taken up by a cell. The high sorption capacity of certain nanoparticles is exploited for the removal of organic and inorganic pollutants while this property may also mobilize sequestered pollutants in the environment. The engineering of nanoparticles that are easily taken up by cells will have a huge impact on medicine and pharmacological research, but the dispersion of such particles in the environment can lead to unwanted and unexpected effects. Also the

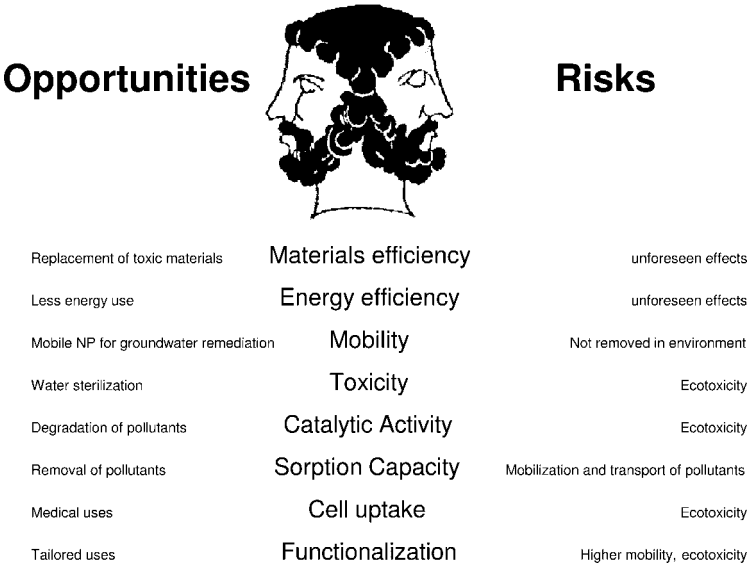


Figure 1.2 The Janus face of nanotechnology.

fact that many engineered nanoparticles are functionalized and therefore have a different surface activity from pristine particles is pivotal for many applications where a tailored property is needed, but such particles may behave in a completely different way from standard particles in the environment and may, for example, be much more mobile or show an increased (or decreased, as the case may be) toxicity. This short list of properties exemplifies the fact that engineered nanoparticles or nanotechnological applications make use of the same properties that are looked for by environmental scientists.

This chapter will give a general overview of potential environmental applications of nanotechnology and nanoparticles and will also give a short overview of the current knowledge about possible risks for the environment.

1.2 More Efficient Resource and Energy Consumption

Pollution prevention by nanotechnology refers on the one hand to a reduction in the use of raw materials, water or other resources and the elimination or reduction of waste and on the other hand to more efficient use of energy or involvement in energy production [1]. The implementation of green chemistry principles for the production of nanoparticles and for nanotechnological applications in standard chemical engineering will lead to a great reduction in waste generation, less hazardous chemical syntheses, improved catalysis and finally an inherently safer chemistry [12]. However, there are very few data that actually show quantitatively that these claims are true and that replacing traditional materials with nanoparticles really does result in less energy and materials consumption and that unwanted or unanticipated side effects do not occur.

Nanomaterials can be substituted for conventional materials that require more raw material, are more energy intensive to produce or are known to be environmentally harmful [8]. Some new nanocatalysts can be used at much lower temperatures than conventional catalysts and therefore require less energy input [13]. The capacity of nanocatalysts to function at room temperature opens the way for broad applications of nanomaterials in many consumer products. Another example of how nanotechnology can reduce energy costs is nanomaterial coatings on ships, which are expected to realize fuel savings on the order of $460 million per year for commercial shipping in the USA [13]. Nanodiamonds are expected to increase the life expectancy of automotive paints and therefore to reduce material costs and expenditure [14]. Nanotechnology may also transform energy production and storage by providing alternatives to current practices. One example is nanoparticulate catalysts for fossil fuels [15], which will lead to reduced emissions or better energy efficiency, higher storage capacity for hydrogen [16, 17], biohydrogen production [18] and more effective and cheaper solar cells or coatings on windows that reduce heat loss [19]. Nanoparticles can increase the storage capacity of batteries and rechargeable batteries [20–22] or are used in flat screens where they reduce the amount of heavy metals [8].

1.3 Pollution Detection and Sensing

Various nanostructured materials have been explored for their use in sensors for the detection of different compounds [23]. An example is silver nanoparticle array membranes that can be used as flow-through Raman scattering sensors for water quality monitoring [24]. The particular properties of carbon nanotubes (CNTs) make them very attractive for the fabrication of nanoscale chemical sensors and especially for electrochemical sensors [25–28]. A majority of sensors described so far use CNTs as a building block. Upon exposure to gases such as NO_2, NH_3 or O_3, the electrical resistance of CNTs changes dramatically, induced by charge transfer with the gas molecules or due to physical adsorption [29, 30]. The possibility of a bottom-up approach makes the fabrication compatible with silicon microfabrication processes [31]. The connection of CNTs with enzymes establishes a fast electron transfer from the active site of the enzyme through the CNT to an electrode, in many cases enhancing the electrochemical activity of the biomolecules [27]. In order to take advantage of the properties of CNTs, they need to be properly functionalized and immobilized. CNT sensors have been developed for glucose, ethanol, sulfide and sequence-specific DNA analysis [27]. Trace analysis of organic compounds, e.g. for the drug fluphenazine, has also been reported [32]. Nanoimmunomagnetic labeling using magnetic nanoparticles coated with antibodies specific to a target bacterium have been shown to be useful for the rapid detection of bacteria in complex matrices [33].

1.4 Water Treatment

Clean water is a requirement for all properly functioning societies worldwide, but is often limited. New approaches are continually being examined to supplement traditional water treatment methods. These need to be lower in cost and more effective than current techniques for the removal of contaminants from water. In this context also nanotechnological approaches are considered. In this section the following application areas will be covered: nanoparticles used as potent adsorbents, in some cases combined with magnetic particles to ease particle separation; nanoparticles used as catalysts for chemical or photochemical destruction of contaminants; nanosized zerovalent iron used for the removal of metals and organic compounds from water; and nanofiltration membranes.

1.4.1 Adsorption of Pollutants

Sorbents are widely used in water treatment and purification to remove organic and inorganic contaminants. Examples are activated carbon and ion-exchange resins. The use of nanoparticles may have advantages over conventional materials due to

the much larger surface area of nanoparticles on a mass basis. In addition, the unique structure and electronic properties of some nanoparticles can make them especially powerful adsorbents. Many materials have properties that are dependent on size [34]. Hematite particles with a diameter of 7 nm, for example, adsorbed Cu ions at lower pH values than particles of 25 or 88 nm diameter, indicating the uniqueness of surface reactivity for iron oxides particles with decreasing diameter [35]. However, another study found that normalized to the surface area the nanoparticles had a lower adsorption capacity than bulk TiO_2 [36]. Several types of nanoparticles have been investigated as adsorbents: metal-containing particles, mainly oxides, carbon nanotubes and fullerenes, organic nanomaterials and zeolites.

For the removal of metals and other inorganic ions, mainly nanosized metal oxides [37, 38] but also natural nanosized clays [39] have been investigated. Also, oxidized and hydroxylated CNTs are good adsorbers for metals. This has been found for various metals such as Cu [40], Ni [41, 42], Cd [43, 44] and Pb [45, 46]. Adsorption of organometallic compounds on pristine multi-walled CNTs was found to be stronger than for carbon black [47].

Chemically modified nanomaterials have also attracted a lot of attention, especially nanoporous materials dues to their exceptionally high surface area [48]. The particle size of such materials is, however, not in the nano-range but normally 10–100 μm. Another option is to modify chemically the nanoparticle itself [49]. TiO_2 functionalized with ethylenediamine was, for example, tested for its ability to remove anionic metals from groundwater [50].

CNTs have attracted a lot of attention as very powerful adsorbents for a wide variety of organic compounds from water. Examples include dioxin [51], polynuclear aromatic hydrocarbons (PAHs) [52–54], DDT and its metabolites [55], PBDEs [56], chlorobenzenes and chlorophenols [57, 58], trihalomethanes [59, 60], bisphenol A and nonylphenol [61], phthalate esters [62], dyes [63], pesticides (thiamethoxam, imidacloprid and acetamiprid) [64] and herbicides such as sulfuron derivatives [65, 66], atrazine [67] and dicamba [68]. Cross-linked nanoporous polymers that have been copolymerized with functionalized CNTs have been demonstrated to have a very high sorption capacity for a variety of organic compounds such as *p*-nitrophenol and trichloroethylene [69]. It was found that purification (removal of amorphous carbon) of the CNTs improved the adsorption [54]. The available adsorption space was found to be the cylindrical external surface; neither the inner cavity nor the inter-wall space of multi-walled CNT contributed to adsorption [70]. Unlike the case with fullerenes, no adsorption–desorption hysteresis was observed, indicating reversible adsorption [70].

Fullerenes have also been tested for adsorption of organic compounds. Adsorption depends to a great extent on the dispersion state of the C_{60} [71], which is virtually insoluble in water [72]. Because C_{60} forms clusters in water, there are closed interstitial spaces within the aggregates into which the compounds can diffuse, which leads to significant adsorption–desorption hysteresis [70, 73]. Fullerenes are only weak sorbents for a wide variety of organic compounds (e.g. phenols, PAHs, amines), whereas they are very efficient for the removal of organometallic compounds (e.g. organolead) [74].

An interesting application is oxide–CNT composites, which have been explored for the removal of metals [75] and also of anions such as arsenate and fluoride [76, 77]. Specially designed polymers and dendrimers are exploited for their potential removal of metals and organics [78, 79].

1.4.2 Magnetic Nanoparticles

Magnetic nanoparticles offer advantages over non-magnetic nanoparticles because they can easily be separated from water using a magnetic field. Separation using magnetic gradients, the so-called high magnetic gradient separation (HGMS), is a process widely used in medicine and ore processing [80]. This technique allows one to design processes where the particles not only remove compounds from water but also can easily be removed again and then be recycled or regenerated. This approach has been proposed with magnetite (Fe_3O_4), maghemite (γ-Fe_2O_3) and jacobsite ($MnFe_2O_4$) nanoparticles for removal of chromium(VI) from wastewater [81–83]. Water-soluble CNTs have been functionalized with magnetic iron nanoparticles for removal of aromatic compounds from water and easy separation from water for re-use [84].

1.4.3 Nanofiltration

Nanofiltration membranes (NF membranes) are used in water treatment for drinking water production or wastewater treatment [85]. NF membranes are pressure-driven membranes with properties between those of reverse osmosis and ultrafiltration membranes and have pore sizes between 0.2 and 4 nm. NF membranes have been shown to remove turbidity, microorganisms and inorganic ions such as Ca and Na. They are used for softening of groundwater (reduction in water hardness), for removal of dissolved organic matter and trace pollutants from surface water, for wastewater treatment (removal of organic and inorganic pollutants and organic carbon) and for pretreatment in seawater desalination.

Carbon nanotubes have been arranged to form a hollow monolithic cylindrical membrane [86], which was efficient for the removal of bacteria or hydrocarbons and that can easily be regenerated by ultrasonication or autoclaving.

1.4.4 Degradation of Pollutants

The semiconductor TiO_2 has been extensively studied for oxidative or reductive removal of organic pollutants [49, 87]. Illumination promotes an electron to the conduction band, leaving a hole in the valence band. This process produces a potent reducing and oxidizing agent. In water, photo-oxidation occurs primarily through hydroxyl radicals. Because TiO_2 requires ultraviolet light for excitation, it has been sensitized to visible light by dyes, through incorporation of transition metal ions [49]

or by doping with nitrogen [88]. The degradation rate of several dyes by nanosized TiO_2 was found to be 1.6–20 times higher than for bulk TiO_2 particles [89]. Several types of compounds such as dyes [88, 90] and organic acids [91] have been shown to be rapidly degraded. A special type of TiO_2 photocatalysts are titania nanotube materials, which were shown to have superior activity [92, 93].

1.4.5 Zerovalent Iron

Laboratory research has established that nanoscale metallic iron is very effective in destroying a wide variety of common contaminants such as chlorinated methanes, brominated methanes, trihalomethanes, chlorinated ethenes, chlorinated benzenes, other polychlorinated hydrocarbons, pesticides and dyes [94]. The basis for the reaction is the corrosion of zerovalent iron in the environment:

$$2Fe^0 + 4H^+ + O_2 \rightarrow 2Fe^{2+} + 2H_2O$$

$$Fe^0 + 2H_2O \rightarrow Fe^{2+} + H_2 + 2OH^-$$

Contaminants such as tetrachloroethane can readily accept the electrons from iron oxidation and be reduced to ethene:

$$C_2Cl_4 + 4Fe^0 + 4H^+ \rightarrow C_2H_4 + 4Fe^{2+} + 4Cl^-$$

However, nanoscale zerovalent iron (nZVI) can reduce not only organic contaminants but also the inorganic anions nitrate, which is reduced to ammonia [95, 96], perchlorate (plus chlorate or chlorite), which is reduced to chloride [97], selenate [98], arsenate [99, 100], arsenite [101] and chromate [102, 103]. nZVI is also efficient in removing dissolved metals from solution, e.g. Pb and Ni [102, 104]. The reaction rates for nZVI are at least 25–30 times faster and also the sorption capacity is much higher compared with granular iron [105]. The metals are either reduced to zerovalent metals or lower oxidation states, e.g. Cr(III), or are surface complexed with the iron oxides that are formed during the reaction. Some metals can increase the dechlorination rate of organics and also lead to more benign products, whereas other metals decrease the reactivity [106].

The reaction rates for nZVI can be several orders of magnitude faster on a mass basis than for granular ZVI [107]. Because the reactivity of ZVI towards lightly chlorinated and brominated compounds is low and because the formation of a passivating layer reduces the reactivity with time, many approaches have been explored where the surface is doped with a catalyst (e.g. Pd, Pt, Cu, Ni) to reduce the activation energy. The same approach has also been tested for nZVI. Surface-normalized reaction rates for such materials were found to be up to 100 times faster than for bulk ZVI [108–111].

The nanoscale iron particles can be produced either by a top-down approach (e.g. milling of iron filings) or by direct chemical synthesis [105]. A common method for synthesis of iron nanoparticles is by reduction of an aqueous ferric solution by reducing agents such as sodium borohydride or sodium hypophosphite [49].

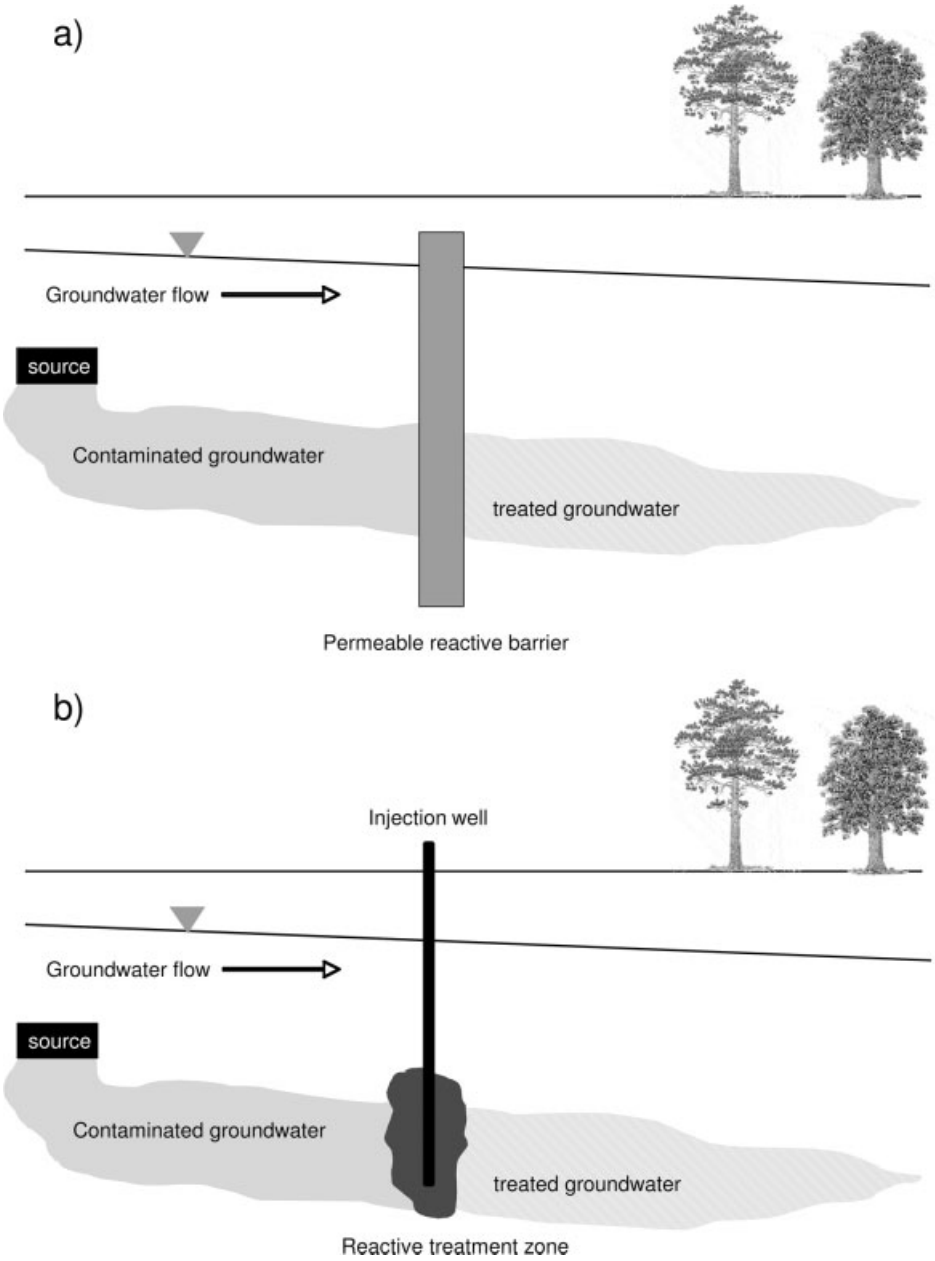


Figure 1.3 Three approaches to application of ZVI for groundwater remediation: (a) conventional reactive barrier using granular ZVI; (b) injection of nZVI to form an immobile reaction zone; (c) injection of mobile nZVI. Modified after [107].

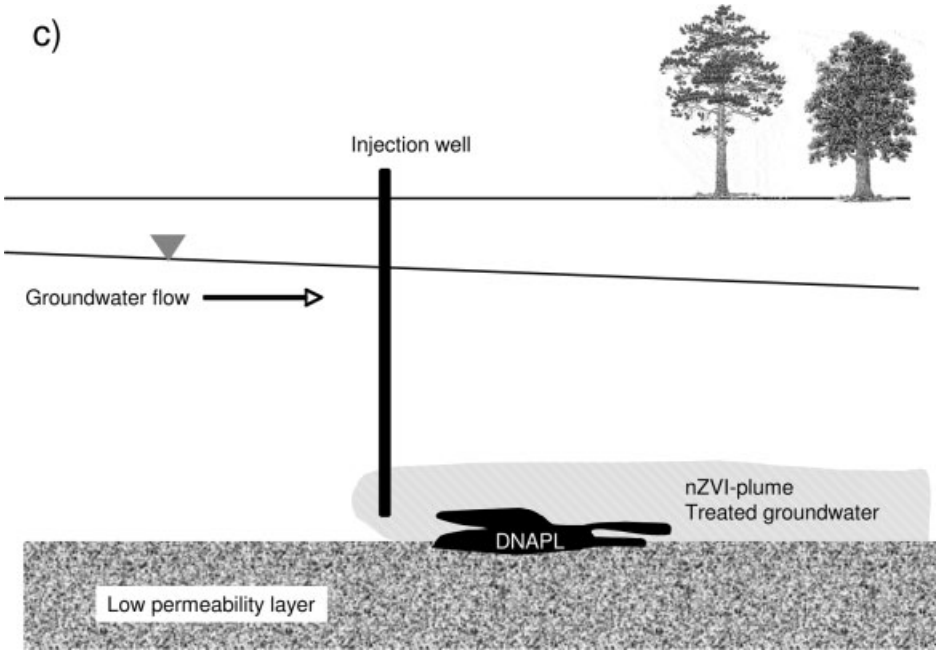


Figure 1.3 *(continued)*

1.5 Soil and Groundwater Remediation

The use of nZVI for groundwater remediation represents the most widely investigated environmental nanotechnological technique. Granular ZVI in the form of reactive barriers has been used for many years at numerous sites all over the world for the remediation of organic and inorganic contaminants in groundwater (see Figure 1.3a). With nZVI, two possible techniques are used: immobile nZVI is injected to form a zone of iron particles adsorbed on the aquifer solids (Figure 1.3b) or mobile nZVI is injected to form a plume of reactive Fe particles that destroy any organic contaminants that dissolve from a DNAPL (dense non-aqueous phase liquid) source in the aquifer (Figure 1.3c). With this technique, the formation of a pollutant plume is inhibited. The successful results of field demonstrations using nZVI have been published, with reported reductions in TCE of up to 96% after injection of 1.7 kg of nanoparticles into the groundwater [112]. A larger test was conducted where 400 kg of nZVI was injected and significant reductions in TCE soil concentration (>80%) and dissolved concentrations (57–100%) were observed [113]. To date approximately 30 projects are under way in which nZVI is used for actual site remediation [105].

Whereas most research using nZVI has been devoted to groundwater, much less has been published about soil remediation. These studies have mostly been done in soil slurries and efficient removal of PAHs by nZVI has been reported [114, 115]. For

PCBs, a removal of only about 40% was attained, caused by the very strong adsorption of PCBs to the soil matrix and limited transfer to the nZVI particles [116]. nZVI has also been used to immobilize Cr(VI) in chromium ore processing residue [117].

Because the iron particles have a strong tendency to aggregate and adsorb on surfaces of minerals, much effort has been directed towards methods to disperse the particles in water and render them mobile. In one approach, water-soluble starch was used as a stabilizer [118], and in another, hydrophilic carbon or poly(acrylic acid) delivery vehicles were used [119]. Modified cellulose, sodium carboxymethylcellulose (CMC), was found to form highly dispersed nZVI [120] and also several polymers have been tested and found to be very effective [121]. In this stabilized form the nZVI was up to 17 times more reactive in degrading trichloroethene than non-stabilized material. However, for other stabilizing agents a decrease in reactivity of up to 9- [121] or 2–10-fold was observed [121]. To deliver the nZVI to the oil/water interface in the case of DNAPL contamination, a copolymer was used to increase colloid stability and at the same time increase phase transfer into the organic phase [122].

1.6 Environmental Risks

1.6.1 Behavior in the Environment

The use of nanoparticles in environmental applications will inevitably lead to the release of nanoparticles into the environment. Assessing their risks in the environment requires an understanding of their mobility, bioavailability, toxicity and persistence. Whereas air-borne particles and inhalation of nanoparticles have attracted a lot of attention [123], much less is known about the possible exposure of aquatic and terrestrial life to nanoparticles in water and soils [2]. Nanoparticles agglomerate rapidly into larger aggregates or are contained within other materials (e.g. polymers). Cations, for example, are able to coagulate acid-treated CNTs with critical coagulation concentrations of 37 mM for Na, 0.2 mM for Ca and 0.05 mM for trivalent metals (e.g. La^{3+}) [124]. Aggregation of CNTs added as a suspension to filtered pond water has been reported [125]. Sedimentation and therefore removal from water can be expected under such conditions. The coagulation and interception by surfaces also determine the fate of nanoparticles in porous media and rapid removal has been observed in many, but not all, cases [126, 127]. However, a recent study shows that humic and fulvic acids are able to solubilize CNTs under natural conditions and that stable suspensions are obtained [128].

Most nanoparticles in technical applications are functionalized and therefore studies using pristine nanoparticles may not be relevant for assessing the behavior of the actually used particles. As mentioned above in Section 1.5 on groundwater remediation, functionalization is often used to decrease agglomeration and therefore increase mobility of particles. Very little is known to date about the influence of functionalization on the behavior of nanoparticles in the environment.

1.6.2 Ecotoxicology

A consistent body of evidence shows that nanosized particles can be taken up by a wide variety of mammalian cell types, are able to cross the cell membrane and become internalized [6, 7, 129–131]. The uptake of nanoparticles is size dependent [132, 133]. Most of the toxicological studies have been carried out with mammalian cells and therefore were carried out in a cell culture medium containing a mixture of proteins and other biological compounds. In this medium, nanoparticles are coated with proteins and have a negative surface charge irrespective of the charge of the pristine particles [132]. Results from such studies therefore cannot be directly transferred to environmental conditions.

Ecotoxicological studies show that nanoparticles are also toxic to aquatic organisms, both unicellular (e.g. bacteria or protozoa) and animals (e.g. daphnia or fish). Whereas bulk TiO_2 is considered to have no health effects on aquatic organisms, this is clearly not the case for nanosized TiO_2 [134]. This was found both for inorganic nanoparticles such as TiO_2 [134–136], CeO_2 [137] and ZnO [136, 138] and for carbon-containing particles such as fullerenes [139–143] and CNTs [144]. The observed effects ranged from higher activity of certain stress-related genes, lipid peroxidation and glutathione depletion and antibacterial activity (growth inhibition) for microorganisms to increased mortality and reduced fertility at high particle concentrations. Inorganic nanoparticular TiO_2 had a toxic effect on bacteria and the presence of light was a significant factor increasing the toxicity [136]. In copepods purified CNTs did not show any effect whereas unpurified CNTs with all their byproducts increased mortality [144]. Organisms are able to use a lipid coating of CNTs as a food source and therefore alter the solubility and toxicity of the CNT in the organism [145].

Nanosized CeO_2 particles were adsorbed on the cell wall of *E. coli* but the microscopic methods were not sensitive enough to discern whether internalization had taken place [137]. Nanosized ZnO was internalized by bacteria [138]. Nanoparticles that damage bacterial cell walls have been found to be internalized, whereas those without this activity were not taken up [146]. CNTs have been shown to be taken up by a unicellular protozoan [125] and they induced a dose-dependent growth inhibition. The CNTs were localized with the mitochondria of the cells.

These results from ecotoxicological studies show that certain nanoparticles will have effects on organisms on the environment, at least at elevated concentrations. The next step towards an assessment of the risks of nanoparticles in the environment will therefore be to estimate the exposure to the different nanoparticles.

1.7 Conclusions

This chapter was intended to give an overview of the various aspects of nanotechnology and the environment, mainly looking at it from the side of applications rather than from the risk side. It should have become clear that nanotechnology in general

and nanoparticles in particular will have important impacts on various fields of environmental technology and engineering. However, we should always keep in mind that nanotechnology has a Janus face and that each positive and desired property of nanomaterials could be problematic under certain conditions and pose a risk to the environment. A careful weighing up of the opportunities and risks of nanotechnology with respect to their effects on the environment is therefore needed.

References

1 Environmental Protection Agency, *US Environmental Protection Agency Report EPA 100/B-07/001*, EPA Washington DC 2007.

2 M. R. Wiesner, G. V. Lowry, P. Alvarez, D. Dionysiou, P. Biswas, *Environ. Sci. Technol.* 2006, **40**, 4336.

3 V. L. Colvin, *Nat. Biotechnol.* 2003, **21**, 1166.

4 M. Siegrist, A. Wiek, A. Helland, H. Kastenholz, *Nat. Nanotechnol.* 2007, **2**, 67.

5 R. Jones, *Nat. Nanotechnol.* 2007, **2**, 71.

6 G. Oberdörster, E. Oberdörster, J. Oberdörster, *Environ. Health Perspect.* 2005, **113**, 823.

7 A. Nel, T. Xia, L. Mädler, N. Li, *Science* 2006, **311**, 622.

8 T. Masciangioli, W. X. Zhang, *Environ. Sci. Technol.* 2003, **37**, 102A.

9 T. Hillie, M. Munasinghe, M. Hlope, Y. Deraniyagala, *Nanotechnology, water and development*, Meridian Institute, 2006.

10 K. A. D. Guzman, M. R. Taylor, J. F. Banfield, *Environ. Sci. Technol.* 2006, **40**, 1401.

11 M. C. Roco, *Environ. Sci. Technol.* 2005, **39**, 106A.

12 M. A. Albrecht, C. W. Evans, C. L. Raston, *Green Chem.* 2006, **8**, 417.

13 *Estimated Energy savings and Financial impacts of nanomaterials by design on selected applications in the chemical industry*, Los Alamos National Laboratory, Los Alamos, NM, 2006.

14 B. Vogt, *Ind. Diamond Rev.* 2004, **3**, 30.

15 T. Garcia, B. Solsona, S. H. Taylor, *Catal. Lett.* 2005, **105**, 183.

16 N. L. Rosi, J. Eckert, M. Eddaoudi, D. T. Vodak, J. Kim, M. O'Keeffe, O. M. Yaghi, *Science* 2003, **300**, 1127.

17 W. Oelerich, T. Klassen, R. Bormann, *J. Alloys Compd.* 2001, **315**, 237.

18 Y. F. Zhang, J. Q. Shen, *Int. J. Hydrogen Energy* 2007, **32**, 17.

19 S. Schelm, G. B. Smith, *Appl. Phys. Lett.* 2003, **82**, 4346.

20 D. G. Rickerby, M. Morrison, *Sci. Technol. Adv. Mater.* 2007, **8**, 19.

21 J. M. Tarascon, M. Armand, *Nature* 2001, **414**, 359.

22 P. Poizot, S. Laruelle, S. Grugeon, L. Dupont, J. M. Tarascon, *Nature* 2000, **407**, 496.

23 A. Vaseashta, M. Vaclavikova, S. Vaseashta, G. Gallios, P. Roy, O. Pummakarnchana, *Sci. Technol. Adv. Mater.* 2007, **8**, 47.

24 J. S. Taurozzi, V. V. Tarabara, *Environ. Eng. Sci.* 2007, **24**, 122.

25 J. Wang, *Electroanalysis* 2005, **17**, 7.

26 M. Trojanowicz, *Trends Anal. Chem.* 2006, **25**, 480.

27 M. Valcarcel, B. M. Simonet, S. Cardenas, B. Suarez, *Anal. Bioanal. Chem.* 2005, **382**, 1783.

28 A. Merkoci, *Microchim. Acta* 2006, **152**, 157.

29 L. Dai, P. Soundarrajan, T. Kim, *Pure Appl. Chem.* 2002, **74**, 1753.

30 N. Sano, F. Ohtsuki, *J. Electrostat.* 2007, **65**, 263.

31 J. Li, J. E. Koehne, A. M. Cassell, H. Chen, H. T. Ng, Q. Ye, W. Fan, J. Han, M. Meyyappan, *Electroanalysis* 2005, **17**, 15.

32 B. Z. Zeng, F. Huang, *Talanta* 2004, **64**, 380.

33 S. C. Chang, P. Adriaens, *Environ. Eng. Sci.* 2007, **24**, 58.

34 M. F. Hochella, *Geochim. Cosmochim. Acta* 2002, **66**, 735.

35 A. S. Madden, M. F. Hochella, T. P. Luxton, *Geochim. Cosmochim. Acta* 2006, **70**, 4095.

36 D. E. Giammar, C. J. Maus, L. Y. Xie, *Environ. Eng. Sci.* 2007, **24**, 85.

37 S. Pacheco, J. Tapia, M. Medina, R. Rodriguez, *J. Non-Cryst. Solids* 2006, **352**, 5475.

38 E. A. Deliyanni, E. N. Peleka, K. A. Matis, *J. Hazard. Mater.* 2007, **141**, 176.

39 G. D. Yuan, L. H. Wu, *Sci. Technol. Adv. Mater.* 2007, **8**, 60.

40 P. Liang, Q. Ding, F. Song, *J. Sep. Sci.* 2005, **28**, 2339.

41 C. Lu, C. Liu, *J. Chem. Technol. Biotechnol.* 2006, **81**, 1932.

42 C. L. Chen, X. K. Wang, *Ind. Eng. Chem. Res.* 2006, **45**, 9144.

43 Y. H. Li, S. G. Wang, Z. K. Luan, J. Ding, C. L. Xu, D. H. Wu, *Carbon* 2003, **41**, 1057.

44 P. Liang, Y. Liu, L. Guo, J. Zeng, H. B. Lu, *J. Anal. At. Spectrosc.* 2004, **19**, 1489.

45 Y. H. Li, Y. Q. Zhu, Y. M. Zhao, D. H. Wu, Z. K. Luan, *Diamond Relat. Mater.* 2006, **15**, 90.

46 Y. H. Li, S. Wang, J. Wei, X. Zhang, C. Xu, Z. Luan, D. Wu, B. Wei, *Chem. Phys. Lett.* 2002, **357**, 263.

47 J. Munoz, M. Gallego, M. Valcarcel, *Anal. Chem.* 2005, **77**, 5389.

48 X. Feng, G. E. Fryxell, L. Q. Wang, A. Y. Kim, J. Liu, K. M. Kemner, *Science* 1997, **276**, 923.

49 S. O. Obare, G. J. Meyer, *J. Environ. Sci. Health A* 2004, **39**, 2549.

50 S. V. Mattigod, G. E. Fryxell, K. Alford, T. Gilmore, K. Parker, J. Serne, M. Engelhard, *Environ. Sci. Technol.* 2005, **39**, 7306.

51 R. Q. Long, R. T. Yang, *J. Am. Chem. Soc.* 2001, **123**, 2058.

52 K. Yang, L. Zhu, B. Xing, *Environ. Sci. Technol.* 2006, **40**, 1855.

53 K. Yang, X. L. Wang, L. Z. Zhu, B. S. Xing, *Environ. Sci. Technol.* 2006, **40**, 5804.

54 S. Gotovac, Y. Hattori, D. Noguchi, J. Miyamoto, M. Kanamaru, S. Utsumi, H. Kanoh, K. Kaneko, *J. Phys. Chem. B* 2006, **110**, 16219.

55 Q. X. Zhou, J. P. Xiao, W. D. Wang, *J. Chromatogr. A* 2006, **1125**, 152.

56 J. X. Wang, D. Q. Jiang, Z. Y. Gu, X. P. Yan, *J. Chromatogr. A* 2006, **1137**, 8.

57 X. J. Peng, Y. H. Li, Z. K. Luan, Z. C. Di, H. Y. Wang, B. H. Tian, Z. P. Jia, *Chem. Phys. Lett.* 2003, **376**, 154.

58 Y. Q. Cai, Y. E. Cai, S. F. Mou, Y. Q. Lu, *J. Chromatogr. A* 2005, **1081**, 245.

59 C. Lu, Y. L. Chung, K. F. Chang, *J. Hazard. Mater.* 2006, **B138**, 304.

60 C. S. Lu, Y. L. Chung, K. F. Chang, *Water Res.* 2005, **39**, 1183.

61 Y. Q. Cai, G. B. Jiang, J. F. Liu, Q. X. Zhou, *Anal. Chem.* 2003, **75**, 2517.

62 Y. Q. Cai, G. B. Jiang, J. F. Liu, Q. X. Zhou, *Anal. Chim. Acta* 2003, **494**, 149.

63 B. Fugetsu, S. Satoh, T. Shiba, T. Mizutani, Y. B. Lin, N. Terui, Y. Nodasaka, K. Sasa, K. Shimizu, T. Akasaka, M. Shindoh, K. I. Shibata, A. Yokoyama, M. Mori, K. Tanaka, Y. Sato, K. Tohji, S. Tanaka, N. Nishi, F. Watari, *Environ. Sci. Technol.* 2004, **38**, 6890.

64 Q. X. Zhou, Y. J. Ding, J. P. Xiao, *Anal. Bioanal. Chem.* 2006, **385**, 1520.

65 Q. X. Zhou, W. D. Wang, J. P. Xiao, *Anal. Chim. Acta* 2006, **559**, 200.

66 Q. X. Zhou, J. P. Xiao, W. D. Wang, *Microchim. Acta* 2007, **157**, 93.

67 Q. X. Zhou, J. P. Xiao, W. D. Wang, G. G. Liu, Q. Z. Shi, J. H. Wang, *Talanta* 2006, **68**, 1309.

68 M. Biesaga, K. Pyrzynska, *J. Sep. Sci.* 2006, **29**, 2241.

69 K. L. Salipira, B. B. Mamba, R. W. Krause, T. J. Malefetse, S. H. Durbach, *Environ. Chem. Lett.* 2007, **5**, 13.

70 K. Yang, B. Xing, *Environ. Pollut.* 2007, **145**, 529.

71 X. Cheng, A. T. Kan, M. B. Tomson, *J. Chem. Eng. Data* 2004, **49**, 675.

72 D. Heymann, *Fullerene Sci. Technol.* 1996, **4**, 509.

73 X. Cheng, A. T. Kan, M. B. Tomson, *J. Nanopart. Res.* 2005, **7**, 555.

74 E. Ballesteros, M. Gallego, M. Valcarcel, *J. Chromatogr. A* 2000, **869**, 101.

75 Z. C. Di, J. Ding, X. J. Peng, Y. H. Li, Z. K. Luan, J. Liang, *Chemosphere* 2006, **62**, 861.

76 X. J. Peng, Z. K. Luan, J. Ding, Z. H. Di, Y. H. Li, B. H. Tian, *Mater. Lett.* 2005, **59**, 399.

77 Y. H. Li, S. G. Wang, A. Y. Cao, D. Zhao, X. F. Zhang, C. L. Xu, Z. K. Luan, D. B. Ruan, J. Liang, D. H. Wu, B. Q. Wei, *Chem. Phys. Lett.* 2001, **350**, 412.

78 Y. H. Xu, D. Y. Zhao, *Ind. Eng. Chem. Res.* 2006, **45**, 1758.

79 J. Y. Kim, S. B. Shim, J. K. Shim, *J. Ind. Eng. Chem.* 2004, **10**, 1043.

80 A. F. Ngomsik, A. Bee, M. Draye, G. Cote, V. Cabuil, *C. R. Chim.* 2005, **8**, 963.

81 J. Hu, G. H. Chen, I. M. C. Lo, *Water Res.* 2005, **39**, 4528.

82 J. Hu, G. H. Chen, I. M. C. Lo, *J. Environ. Eng.* 2006, **132**, 709.

83 J. Hu, I. M. C. Lo, G. H. Chen, *Langmuir* 2005, **21**, 11173.

84 J. Jin, R. Li, H. L. Wang, H. N. Chen, K. Liang, J. T. Ma, *Chem. Commun.* 2007, 386.

85 N. Hilal, H. Al-Zoubi, N. A. Darwish, A. W. Mohammad, M. Abu Arabi, *Desalination* 2004, **170**, 281.

86 A. Srivastava, O. N. Srivastava, S. Talapatra, R. Vajtai, P. M. Ajayan, *Nat. Mater.* 2004, **3**, 610.

87 M. R. Hoffmann, S. T. Martin, W. Choi, D. W. Bahnemann, *Chem. Rev.* 1995, **95**, 69.

88 Y. Liu, J. Li, X. Qiu, C. Burda, *Water Sci. Technol.* 2006, **54**, 47.

89 K. Nagaveni, G. Sivalingam, M. S. Hedge, G. Madras, *Appl. Catal. B* 2004, **48**, 83.

90 R. Comparelli, P. D. Cozzoli, M. L. Curri, A. Agostiano, G. Mascolo, G. Lovecchio, *Water Sci. Technol.* 2004, **49**, 183.

91 K. Nagaveni, G. Sivalingam, M. S. Hegde, G. Madras, *Environ. Sci. Technol.* 2004, **38**, 1600.

92 Y. S. Chen, J. C. Crittenden, S. Hackney, L. Sutter, D. W. Hand, *Environ. Sci. Technol.* 2005, **39**, 1201.

93 H. M. Zhang, X. Quan, S. Chen, H. M. Zhao, *Environ. Sci. Technol.* 2006, **40**, 6104.

94 W. X. Zhang, *J. Nanopart. Res.* 2003, **5**, 323.

95 K. Sohn, S. W. Kang, S. Ahn, M. Woo, S. K. Yang, *Environ. Sci. Technol.* 2006, **40**, 5514.

96 Y. H. Liou, S. L. Lo, W. H. Kuan, C. J. Lin, S. C. Weng, *Water Res.* 2006, **40**, 2485.

97 J. S. Cao, D. Elliott, W. X. Zhang, *J. Nanopart. Res.* 2005, **7**, 499.

98 K. Mondal, G. Jegadeesan, S. B. Lalvani, *Ind. Eng. Chem. Res.* 2004, **43**, 4922.

99 S. R. Kanel, J. M. Greneche, H. Choi, *Environ. Sci. Technol.* 2006, **40**, 2045.

100 G. Jegadeesan, K. Mondal, S. B. Lalvani, *Environ. Prog.* 2005, **24**, 289.

101 S. R. Kanel, B. Manning, L. Charlet, H. Choi, *Environ. Sci. Technol.* 2005, **39**, 1291.

102 S. M. Ponder, J. G. Darab, T. E. Mallouk, *Environ. Sci. Technol.* 2000, **34**, 2564.

103 B. A. Manning, J. R. Kiser, H. Kwon, S. R. Kanel, *Environ. Sci. Technol.* 2007, **41**, 586.

104 X. Q. Li, W. X. Zhang, *Langmuir* 2006, **22**, 4638.

105 X. Q. Li, D. W. Elliott, W. X. Zhang, *Crit. Rev. Solid State Mater. Sci.* 2006, **31**, 111.

106 H. L. Lien, Y. S. Jhuo, L. H. Chen, *Environ. Eng. Sci.* 2007, **24**, 21.

107 P. G. Tratnyek, R. L. Johnson, *Nanotoday* 2006, **1**, 44.

108 W. X. Zhang, C. B. Wang, H. L. Lien, *Catal Today* 1998, **40**, 387.

109 H. L. Lien, W. X. Zhang, *J. Environ. Eng.* 2005, **131**, 4.

110 B. Schrick, J. L. Blough, A. D. Jones, T. E. Mallouk, *Chem. Mater.* 2002, **14**, 5140.

111 T. T. Lim, J. Feng, B. W. Zhu, *Water Res.* 2007, **41**, 875.

112 D. W. Elliott, W. X. Zhang, *Environ. Sci. Technol.* 2001, **35**, 4922.

113 J. Quinn, C. Geiger, C. Clausen, K. Brooks, C. Coon, S. O'Hara, T. Krug, D. Major, W. S. Yoon, A. Gavaskar,

T. Holdsworth, *Environ. Sci. Technol.* 2005, **39**, 1309.

114 M. C. Chang, H. Y. Shu, W. P. Hsieh, M. C. Wang, *J. Air Waste Manage. Assoc.* 2005, **55**, 1200.

115 M. C. Chang, H. Y. Shu, W. P. Hsieh, M. C. Wang, *J. Air Waste Manage. Assoc.* 2007, **57**, 221.

116 P. Varanasi, A. Fullana, S. Sidhu, *Chemosphere* 2007, **66**, 1031.

117 J. S. Cao, W. X. Zhang, *J. Hazard. Mater. B* 2006, **132**, 213.

118 F. He, D. Y. Zhao, *Environ. Sci. Technol.* 2005, **39**, 3314.

119 B. Schrick, B. W. Hydutsky, J. L. Blough, T. E. Mallouk, *Chem. Mater.* 2004, **16**, 2187.

120 F. He, D. Y. Zhao, J. C. Liu, C. B. Roberts, *Ind. Eng. Chem. Res.* 2007, **46**, 29.

121 N. Saleh, K. Sirk, Y. Q. Liu, T. Phenrat, B. Dufour, K. Matyjaszewski, R. D. Tilton, G. V. Lowry, *Environ. Eng. Sci.* 2007, **24**, 45.

122 N. Saleh, T. Phenrat, K. Sirk, B. Dufour, J. Ok, T. Sarbu, K. Matyiaszewski, R. D. Tilton, G. V. Lowry, *Nano Lett.* 2005, **5**, 2489.

123 P. Biswas, C. Y. Wu, *J. Air. Waste Manage. Assoc.* 2005, **55**, 708.

124 M. Sano, J. Okamura, S. Shinkai, *Langmuir* 2001, **17**, 7172.

125 Y. Zhu, Q. Zhao, Y. Li, X. Cai, W. Li, *J. Nanosci. Nanotechnol.* 2006, **6**, 1357.

126 X. Cheng, A. T. Kan, M. B. Tomson, *J. Mater. Res.* 2005, **20**, 3244.

127 K. A. Dunphy Guzman, D. L. Finnegan, J. F. Banfield, *Environ. Sci. Technol.* 2006, **40**, 7688.

128 H. Hyung, J. D. Fortner, J. B. Hughes, J. H. Kim, *Environ. Sci. Technol.* 2007, **4**, 179.

129 S. K. Smart, A. I. Cassady, G. Q. Lu, D. J. Martin, *Carbon* 2006, **44**, 1034.

130 I. Lynch, K. A. Dawson, S. Linse, *Sci. STKE* 2006, pe14.

131 B. M. Rothen-Rutishauser, S. Schürch, B. Haenni, N. Kapp, P. Gehr, *Environ. Sci. Technol.* 2006, **40**, 4353.

132 L. K. Limbach, Y. Li, R. N. Grass, T. J. Brunner, M. A. Hintermann, M. Muller, D. Gunther, W. J. Stark, *Environ. Sci. Technol.* 2005, **39**, 9370.

133 B. D. Chithrani, A. A. Ghazani, W. C. W. Chan, *Nano Lett.* 2006, **6**, 662.

134 S. B. Lovern, R. Klaper, *Environ. Toxicol. Chem.* 2006, **25**, 1132.

135 K. Hund-Rinke, M. Simon, *Environ. Sci. Pollut. Res.* 2006, **13**, 225.

136 L. K. Adams, D. Y. Lyon, P. J. J. Alvarez, *Water Res.* 2006, **40**, 3527.

137 A. Thill, O. Zeyons, O. Spalla, F. Chauvat, J. Rose, M. Auffan, A. M. Flank, *Environ. Sci. Technol.* 2006, in press.

138 R. Brayner, R. Ferrari-Illiou, N. Brivois, S. Djediat, M. F. Benedetti, F. Fiévet, *Nano Lett.* 2006, **6**, 866.

139 D. Y. Lyon, J. D. Fortner, C. M. Sayes, V. L. Colvin, J. B. Hughes, *Environ. Toxicol. Chem.* 2005, **24**, 2757.

140 E. Oberdorster, *Environ. Health Perspect.* 2004, **112**, 1058.

141 S. Q. Zhu, E. Oberdorster, M. L. Haasch, *Mar. Environ. Res.* 2006, **62**, S5.

142 D. Y. Lyon, L. K. Adams, J. C. Falkner, P. J. J. Alvarez, *Environ. Sci. Technol.* 2006, **40**, 4360.

143 E. Oberdörster, S. Zhu, T. M. Blickley, P. McClellan-Green, M. L. Haasch, *Carbon* 2006, **44**, 1112.

144 R. C. Templeton, P. L. Ferguson, K. M. Washburn, W. A. Scrivens, G. T. Chandler, *Environ. Sci. Technol.* 2006, **40**, 7387.

145 A. P. Roberts, A. S. Mount, B. Seda, J. Souther, R. Qiao, S. Lin, P. C. Ke, A. M. Rao, S. J. Klaine, *Environ. Sci. Technol.* 2007, **41**, 3025.

146 P. K. Stoimenov, R. L. Klinger, G. L. Marchin, K. J. Klabunde, *Langmuir* 2002, **18**, 6679.

2
Photocatalytic Surfaces: Antipollution and Antimicrobial Effects

Norman S. Allen, Michele Edge, Joanne Verran, John Stratton, Julie Maltby, and Claire Bygott

2.1
Introduction to Photocatalysis: Titanium Dioxide Chemistry and Structure–Activity

For many years, titanium dioxide pigments have been used successfully for conferring opacity and whiteness to a host of different materials. Their principal usage is in applications such as paints, plastics, inks and paper, but they are also incorporated into a diverse range of products, such as foods and pharmaceuticals. The fundamental properties of titanium dioxide have given rise to its supreme position in the field of white pigments. In particular, its high refractive index permits the efficient scattering of light. Its absorption of UV light has conferred durability on products. Its non-toxic nature has meant that it can be widely used in almost any application without risk to health and safety. However, the primary reason for its success is the ability to reflect and refract or scatter light more efficiently than any other pigment, due to its high refractive index in comparison with extenders, fillers and early pigments [1–5] (see Table 2.1).

Titanium dioxide exists in three crystalline modifications, rutile, brookite and anatase, all of which have been prepared synthetically. In each type, the titanium ion coordinates with six oxygen atoms, which in turn are linked to three titanium atoms and so on. Anatase (Figure 2.1) and rutile (Figure 2.2) are tetragonal whereas brookite is orthorhombic. Brookite and anatase are unstable forms. Brookite is not economically significant since there is no abundant supply in nature.

Examination (Table 2.2) of the basic properties of the two main crystal forms shows differences in specific gravity, hardness, refractive index and relative tint strength. The oil absorption of commercial anatase and rutile pigments also varies, in part due to the different types of surface treatments applied to them.

Titanium dioxide has the highest average refractive index known. For anatase, it is 2.55 and for rutile it is 2.76. These high values account for the exceptional light scattering ability of pigmentary titanium dioxide when dispersed in various media, which in turns yields the high reflectance and hiding power associated with this pigment. Although single-crystal titanium dioxide is transparent, as a finely divided

Nanotechnology. Volume 2: Environmental Aspects. Edited by Harald Krug

ISBN: 978-3-527-31735-6

Table 2.1 Refractive index of TiO_2 in comparison with extenders, filler and other pigments.

Material	Refractive index
Rutile TiO_2	2.76
Anatase TiO_2	2.52
Lithopone	2.13
Zinc oxide	2.02
White lead	2.00
Calcium carbonate	1.57
China clay	1.56
Talc	1.50
Silica	1.48

powder it has a very high reflectance and it is intensely white because its high reflectance is substantially uniform throughout the visible spectrum. This white color is different in tone for the two crystal structures due to their different reflectance curves across the visible and near-visible spectrum (Figure 2.3).

From examination of Figure 2.3, is evident that:

- Rutile TiO_2 reflects the radiation slightly better that anatase and is therefore brighter.
- The higher absorption of rutile at the very blue end of the visible spectrum and in the UV region accounts for the yellower tone of rutile pigment. This higher UV

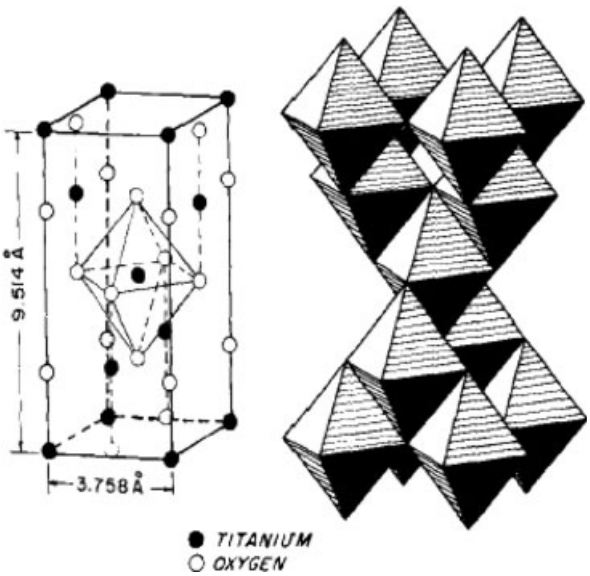


Figure 2.1 Crystal structure of anatase.

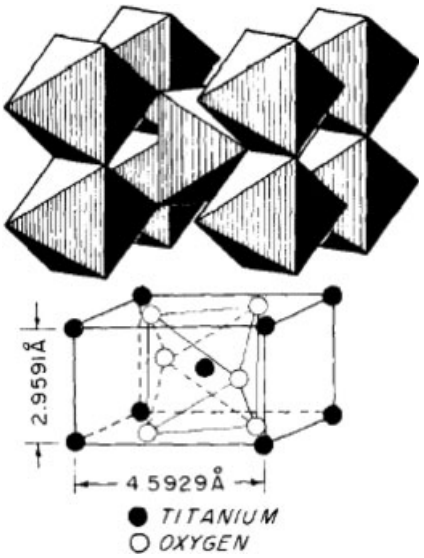


Figure 2.2 Crystal structure of rutile.

absorption also provides relatively better durability to a system as it reduces the amount of energy available to degrade the binder.

Although the difference in refractive index gives rutile pigments up to 15% opacity benefit over the anatase pigment, the bluer tone and lower hardness of anatase pigments are beneficial in some applications, especially where low abrasivity may be an issue. Where the highest possible optical efficiency (opacity) and durability are required, rutile pigments are superior.

Refractive index varies with wavelength and for titanium dioxide the refractive index is greater for shorter wavelengths (blue region) and lower for longer wavelengths (red region) of the visible spectrum.

Table 2.2 Typical properties of anatase and rutile polymorphs.

Property	Anatase	Rutile
Pigment form		
Appearance	Brilliant white powders	
Density ($g\,cm^{-3}$)	3.8–4.1	3.9–4.2
Refractive index	2.55	2.76
Oil absorption (1b/100b)	18–30	16–48
Tinting strength (Reynolds)	1200–1300	1650–1900
Crystal form		
Density ($g\,cm^{-3}$)	3.87	4.24
Hardness (Moh)	5–6	6–7

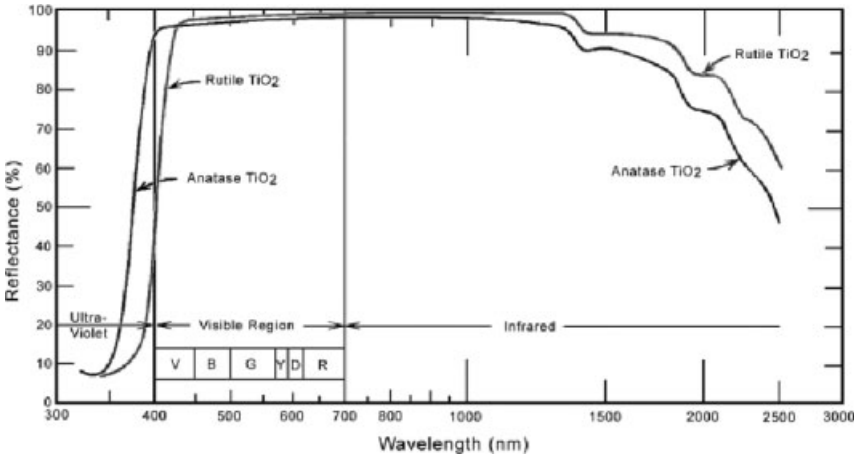


Figure 2.3 Reflectance of anatase and rutile pigments through the near-UV, visible and IR regions.

The performance of a pigment in a surface coating is significantly affected by the interaction of the medium with the pigment surface. The consequences are felt at all stages, but are particularly relevant for dispersion, shelf stability and exterior durability. Treated TiO_2 absorbs UV radiation and protects the polymer photochemically; untreated TiO_2, however, is itself photocatalytic. Although it converts most of the UV energy into heat, the remaining energy creates radicals, which accelerate the breakdown of the polymer. Almost all titanium dioxide used in plastics applications is surface treated. Treatments are essentially the same, whether the base pigment is produced by the chloride or the sulfate route. The level of photocatalytic activity may be reduced by surface treatment of the base pigment with suitable inorganic compounds [1]. The most common precipitates are oxyhydrates of aluminum and silicon. Also used are oxides and oxyhydrates of zirconium, tin, zinc, cerium and boron. The treatment functions by placing a physical barrier between the pigment surface and the polymer matrix, blocking the active sites and minimizing degradation. The treatment may also aid and reduce the requirements of power and shear when mixing. Many TiO_2 pigments also have a final organic treatment, such as with trimethylolpropane or pentaerythritol. Its primary function is to modify the interfacial region between the hydrated inorganic oxide TiO_2 particle surface and various less polar organic polymers.

During the manufacturing process of TiO_2, the pigment is formed as discrete particles of around 0.2–0.4 μm. The titanium dioxide manufacturers control the operational variables to produce particles of a uniform size and distribution. These 0.2–0.4 μm particles have been engineered to maximize the scattering of light, resulting in optimum brightness and opacity.

However, as soon as the particles are manufactured, they begin to combine into aggregates, agglomerates and flocs. *Aggregates* are associations of pigment particles that are fixed together along the crystal faces. Bonds between particles are strong and cannot be broken by conventional grinding devices. *Agglomerates* are associations of pigment particles and aggregates that are weakly bonded together. *Flocs* are

associations of crystallites, aggregates and agglomerates joined across corners or held together by short range attractive forces. These flocs disperse under moderate shear (Figure 2.4).

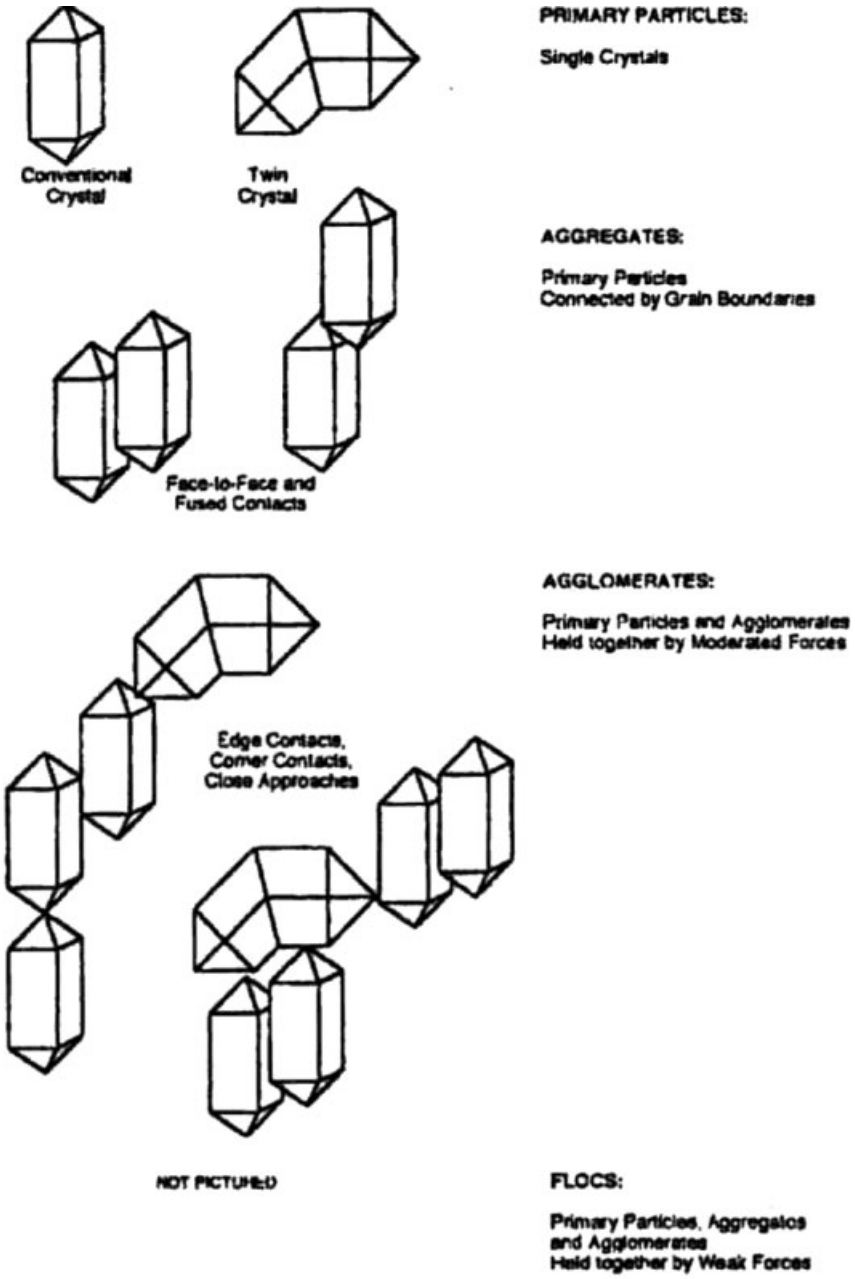


Figure 2.4 Aggregates and agglomerates.

Aggregates can only be broken into individual pigment particles with intensive milling. One of the last manufacturing steps performed by the TiO_2 manufacturer is micronization and/or milling to dissociate as many aggregates as possible. Aggregates will not reform unless the pigment is heated to over 500 °C. Agglomerates are also broken up in the milling step. However, agglomerates will easily re-form during packing, storage and transportation. The disruption of these inter-particle bonds is generally understood to be the dispersion that needs to be performed by the TiO_2 consumer.

It is possible to manipulate the TiO_2 particle size to within a very narrow range around a predetermined optimum. Generally, in paint applications this optimum is approximately 0.2–0.3 μm, as it is within this range that TiO_2's light scattering ability is at its peak, which in turn maximizes the level of gloss finish.

TiO_2 pigment particles are submicroscopic with size distributions narrower than many so-called monodisperse particulates. Appropriately ground, pigment dispersions contain less than 5 wt.% of particles smaller than 0.10 μm and larger than 1.0 μm. Optical effectiveness, that is, light scattering is controlled by the mass/volume frequency of particles in the size range from 0.1 to 0.5 μm. Gloss is diminished by a relatively small mass/volume fraction of particle larger than about 0.5 μm. Dispersibility and film fineness is degraded by a very small mass/volume fraction of particles larger than about 5 μm. Important optical properties such as opacity, hiding power, brightness, tone, tinting strength and gloss are all dependent upon the particle size and particle size distribution.

Pure titanium dioxide possesses by nature an internal crystal structure that yields an innately high refractive index. When the particle size and particle size distribution are to be optimized so as to contribute along with its high refractive index to a maximum light scattering, *conventional or pigmentary* titanium dioxide is obtained. It reflects all the wavelength of the visible light to the same degree, producing the effect of whiteness to the human eye. All these attributes, together with its opacity, are achieved for an optimal particle diameter which is approximately 0.2–0.4 μm, that is, in the order of half the wavelength of visible light. This fact can also be demonstrated on the basics of Mie theory [6].

There exists, however, another type of titanium dioxide whose median crystal size has been explicitly reduced up to 0.02 μm. This is the so-called *nanoparticles or ultrafine* TiO_2 and will be the subject of this chapter.

The history of nanoparticle titanium dioxide dates back to the late 1970s when the first patent on the preparation of these materials was issued in Japan. It is in principle possible to obtain nanoparticle TiO_2 by simple milling of the pigmentary TiO_2 to a finer particle [4]. However, the properties of the fine powders in terms of purity, particle size distribution and particle shape remain highly unsatisfactory.

Several wet-chemical processes were developed during the 1980s by TiO_2 pigment manufacturers such as Ishihara, Tioxide and Kemira. The first part of the process, the production of the nanoparticle base material, uses after-wash titanium hydroxylate as the raw material. After subsequent process steps involving the decomposition of the hydroxylate crystal structure and the reprecipitation of the TiO_2, the product is calcined to obtain oval-shaped particles with a desired primary crystal size and narrow

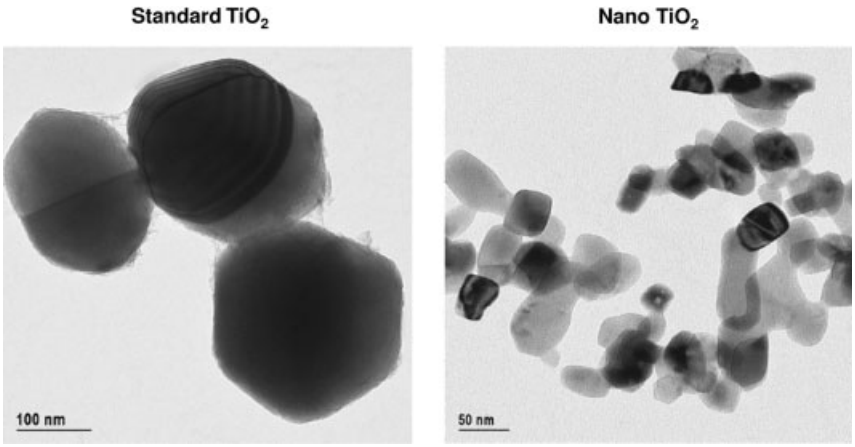


Figure 2.5 Typical TEM images of pigmentary and nanoparticles.

size distribution. The base crystals are coated in the after-treatment unit according to the requirements of the end-use. One of the primary tasks of the after-treatment is to ensure good dispersibility of extremely fine particles in the final application.

TiO_2 nanoparticles are also routinely produced by the gas-to-particle conversion in flame reactors because this method provides good control of particle size, particle crystal structure and purity [4].

Typically, the crystal size of these products is about one-tenth of the size of the normal pigmentary grade. Figure 2.5 shows typical transmission electron micrographs for pigmentary and nanoparticulate titanium dioxide at the same magnification.

Table 2.3 shows a comparison of some typical values of the physical properties of nanoparticle and conventional titanium dioxide products.

The smaller crystal size influences various properties and leads to higher values for the surface area and oil absorption. Lower values for specific gravity and bulk density are also achieved. Otherwise, it has many of the properties of conventional TiO_2 pigments: non-toxic, non-migratory, inert and stable at high temperatures.

The optical behavior of ultrafine TiO_2 differs dramatically from that of conventional TiO_2 pigment. The optical properties of nanoparticle TiO_2 are governed by the

Table 2.3 Typical properties of nanoparticle and conventional titanium dioxide.

Property	Nanoparticle	Pigmentary
Appearance	White powder	White powder
Crystal structure	Anatase or rutile	Anatase or rutile
Crystal size (μm)	0.005–0.05	0.15–0.3
Specific surface area ($m^2\,g^{-1}$)	50–>300	15
Bulk density ($g\,mL^{-1}$)	3.3	4.0
Oil absorption (g per 100 g)	30	16

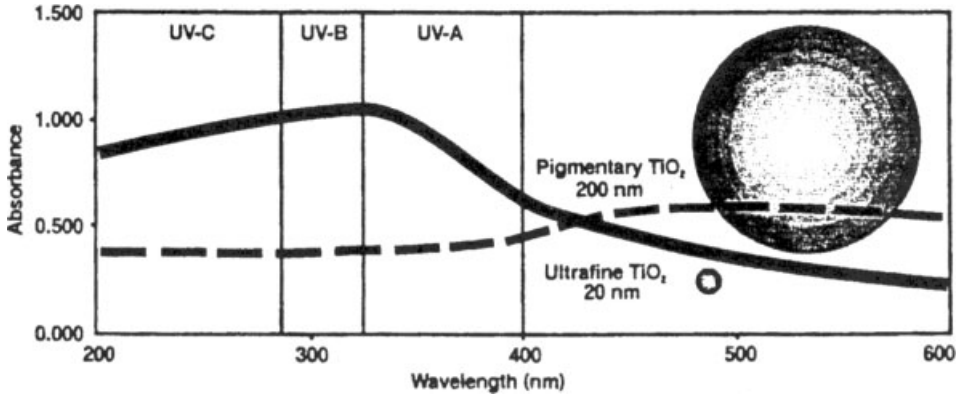


Figure 2.6 Comparison of the optical behavior of ultrafine TiO_2 and pigmentary TiO_2.

Rayleigh theory of light scattering. A simple interpretation of this theory is that the shorter wavelengths of light are more efficiently scattered by very small particles. The intensity of the scattered light is inversely proportional to the fourth power of the wavelength. In practical terms, a reduction in the crystal size of a TiO_2 product leads to an optimum size of TiO_2 of the order of 20–50 nm where the UV spectrum of light (200–400 nm) is effectively scattered from the particles while the visible wavelengths are transmitted through the material. The material thus appears virtually transparent to the naked eye. The behavior is demonstrated in Figure 2.6, which shows the difference between pigmentary and nanoparticle TiO_2.

The complete picture of the optical behavior of TiO_2 becomes more complete by recognizing that TiO_2 is a semiconductor. TiO_2 exhibits a characteristic energy gap of 3.23 or 3.06 eV between the valence band and the conduction band for anatase and rutile, respectively. Wavelengths shorter than 390 nm for anatase and 405 nm for rutile – corresponding to higher energy than the threshold energy – will excite electrons from the valence to the conduction band. Summarizing, titanium dioxide exhibits various mechanisms on exposure to light depending on the wavelength and the particle size: (see Table 2.4 and Figure 2.7). Electron–hole pairs are formed, giving rise to various sensitization processes.

Based on the light scattering property described earlier, nanoparticle titanium dioxide can be used to impart excellent UV protection. Compared with the available

Table 2.4 Optical behavior of pigmentary and nanoparticle TiO_2 under visible and UV light.

Particle size	Wavelength $<$400 nm	Wavelength $>$400 nm
Pigmentary TiO_2	Semiconductor absorption	Scattering and reflection (Mie scattering)
Nanoparticle TiO_2	Semiconductor absorption Scattering and reflection (Raleigh's theory)	Transmission of light Particle diameter $\ll$ wavelength

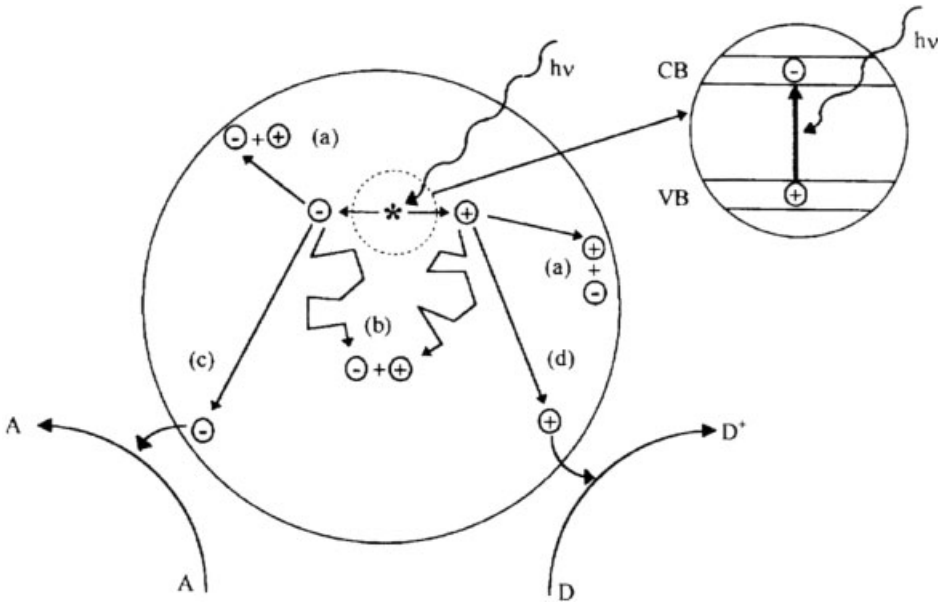


Figure 2.7 Illustration of the major processes occurring on a semiconductor particle following electronic excitation. Electron –hole recombination can occur at the surface [reaction (a)] or in the bulk [reaction (b)] of the semiconductor. At the surface of the particle, photogenerated electrons can reduce an electron acceptor A [reaction (c)] and photogenerated holes can oxidize an electro donor D [reaction (d)]. The combination of reactions (c) and (d) represent the semiconductor sensitization of general redox reactions given later in the text.

UV absorbers, ultrafine TiO_2 possesses effective UV filter properties over the entire UV spectrum (UVC + UVB + AVA). For example, it is gaining wide acceptance for use in sun creams. Nanoparticle TiO_2, apart from its effective attenuating characteristics, is extremely inert and, therefore, safe to use next to the skin [5]. Nanoparticle TiO_2 can also be used in clear plastic films to provide UV protection to foodstuffs. UV radiation from artificial lighting in a grocery store induces auto-oxidation in, e.g., meat and cheese, resulting in discoloration. In this regard it also exhibits antibacterial behavior, which will be discussed later.

It is also possible to use ultrafine TiO_2 as a light stabilizer in plastics to protect the material itself from yellowing and to retard the deterioration of the mechanical properties. A further example of the potential of nanoparticle TiO_2 as a UV filter is found in clear wood finishes. The original color of wood panels can be retained by a clear lacquer made with 0.5–4% nanoparticle TiO_2 [7]. In addition to preventing wood from darkening, ultrafine TiO_2 also enhances its lifetime.

An exciting and increasing popular application of the optical properties of ultrafine TiO_2 is found in automotive coatings, where the ultrafine powder is used as an effect pigment in combination with mica flakes to create the so-called titanium opalescent effect. For UV protection applications and due to the intrinsic photoactivity of TiO_2 pigments, mainly, *nanoparticle surface-treated rutile* pigments are used. *Nanoparticle anatase TiO_2* finds applications in the field of photocatalysis.

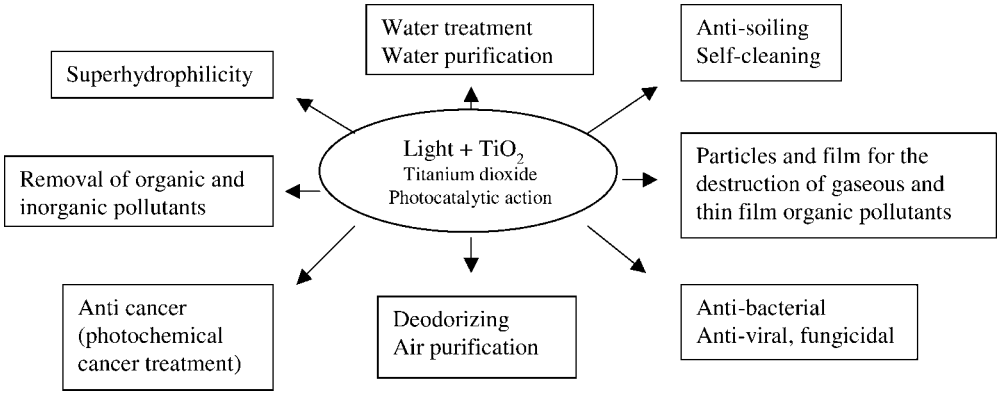


Figure 2.8 Major areas of activity in titanium dioxide photocatalysis.

2.2 Applications

The field of heterogeneous photocatalysis is very diverse and involves many research groups throughout the world. A number of research themes have emerged which offer real potential for commercial developments and merit much greater research.

Figure 2.8 displays and summarizes many of the major fields in which TiO_2 is used as a photocatalyst, all related in a sense towards solving environmental issues.

For the purpose of this chapter, we will just focus on our application studies of TiO_2 for antibacterial, self-cleaning and depollution [atmospheric contaminants such as volatile organic compounds (VOCs) and nitrogen oxides]. For example, in indoor environments, most surfaces, e.g. ceramic tiles, window glass or paper, are gradually covered with organic matter such as oils, dirt and smoke residue and become fouled [8]. Transparent TiO_2 coatings can be completely unobtrusive, causing no readily discernible changes in the substrate color or transparency, but they can decompose organic matter as it deposits. Thus, various types of surfaces with TiO_2 can be covered to make them self-cleaning under sunlight as well as room light (Figure 2.9). Thus, surfaces based on paints, ceramics, glass and cementitious materials containing active photocatalytic titania nanoparticles have widespread applications to create environmentally clean areas within their proximity.

2.3 Photocatalytic Chemistry

The overall catalytic performance of titanium dioxide particles has been found to be dependent on a number of parameters, including preparation method, annealing temperature, particle/crystal size, specific surface area, ratio between the anatase and rutile crystal phases, light intensity and the substrate to be degraded [9]. Furthermore, the electrons confined in the nanomaterial exhibit a different behavior to that in the

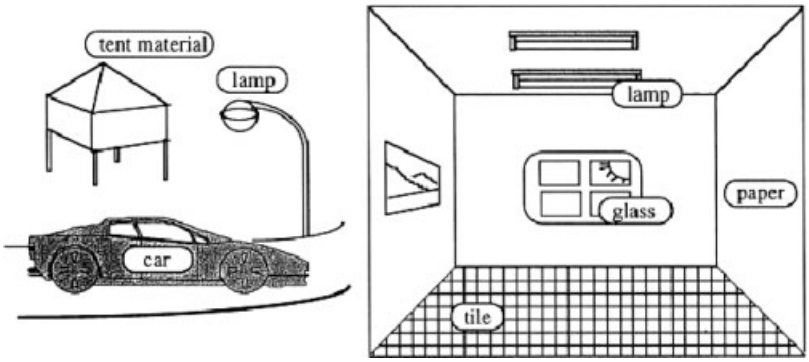


Figure 2.9 Schematic representation of the possible applications of transparent TiO_2 thin film photocatalyst in the indoor and outdoor environments.

bulk materials. The properties of the electrons in small semiconductors should be dependent on the crystallite size and the shape due to quantized motion of the electron and hole in a confined space. This phenomenon is called the quantum size effect. However, the quantization effect does not exist in amorphous phases [10]. As a result of the confinement, the bandgap increases and the band edges shift to yield larger redox potentials. Hence the use of size-quantized semiconductor TiO_2 particles may result in increased photoefficiencies [11].

However, other workers [12] reported that the photocatalytic activity increased greatly and the blue shift was significant only at particle diameters less than 10 nm. On the other hand, the small size effect can improve the photocatalytic activity of the TiO_2 due to the increasing specific surface area, which gives more reactive sites to absorb pollutants. Meanwhile, the diffusion of the photoinduced electrons or holes from bulk to surface becomes fast with a decrease in the particle size [13], which will also lead to an enhancement of the photocatalytic activity. On the other hand, the surface tension increases and causes a crystal lattice distortion with decreasing particle size [14] and consequent change in the structure of the energy band.

The anatase TiO_2 phase is more active than the rutile phase in photocatalysis. The reason for the lower photocatalytic efficiencies in the rutile TiO_2 phase is because the recombination of the electron-hole pair produced by UV irradiation occurs more rapidly on the surface of the rutile phase and the amounts of reactants and hydroxides attached to the surface of the rutile phase are smaller than those of the anatase TiO_2 phase [17]. However, according to other work, the decrease in the photocatalytic effect during the transformation from the anatase to rutile TiO_2 phase was not due to the change in the crystalline structure, but mainly to changes in the specific surface area and porosity [16]. The photocatalytic processes on a titanium dioxide particle are displayed simply in Figure 2.10. Primarily following photoexcitation, a number of surface processes can take place providing activation and further reactions depending on the nature of the environment in question. Holes can generate active hydroxyl radicals whereas active oxygen species are generated through electron transfer processes. All exhibit high activity that can react with surrounding organic and gaseous environments.

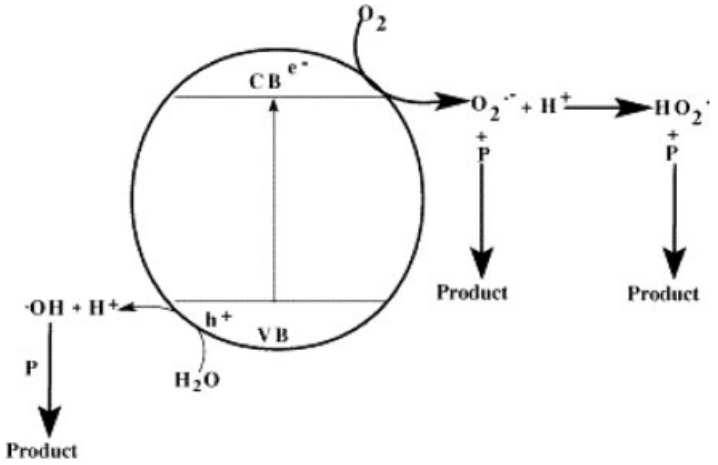


Figure 2.10 Surface photocatalytic activity of titanium dioxide.

Polymeric and organic coatings systems are commonly utilized by titania doping for various applications. In outdoor applications all polymers degrade. The degradation rate depends on the environment (especially sunlight intensity, temperature and humidity) and on the type of polymer. This so-called photo-oxidative degradation is due to combined effects of photolysis and oxidative reactions. Sunlight photolytic degradation and/or photo-oxidation can only occur when the polymer contains chromophores which absorb wavelengths of the sunlight spectrum on Earth (>290 nm). These wavelengths have sufficient energy to cause a dissociative (cleavage) processes resulting in degradation.

Chromophores that can absorb sunlight are:

- internal in-chain impurities such as hydroperoxides or carbonyls formed during storage, processing or weathering;
- external impurities such as polymerization catalyst residues, additives (e.g. pigments, dyes or antioxidants), pollutants from the atmosphere or metal traces from processing equipment;
- parts of the molecular structure of the polymer, i.e. polyaromatics;
- charge-transfer complexes between oxygen and the polymer chain.

Photo-oxidative degradation is due to a radical-based auto-oxidative process (Figure 2.11), which can be divided into four stages.

In the initiation step. free radicals are generated. During photo-oxidation these radicals are formed due to a photolysis reaction of one of the chromophores present. The propagation reactions are thermal reactions and these have been studied in more detail. The rate of the reaction of oxygen with alkyl radicals is very high and that is why the rate of the propagation is largely determined by the ease of hydrogen atom abstraction in the second step of the propagation. The propagation reaction is a repeating reaction; photochemically, hydroperoxides can decompose

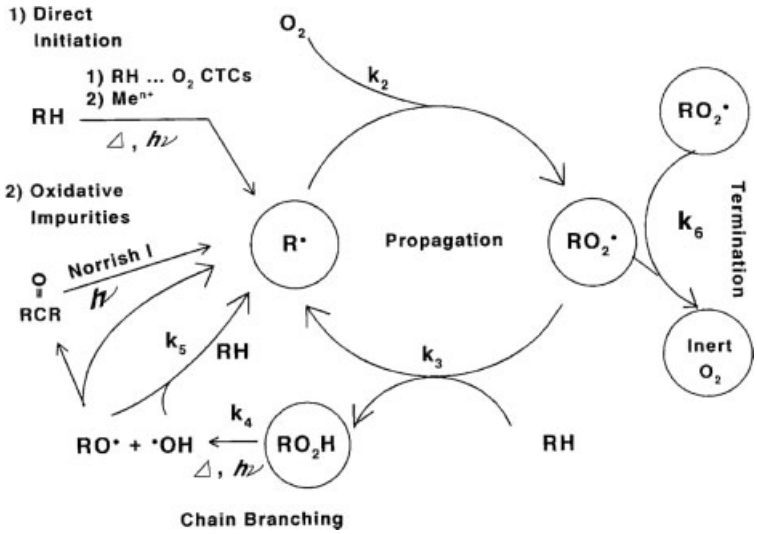


Figure 2.11 Auto-oxidation mechanism for almost all polymers (R = polymer chain, H = most labile hydrogen, $X^{\bullet}$ = any radical, k_i = reaction rate).

homolytically into alkoxy and hydroxy radicals, which can initiate another propagation cycle [18, 19]. Termination reactions are bimolecular. In the presence of sufficient air, which is normally the case for the long-term degradation of polymers, only the reaction of two peroxy radicals has to be considered. Here the reaction depends on the type of peroxy radical present. Aside from these processes, polyaromatics and heterochain polymers exhibit further complex reactions but for the purposes of this chapter the main processes of concern are those induced by the catalytic effect of the titanium dioxide.

The ability of pigments to catalyze the photo-oxidation of polymer systems has also attracted significant attention in terms of their mechanistic behavior. In this regard, much of the information originates from work carried out on TiO_2 pigments in both polymers and model systems [20–27]. To date their are three current mechanisms of the photosensitized oxidation of polymers by TiO_2 and, for that matter, other white pigments such as ZnO:

1. The formation of an oxygen radical anion by electron transfer from photoexcited TiO_2 to molecular oxygen [20]. A recent modification of this scheme involves a process of ion annihilation to form singlet oxygen, which then attacks any unsaturation in the polymer [28].

$$TiO_2 + O_2 \xrightarrow{h\nu} {TiO_2}^{+\bullet} + {O_2}^{-\bullet}(I)$$

$$(I) \rightarrow TiO_2 + {}^1O_2(\text{ion annihilation})$$

$$(I) + H_2O \rightarrow TiO_2 + HO^{\bullet} + HO_2^{\bullet}$$

$$2HO_2^{\bullet} \rightarrow H_2O_2 + O_2$$

$$RCH_2 = CHR' + {}^1O_2 \rightarrow RCH = CHCH(OOH)R'$$

2. Formation of reactive hydroxyl radicals by electron transfer from water catalyzed by photoexcited TiO_2 [29]. The Ti^{3+} ions are reoxidized back to Ti^{4+} ions to start the cycle over again.

$$H_2O \xrightarrow{(TiO_2)h\nu} H^+ + e'(Aqu) + rOH$$

$$[Ti^{4+}] + e' \rightarrow [Ti^{3+}]$$

$$[Ti^{3+}] + O_2 \rightarrow [Ti^{4+}]$$

3. Irradiation of TiO_2 creates an exciton (p) which reacts with the surface hydroxyl groups to form a hydroxyl radical [20]. Oxygen anions are also produced, which are adsorbed on the surface of the pigment particle. They produce active perhydroxyl radicals.

$$TiO_2 \xrightarrow{h\nu} e' + (p)$$

$$OH^{\bullet} + (p) \rightarrow HO^{\bullet}$$

$$Ti^{4+} + e' \rightarrow Ti^{3+}$$

$$Ti^{3+} + O_2 \rightarrow [Ti^{4+} \cdots O^{2-}]_{adsorbed}$$

$$[Ti^{4+} \cdots O^{2-}]_{adsorbed} + H_2O \rightarrow Ti^{4+} + HO^{\bullet} + HO_2^{\bullet}$$

As mentioned previously, titanium dioxide particles are often coated with various media. For example, to improve pigment dispersion and reduce photoactivity, the surface of the pigment particles is coated with precipitated aluminosilicates. Zirconates are also used in some instances whereas for other applications such as in nylon polymers and fibers the anatase is coated with manganese silicates or phosphates. Anatase will photosensitize the oxidation of a polymer, the effect being dependent on the nature and density of the coating and increasing with pigment concentration. Uncoated rutiles are also photosensitizers but again the effect is reduced and proportional to the effectiveness of the coating. In this case, stabilization increases with increasing coated rutile concentration. Thus, the surface characteristics of the titania pigment are an important factor in controlling photoactivity. As discussed for Figure 2.11, the surface is covered with hydroxyl groups of an amphoteric character formed by the adsorption of water. These groups are more acidic in character on the surface of anatase and less effectively bound than those on rutile. The surface carriers

(excitons) therefore react more slowly with the hydroxyl groups in the case of rutile. Infrared analysis has been used to characterize the different species on the particle surfaces. At 3000–3700 cm^{-1} free and hydrogen-bonded OH groups can be detected whereas in the region 1200–1700 cm^{-1} H−O−H bending and carbonates can be seen.

Surface modifications of the TiO_2 particles with inorganic hydrates may reduce the photochemical reactivity of titanium pigments. This can reduce the generation of free radicals by physically inhibiting the diffusion of oxygen and preventing release of free radicals. The often simultaneous chemical effects of surface modification can involve provision of hole and electron recombination sites or hydroxyl radical recombination sites. In addition to the latter effects, the surface treatment or coating, as mentioned above, can improve other properties such as wetting and dispersion in different media (water, solvent or polymer), to improve compatibility with the binder and dispersion stability and color stability. The photosensitivity of titanium dioxide is considered to arise from localized sites on the crystal surface and occupation of these sites by surface treatments inhibits photo-reduction of the pigment by UV radiation and hence the destructive oxidation of the binder is inhibited. Coatings containing 2–5 wt.% of alumina or alumina and silica are satisfactory for general-purpose paints. If greater resistance to weathering is desired, the pigments are coated more heavily to about 7–10 wt.%. The coating can consist of a combination of several materials, e.g. alumina, silica, zirconia, aluminum phosphates or other metals. For example, the presence of hydrous alumina particles lowers van der Waals forces between pigment particles by several orders of magnitude, decreasing particle–particle attractions. Hydrous aluminum oxide phases appear to improve dispersibility more effectively than most of the other hydroxides and oxides. Coated and surface-treated nanoparticles of titania also have commercial uses in, for example, enhanced stabilization of polymers and coatings [7].

During the weathering of commercial polymers containing white pigments such as titania, oxidation occurs at the surface layers of the material, which eventually erode away, leaving the pigment particles exposed. This phenomenon is commonly referred to as “chalking” and has been confirmed by scanning electron microscopy. Methods of assessing pigment photoactivities have attracted much interest from both scientific and technological points of view. Artifical and natural weathering studies are tedious and very time consuming. Consequently, numerous model systems have been developed to assess their photochemical activities rapidly. Most of these systems undergo photocatalytic reactions to give products which are easily determined, usually by UV absorption spectroscopy, HPLC, GC, microwave spectroscopy, etc.

2.4
Photoactivity Tests for 2-Propanol Oxidation and Hydroxyl Content

These are specific tests to ascertain titanium dioxide photoactivity. The various types and grades of titania discussed in this chapter are listed in Table 2.5. The oxidation of 2-propanol to yield acetone is a specific methodology and this has been related to

Table 2.5 Properties of pigmentary and nanoparticulate titanias.

Sample	BET surface area ($m^2 g^{-1}$)	Particle size	Surface treatment	% Surface treatments
A anatase normal	10.1	0.24 μm	None	
B rutile normal	6.5	0.28 μm	Al	1
C rutile normal	12.5	0.25 μm	Al	2.8
D rutile normal	12.5	0.29 μm	Al	3.4
E nano anatase	44.4	20–30 nm	None	
F nano anatase	77.9	15–25 nm	None	
G nano anatase	329.1	5–10 nm	None	
H nano anatase	52.1	70 nm	Hydroxyapatite	5
I nano rutile	140.9	25 nm	None	
J nano rutile	73.0	40 nm	Al, Zr	13
K nano anatase	190.0	6–10 nm	Al, Si, P	20
L nano rutile	73.0	30–50 nm	Al, Zr	13
M nano anatase	239.0	71 nm	Al, Si, P	12
N nano anatase	190.0	92 nm	Al, Si, P	20
O rutile normal	12.5	250 nm	Al, Si, P	3.5

oxygen consumption during irradiation of the medium in the presence of the titania particles [30]. The hydroxyl content relates to the concentration of hydroxyl functionalities present on the pigment particles and is often related to activity [31]. The data for both tests are compared in Figure 2.12. There are a number of correlations and trends within the data. First, all the nanoparticle grades exhibit higher photoactivities than the pigmentary grades. Thus, for oxygen consumption the anatase A is more active than the rutile types B and D, the last being the least active and most

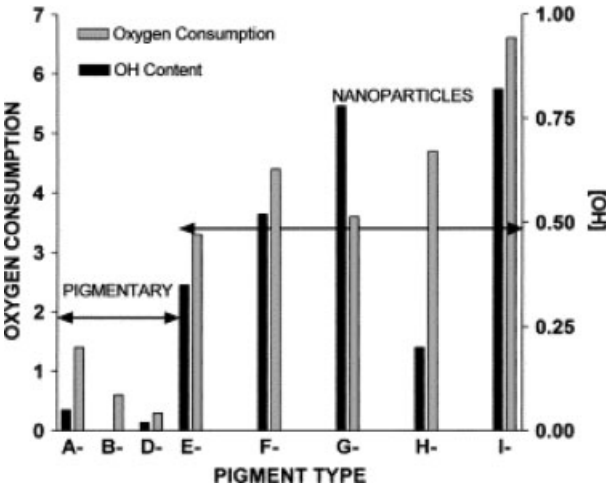


Figure 2.12 Comparison of hydroxyl content and oxygen consumption rates for 2-propanol oxidation with pigmentary and nanoparticle titanium dioxide.

durable pigment. Second, of the nanoparticles the rutile grade I is the most active in both tests. The three anatase grades E, F and G exhibit increasing activity with hydroxyl content whereas for oxygen consumption F is the greater. It is nevertheless clear from the data that nanoparticulates are significantly more active.

Environmental issues play an important role in the applications of titania fillers. These include the use of their photocatalytic behavior in the development of self-cleaning surfaces for buildings, i.e. antisoiling and antifungal growth and VOC/NO_x reduction (emissions). The latter can cause lung damage by lowering resistance to diseases such as influenza and pneumonia whereas in combination with VOCs it produces smog and contributes to acid rain, causing damage to buildings. From a commercial point of view, such benefits have enormous implications. Japanese scientists [32] have been actively exploiting the development of a variety of materials and a number of European ventures have followed suit, most notably in Italy, with Global Engineering and Millennium Chemicals as examples, and the European-funded PICADA Consortium [33]. Here, developments range from self-cleaning and depolluting surfaces and facades based on nano-titania activated coatings and cementitious materials. These applications include antisoiling, depollution of VOCs and NO_x contaminants and antifungal/microbial activities. Numerous reports have appeared in newspapers and magazine articles highlighting such applications, e.g. self-cleaning paving and building blocks and facades that can also depollute the surrounding atmosphere, internal coatings and paints for sanitization and elimination of MRSA and also, for example, clothes and textiles that supposedly never need cleaning (although in many cases this is undoubtedly an exaggeration and was promoted for public awareness of the potential). The cements are normally loaded up to 3% w/w for optimum activation and cost efficiency.

2.5
Self-Cleaning Effects: Paints/Cementitious Materials

The relative photoactivities of the nanoparticles and pigments may be compared by measuring their influence in the first instance, on the durability for example, of an 18% w/w PVC-based alkyd paint matrix. Mass loss and gloss loss are the two industrial parameters often used and the former is compared in Figure 2.13 for the range of pigments and nanoparticles given here. These results clearly show that photoactivity is divided into two main trends. For mass loss, all four nanoparticle anatase pigments are the most photoactive. The rutile pigments B and D and the anatase pigment A exhibit similar activity to that of the nanoparticle rutile grade I.

In terms of self-cleaning, this can be described in a number of formats. For cementitious materials this would imply a surface which under light activation would have the ability continuously to destroy or "burn off" by oxidation the surface dirt layers, whether they be carbonaceous, oil or soil. This can be clearly seen visually in some of the commercial trials undertaken by Millennium Chemicals in tunnels in Italy (in conjunction with Global Engineering, Milan). The photocatalytic activity of Eco-cements can be measured by, for example, determining the fading rate of an

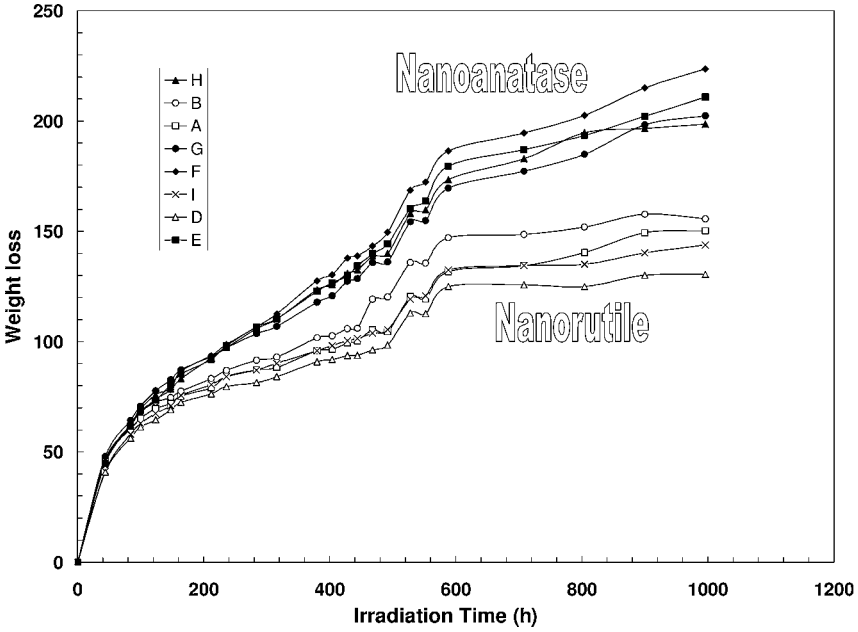


Figure 2.13 Weight loss (mg per 100 cm^2) of PVC alkyd paint films during irradiation in an Atlas Weatherometer containing equivalent amounts of titania pigments.

impregnated dye such as Rhodamine B. This is illustrated in Figure 2.14, where over a given period of light exposure the cement with photocatalyst exhibits a rapid dye fade compared with the undoped material. In reality, this is further illustrated by the photographs of real trials in Milan (shown in Figure 2.15), where the photocatalytic cement remains clean after a period of use compared with that for undoped cement.

For paints and coatings, the idea is to limit the oxidation and chalking of the paint film to the very near surface layers such that over time with weathering rain water will

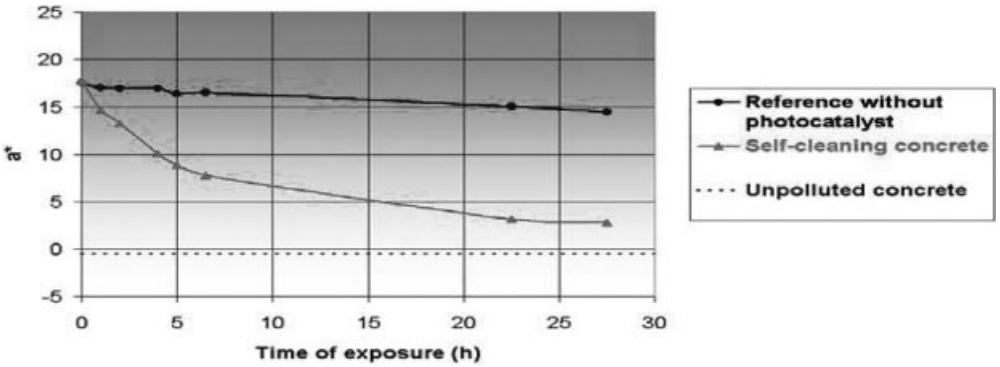


Figure 2.14 Fading rate of Rhodamine B dye impregnated into concrete with and without titania photocatalyst.

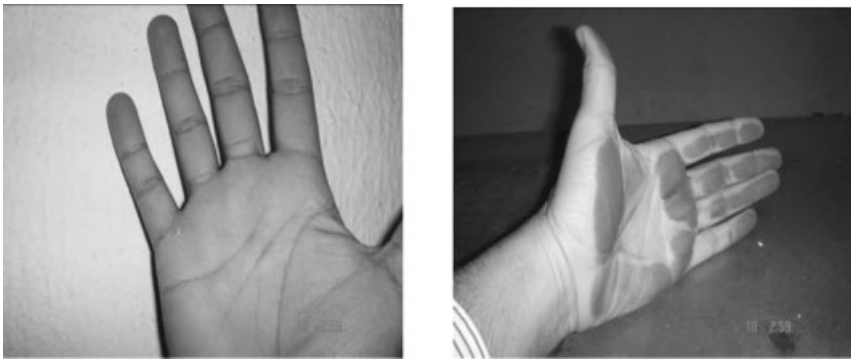


Figure 2.15 Cementitious materials with and without photocatalyst in a motorway tunnel in Italy after 3 months of exposure.

wash the top layer, leaving an underlying clean, fresh surface. The other is like the cementitious materials the surface deposits are oxidized or "burnt off" leaving the surface layer clean. In the former case it is important that the coating exhibits high durability for a reasonably cost-effective stable system. Also, the paints chosen must be stable to flocculation and viscosity changes, cure or dry at ambient temperature and ideally be water based to avoid further environmental problems. Most polymers are carbon based and are unlikely to be photo-resistant, but water-based acrylic latex paints have been evaluated. In the first instance four types of acrylic water-based paints were evaluated in terms of relative stability toward photoactive nanoparticles. Here a special sol–gel grade of anatase was prepared in the laboratory with no post-firing. Particles of varying sizes were also prepared via this route. The relative paint stabilities with and without the anatase sol particles (10–20 nm) at 5% w/w after 567 hours of weathering are shown in Table 2.6. Of these paint formulations, only the polysiloxane BS45 (Wacker) proved to be resistant to the photocatalytic effects of the titania particles. The styrene–acrylic, poly(vinyl acetate) and acrylic copolymers all showed high degrees of chalking (weight loss).

Table 2.6 Weight loss for paints after 567 h of Atlas exposure: various polymers plus 5% anatase sol particles, 10–20 nm.

Paint composition	Weight loss (%)
Styrene acrylic	12.2
Styrene acrylic + anatase sol	97.3
PVA copolymer	11.4
PVA copolymer + anatase sol	97.9
Acrylic copolymer	7.4
Acrylic copolymer + anatase sol	101.0
Polysiloxane BS45	23.3
Polysiloxane BS45 + anatase sol	13.6

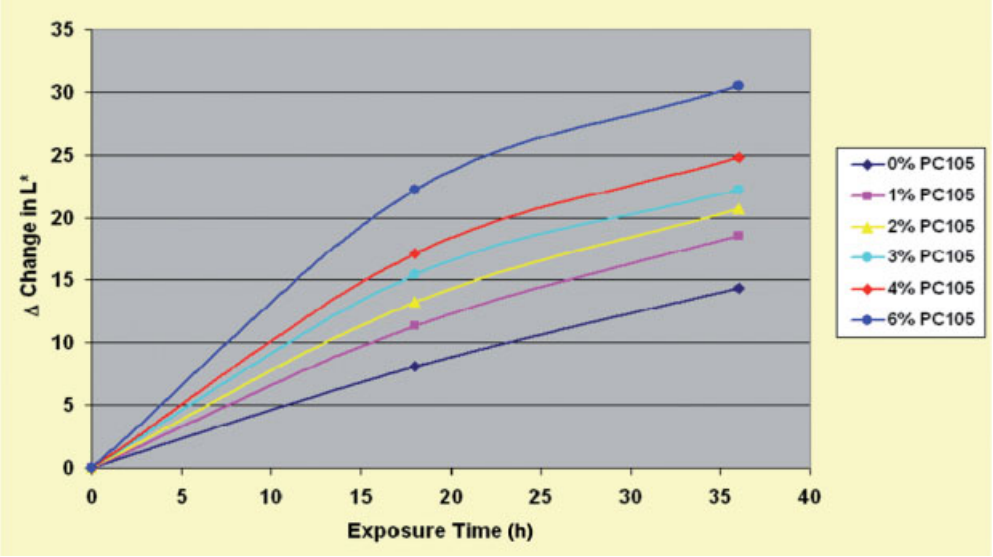


Figure 2.16 Change in color difference factor L* with exposure time for methylene blue-impregnated silicate paint films with increasing nanoparticle titanium dioxide PC105.

From the point of view of surface cleaning, paint films can also be impregnated (like cement) with dyes and fading rates measured. Aside from organic-based paints, a number of inorganic paints are commercially available, a number made from complex alkali metal silicates. Because of their inorganic nature, they tend to be significantly light stable. An example of the self-cleaning effect of a typical silicate-based paint is demonstrated by the fading data on methylene blue dye in Figure 2.16. It is seen that photobleaching of the dye occurs more rapidly in film with photocatalyst than an undyed (undoped) film and that this increases with increasing concentration of PC105 nanoparticles.

Another method used potentially to enhance the durability of a substrate while simultaneously controlling photocatalytic activity is to dope the paints with mixtures of durable and catalytically active grades of titanium dioxide. In this regard, mixtures of pigmentary rutile O and nanoparticle anatase F pigments appear to provide one interesting illustration option, with the former inducing some level of base stability while the presence of the latter gives rise to surface activity. Figures 2.17 and 2.18 illustrate this effect for a siliconized polyester coating exposed in a QUV weatherometer for gloss and mass loss, respectively. Gloss loss is seen to be gradually reduced with time the effect increasing with increasing loading of anatase nanoparticle F. Mass loss is also seen to increase gradually with increasing levels of the same nanoparticle. In this case, it is evident that only low levels of shedding/chalking occur with time such that the paint film retains some level of durability except for the very near surface layer.

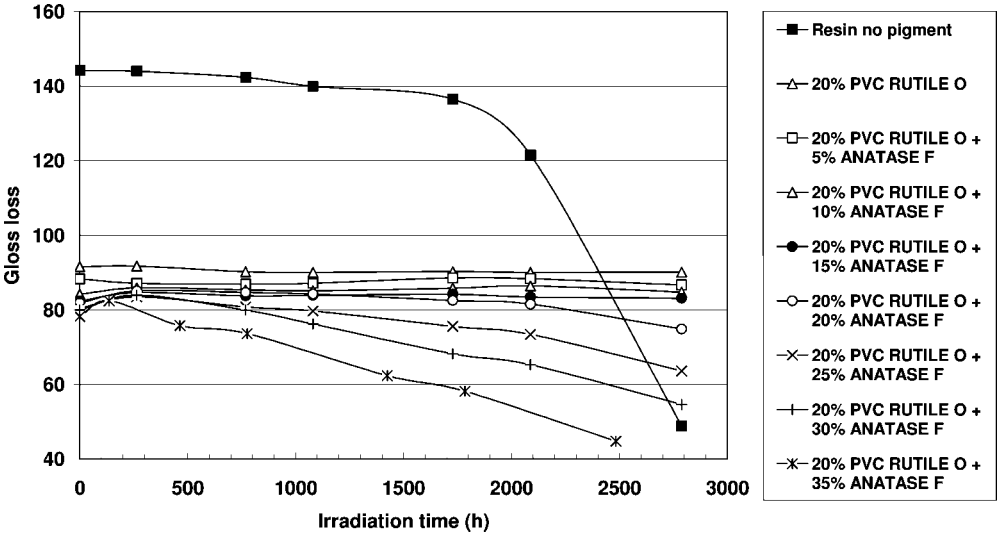


Figure 2.17 Gloss loss versus irradiation time in a QUV weatherometer for a DSM siliconized polyester resin with 20% w/w rutile pigment O plus increasing levels of 5, 10, 15, 20, 25, 30 and 35% w/w of nanoparticle anatase F.

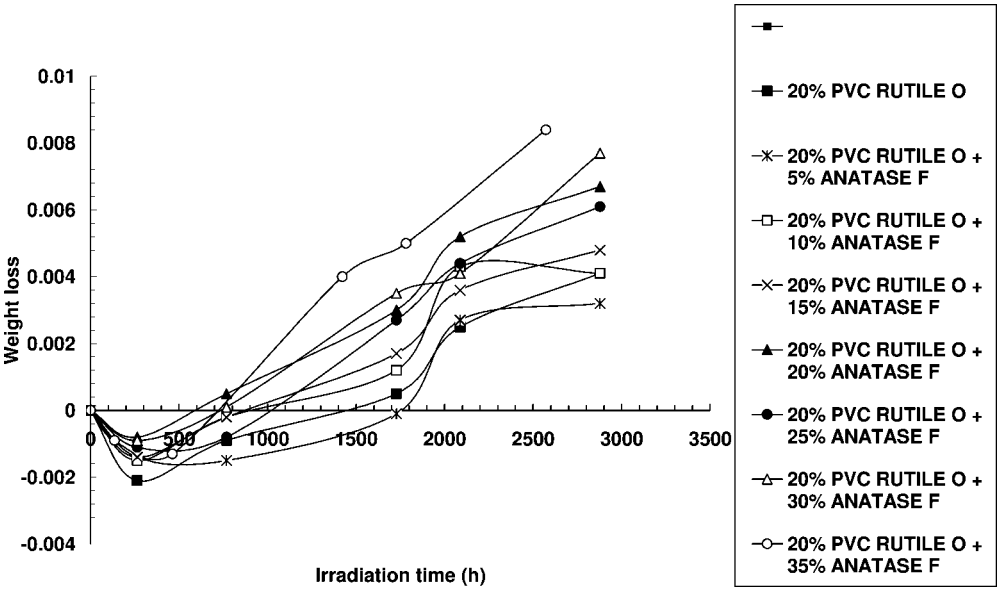


Figure 2.18 Mass loss versus irradiation time in a QUV weatherometer for a DSM siliconized polyester resin with 20% w/w rutile pigment O plus increasing levels of 5, 10, 15, 20, 25, 30 and 35% w/w of nanoparticle anatase F.

Table 2.7 Weight loss for Lumiflon paint pigmented with RCL-696/nano-TiO_2 after 546 h of Atlas exposure.

Nano-TiO_2	Pigmentary TiO_2	Weight loss (%)
10 wt.% PC500	RCL-696	19.0
20 wt.% PC500	RCL-696	66.5
10 wt.% PC105	RCL-696	31.0
20 wt.% PC105	RCL-696	62.8
10 wt.% PC50	RCL-696	30.4
20 wt.% PC50	RCL-696	39.0
10 wt.% Showa Denko	RCL-696	77.0
20 wt.% Showa Denko	RCL-696	105.4
10 wt.% AT1	RCL-696	16.6
20 wt.% AT1	RCL-696	43.2
20 wt.% PC500	None	97.6
20 wt.% PC105	None	128.7
20 wt.% PC50	None	121.4
20 wt.% Showa Denko	None	146.8
20 wt.% AT1	None	138.7
None	RCL-696	4.7
Clear resin blank	None	5.4

A similar but perhaps more extreme effect is shown in Table 2.7 for a Lumiflon fluorinated acrylic paint film. At 10 and 20% concentrations of the nanoparticles G, F, E and H, chalking is fairly high, whereas the pigmentary rutile O at 20% w/w only gives a 4.7 mass loss value. The pigmentary uncoated anatase A is also an option, giving high levels of chalking at 10 and 20% w/w. Thus, control of pigment type and particle size in addition to their concentrations is a critical area of development for effective self-cleanable paint surfaces, the effect varying also with the paint formulation. In this regard, coatings could effectively be developed to suit a particular type of environment.

2.5.1 Antibacterial Effect

The ability of the nanoparticles to destroy bacteria and fungi has also been actively pursued and some data from our laboratories are demonstrated. The type of photocatalytic medium, nanoparticle and bacteria/fungi all play an intimate key role in performance. There are a number of tests one can apply [34–36], the simplest evaluation being the typical zone of inhibition on agar plates where the growth of bacteria is measured around a paint film. *Staphylococcus aureus* growth is shown on agar plates in Figure 2.19 for typical silicate paint films with and without PC105 nanoparticles. On the right-hand picture plate a clear zone of inhibition is seen to develop compared with that for the undoped film.

Similar tests have also been undertaken in our laboratories on titanium dioxide powders. Here *E. coli* were used where their destruction (measured in terms of

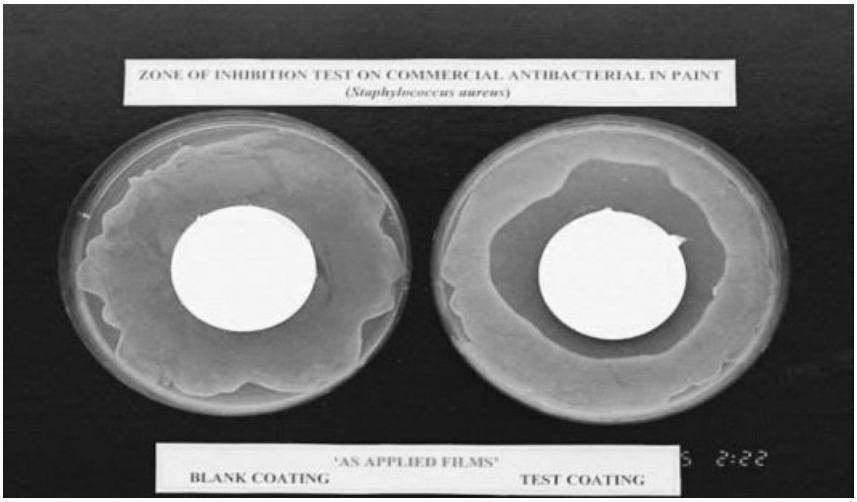


Figure 2.19 A typical antibacterial evaluation on an agar plate medium for Eco-paint films with and without nanoparticulate photocatalyst titania.

colony-forming units) after irradiation with UV light in the presence of the titania particles is plotted against irradiation time. A study on the range of titania powders showed (Figure 2.20) that there was an inverse relationship between antibacterial activity and particle size: for the pigment powders, pigment E > H > F > G = A, with a

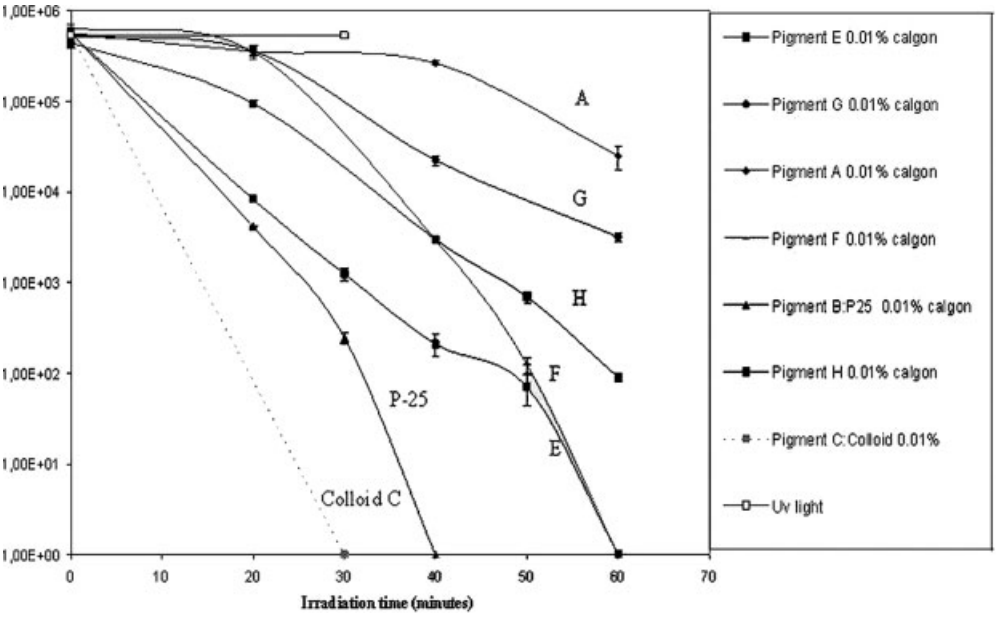


Figure 2.20 Photocatalytic bactericidal effect of pigments A, B, E, F, G and H powders and C (colloid) at 0.01% in stirred conditions, using unwashed cell suspensions.

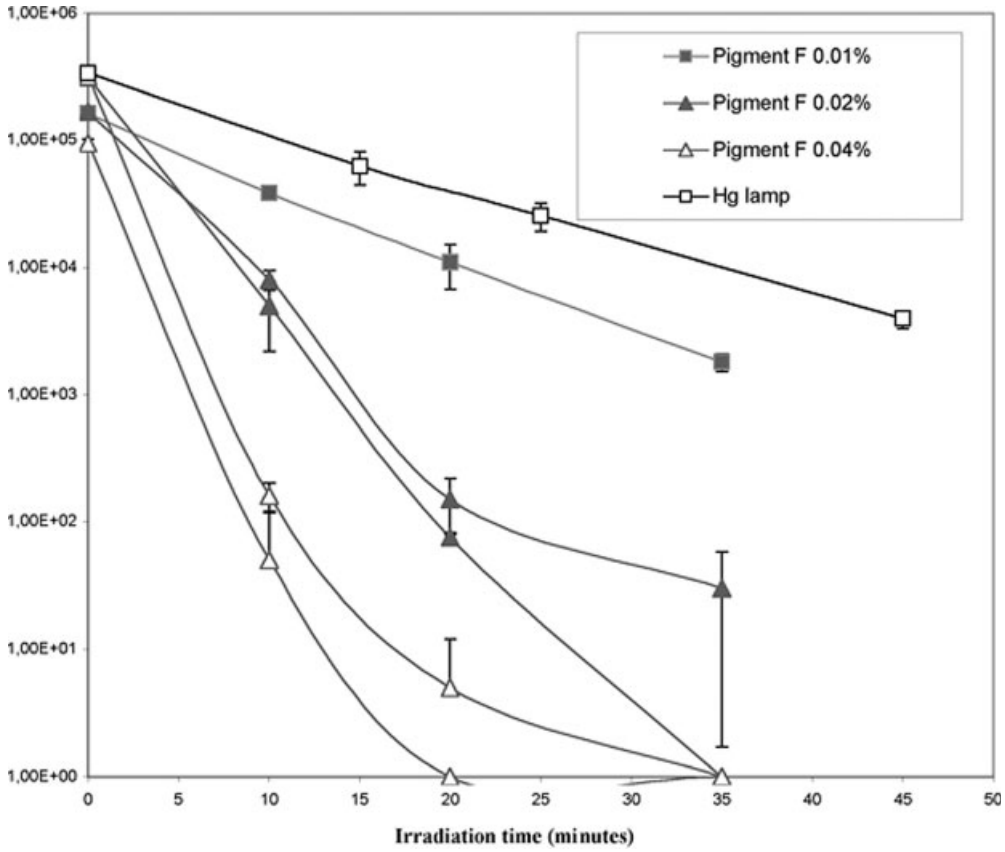


Figure 2.21 Effect of concentration of pigment F on photocatalytic bactericidal effect.

sol–gel colloid dispersion C and Degussa P25 having the greatest effects. However, the experimental conditions used provided some confounding factors which required clarification in order to identify the best experimental method and the most effective pigments in terms of antibacterial activity, as mentioned previously. Here the particles were dispersed in a Calgon medium and this evidently reduced activity on the plate.

In general, the antimicrobial effect increased with increasing concentration of nanoparticles (Figure 2.21) up to 0.04%. In the absence of any dispersion effect, the activity of the nanoparticle E was comparable to that of the Degussa P25, followed by F and G and with little difference between A, H and D and the mercury lamp control (Figure 2.22). Similar results were observed at 0.02% in our work. For the nanopigments E, F and G, a further enhanced effect was noted at 0.1%, but the effect of P25 was reduced.

The enhanced activity of C over its derivative pigment G was lost when C was dried and ground (Figure 2.23). This finding demonstrates that the drying process has a

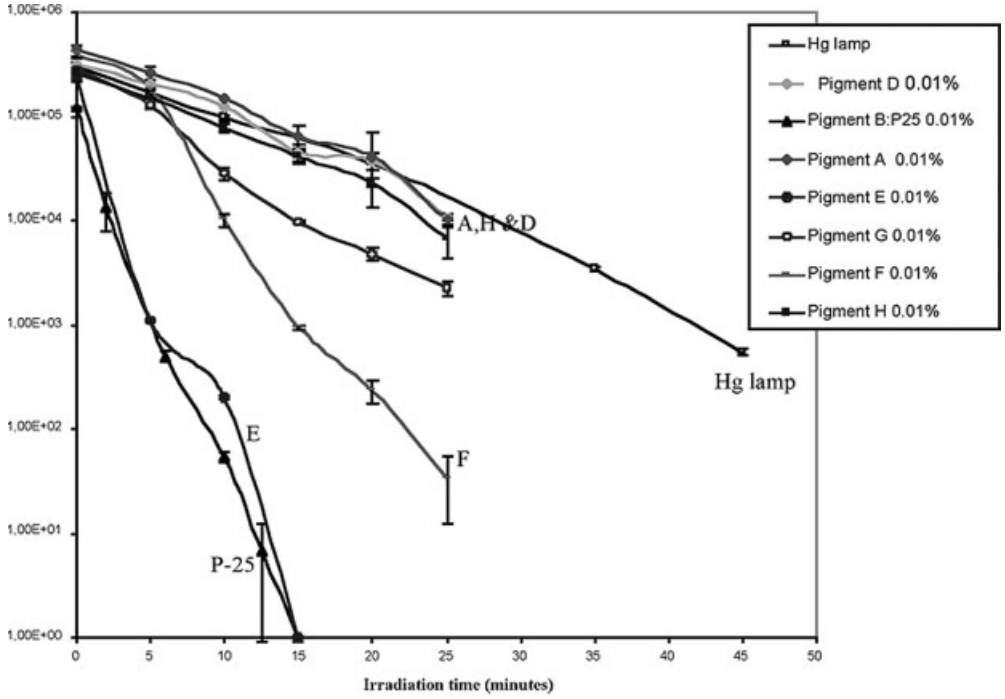


Figure 2.22 Photocatalytic bactericidal effect of pigments A, B D, E, F, G and H at 0.01%, using washed cell suspensions.

marked effect on activity, due to a decrease in surface area during aggregation and to a decrease in dispersion stability in water.

Pigment E has the most reactive surface because it has fewer defects, which increases the efficiency of the photogenerated radicals form our microwave analysis [33]. Thus pigments calcined at higher temperatures (E > F > G) have better crystallinity and therefore higher antibacterial activities.

The UV light itself has little effect on the bacteria; the pigmentary grade of anatase A has a small effect whereas the nanoparticle G has a somewhat greater effect. However, the most interesting feature of these data is the very high destructive effect of the mixed phase nanoparticle grade made by Degussa (P25). This nanoparticle grade of titania is well established is the literature in terms of its high photoactivity [32]. In this work, a grade of nanoparticle anatase G was prepared in the laboratory whereby the particles were seeded form solution and then dried but not subsequently oven fired. This so-called washed form of titania is seen in the data to be higher in activity than that of the Degussa material. This effect is currently being investigated further in terms of hydroxyl content and hydrogen peroxide generation.

The overall effect of activity will depend on whether more TiO_2 is activated as a consequence of increased surface area or whether less TiO_2 is activated because less light passes through the suspension due to light scattering. Larger aggregates of particles sediment in a liquid system and an increased concentration of pigments

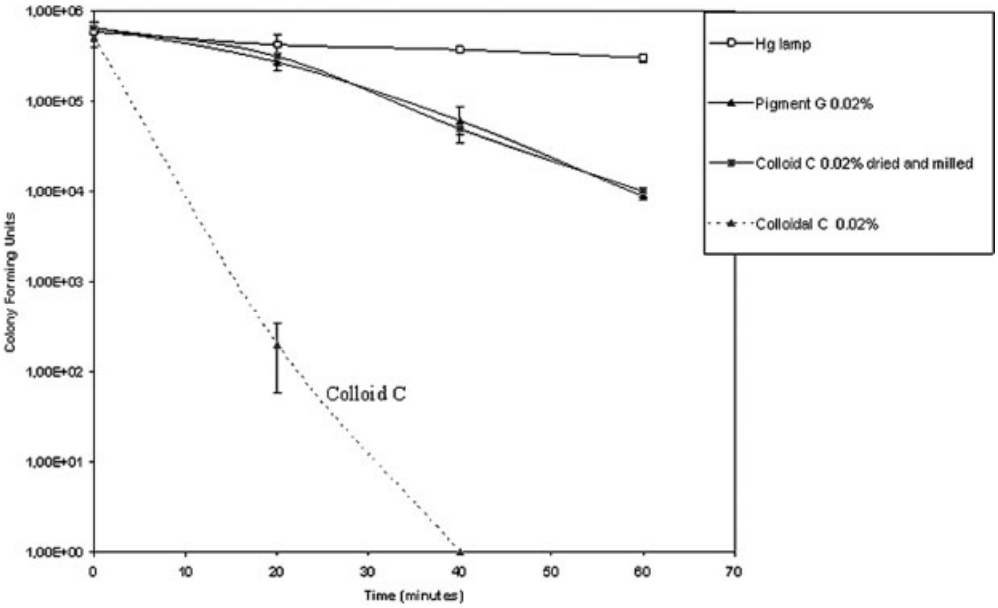


Figure 2.23 Photocatalytic bactericidal effect for C colloid suspension, C dried and milled and pigment G powder (unstirred).

shows less of an antimicrobial effect since less light passes through the suspension if the cell–particle mixture is not stirred. Conversely, Calgon milled pigments, which are nanometer sized, also scatter light significantly at high concentrations and decrease activity (optimum loading 0.01%), hence the optimum activity is presented by nanoparticle powder aggregates in this work. The most important aspect to consider in terms of antibacterial inactivation is the relative sizes of the titanium particles/aggregates and the bacterial cell. *E. coli* measures approximately $1 \times 3\ \mu m$; a benzene molecule is $0.00043\ \mu m$. The porosity of the pigment has no bearing on the antimicrobial effect, whereas the chemical pollutant can diffuse into the porous particle structure. Thus, the higher surface area of pigment G did not enhance any antibacterial effect. It has been verified using a disc centrifuge in our study that the three nanoparticulate powder pigments E, F and G were aggregated into $0.7\ \mu m$ particles. In this case, they would all offer comparable active areas to bacteria. Only the inherent ability of the pigments to generate radicals will affect antibacterial activity. Hence the process is more sensitive to structure (crystallinity) than to texture (surface area) and follows a clear inverse relationship with particle surface area.

2.5.2
Depollution: NO_x/VOC Removal

The ability of photocatalytic surfaces to depollute the surrounding atmosphere with a certain radius has been well documented recently in the literature [32, 33, 36].

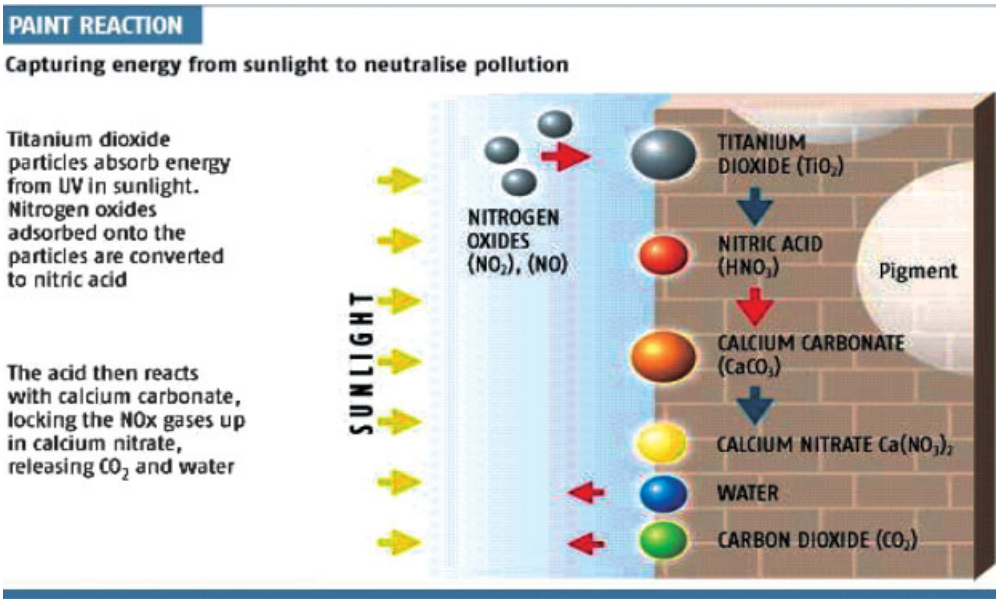


Figure 2.24 Depollution scheme for photocatalytic paints and surfaces.

Japanese scientists have been particularly prolific in this area for some time now. In this part of the research work it was important to be able to develop coatings and cementitious materials that remove NO_x, SO_2, VOCs and potentially ozone especially in areas where such contamination is likely to be above recommended standards. Examples include motorway tunnels, underground car parks, busy highways, chemical factories and city dwellings such as schools. A pictorial representation of the key mechanistic features of the depolluting paint coatings is illustrated in Figure 2.24.

The materials should in this regard be durable and show little or no loss in activity with aging, in addition to having the ability to inactivate nitric acid reaction products. Also, as above it should be self-cleaning. The coating may also, in some cases, need to be translucent so that existing coatings or stonework can be over-coated without any change in appearance. To some extent the coating must be photo-resistant to the effects of the nano-TiO_2 and would probably need to be porous to allow contact between the TiO_2 surface and the gaseous pollutants. Nano-TiO_2 is an excellent scatterer of light and if the coating is porous this further increases light scatter. Some potential problems in the design of such coatings have been circumvented in our laboratories, such as poor adhesion and poor durability. Also, the nitric acid formed in the reaction could damage the substrate or poison the catalytic reaction. A suitable test method was developed to measure the efficacy of the coatings studied via a "Signal" detection system (Figure 2.25).

In the diagram shown, test films of paint are irradiated in a cell through which a standard flow rate of nitrogen is passed with a set concentration of NO_x gases. NO_x

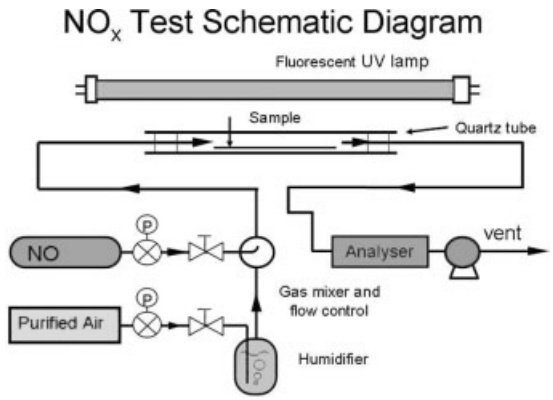


Figure 2.25 Schematic diagram of NO_x gas detection system for irradiated photocatalytic surfaces.

levels are measured via a chemiluminescence detector system before and after irradiation to give a measure of depollution.

Commercial dry nano-TiO_2 products with a range of particle size and surface area were developed with surface areas ranging from 20 to 300 $m^2 g^{-1}$ for evaluation as indicated in Table 2.5. In addition to these, colloidal sol–gel particle media were also developed for easy dispersion. Even with the smallest crystallite size it is difficult to eliminate light scattering at levels above 5% at conventional coatings thickness (25 μm) due to aggregation. With special non-dried sol–gel nano-TiO_2 there is less light scattering because of reduced particle aggregation. It appeared that the coatings had to be porous before there was a significant activity towards gaseous reductions such as NO_x.

From the data in Figures 2.26 and 2.27, the efficacy of NO_x removal increases significantly with both an increase in particle surface area and concentration of nanoparticle titania (anatase) in a polysiloxane paint substrate. Porosity can also be introduced by using materials other than TiO_2 itself. Nanoparticle calcium carbonate offered the possibility of high translucency and the ability to react with nitric acid. The results are confirmed in Figure 2.28, where it is seen that the NO_x is reduced

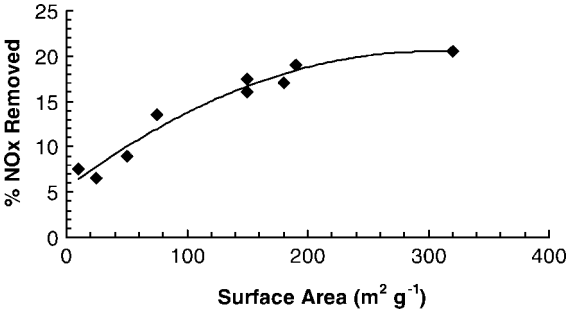


Figure 2.26 Percentage concentration of NO_x removed versus the surface area of anatase sol–gel particles at 5% w/w in a polysiloxane Wacker BS 45 paint system.

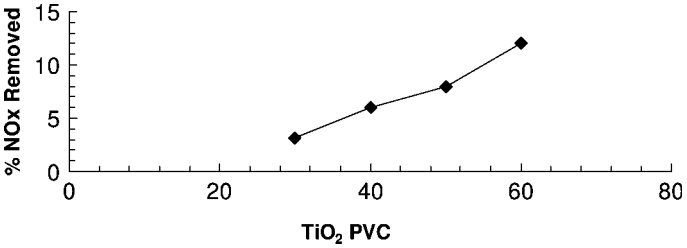


Figure 2.27 Percentage concentration of NO_x removed versus the concentration of anatase sol–gel particles (10–20 nm) at 5% w/w in a polysiloxane Wacker BS 45 paint system.

not only with increasing titania doping but also with increasing levels of calcium carbonate addition.

The most interesting feature of the results however, is the influence of titania and calcium carbonate loading on the extent of degradation of the polysiloxane paint films as measured by percentage weight loss. The data in Figure 2.29 show that in the absence of calcium carbonate the extent of degradation is low, as indicated above, whereas in its presence the rate of degradation increases with concentration from 2.5 to 5.0% w/w. At 10% w/w of titania the extent of degradation is significant in the presence of the calcium carbonate. In this case the access of both moisture and oxygen through the film matrix will be enhanced. Film translucency also decreases with increasing loadings of titania and calcium carbonate particles, as shown by the data on contrast ratio in Figure 2.30.

Measurements on NO_x reductions have also been obtained in terms of NO and NO_2 gases, where it is seen that the rate of NO_x destruction is clearly greater in the presence of the nanoparticles alone whereas the paint matrix gives rise to a barrier effect, as might be expected (Table 2.8, Figure 2.31). Nevertheless, the efficacy of the paint films in destroying the NO_x gases is high. The durability of a paint film in terms of NO_x reduction is also important and this is illustrated by the plot in Figure 2.32 for a typical Eco-silicate paint system. Here the percentage reduction in NO_x ability is reduced by only about 10% and thereafter stabilizes after 12 months.

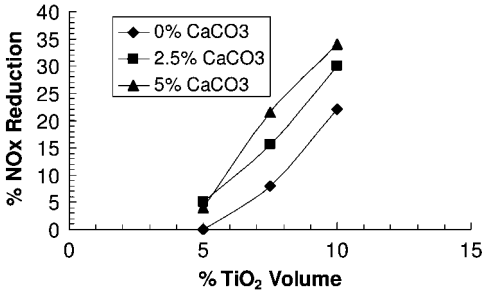


Figure 2.28 Percentage NO_x reduction versus volume of titania (anatase 10–20 nm) for a polysiloxane BS 45 paint substrate with 0, 2.5 and 5.0% w/w of nanoparticle calcium carbonate.

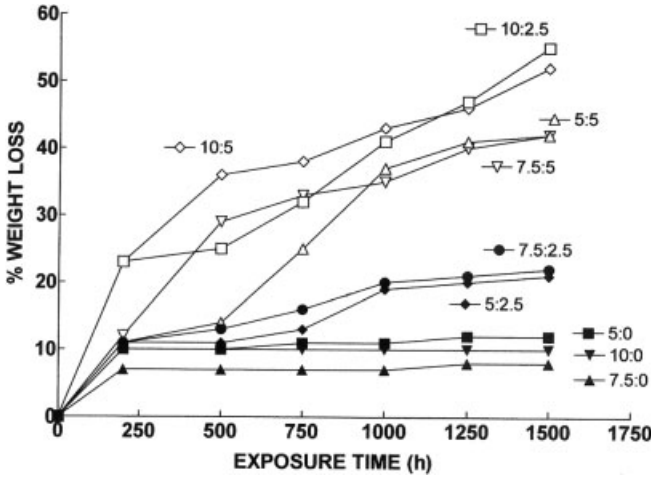


Figure 2.29 Percentage weight loss versus exposure time in an Atlas Ci65 weatherometer for polysiloxane paint films (BS 45) containing different ratios of nanoparticle anatase (10–20 nm) sol–gel titania (5/7.5/10) to calcium carbonate (0/2.5/10) particles.

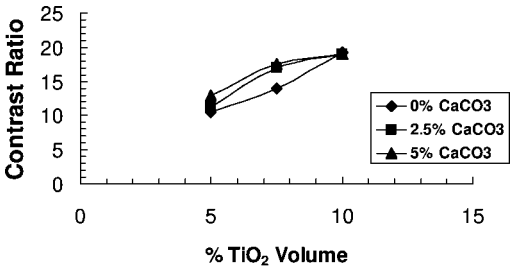


Figure 2.30 Translucency (contrast ratio) for BS 45 paint films with volume addition of sol–gel anatase titania (10–20 nm) particles versus calcium carbonate addition.

Table 2.8 NO_x reduction comparisons for polysiloxane latex with and without titania sol–gel and sol–gel alone using steady-state signal detection apparatus.

Composition	NO_x reduction (%)		NO_x reduction ($\mu g\, m^{-2}\, s^{-1}$)	
	NO	NO_2	NO	NO_2
BS45 latex	0	0	0	0
BS45 latex + 5% sol	84.9	9.3	0.060	0.055
Sol	84.9	55.8	0.320	0.409

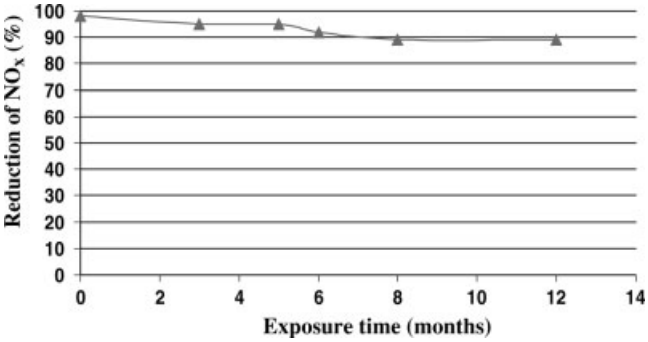


Figure 2.31 Plot of reduction of NO_x versus exposure time.

Another important factor is irradiation power or light flux density. Optimum power appears to be achieved at 0.6 $mW\,cm^{-2}$ of film, as shown in Figure 2.32.

The effectiveness of the Eco-paint and cemetitious materials in terms of depollution in surrounding areas is also important and effectively demonstrated by the tunnel experiment in Figure 2.33. Here one wall is effectively coated with a titania-doped cement whereas the other is undoped. NO_x measurements under stedy-state irradiation show significantly less ppb concentrations for the titania-doped left wall under both actual conditions and also via mathematical modeling experiments. With the same experiment using the Eco-cement coating, VOC reductions can also be measured in relation to the air velocity. The data in Figure 2.34 shows that as the air velocity is reduced so the VOC concentrations are effectively reduced. The benzene being unsubstituted is more difficult to decompose photocatalytically and therefore requires a slower abatement air speed for effective decomposition. The greater the degree of alkyl group substitution, the more effective and easier is the decomposition rate. Alkyl groups are more easily oxidized than benzene rings.

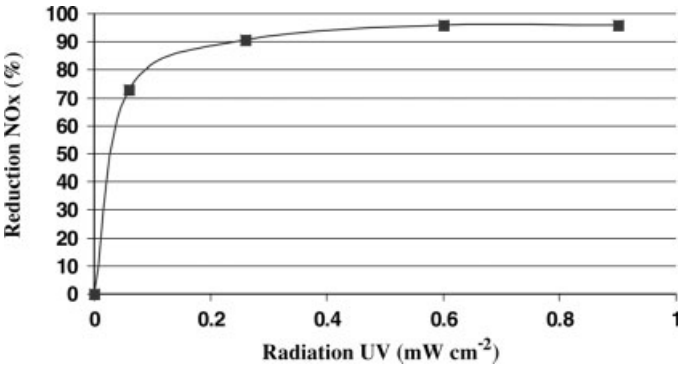


Figure 2.32 Reduction in NO_x with UV radiation intensity for a typical Eco-silicate paint film.

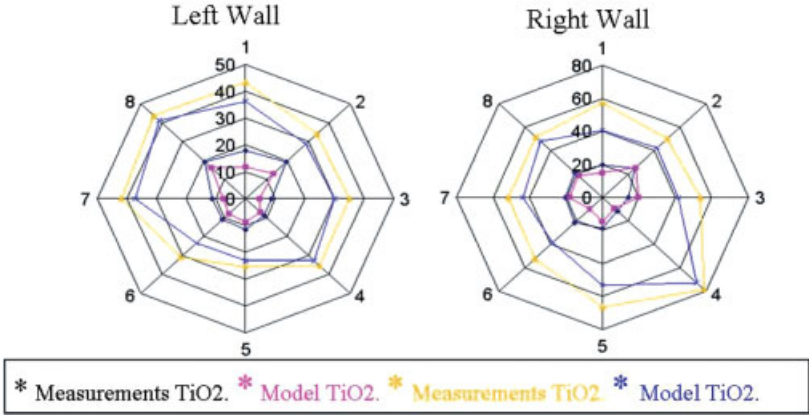


Figure 2.33 NO_x concentrations around a tunnel wall with and without photocatalytic titanium dioxide doping for a cementitious facade. NO_x mean concentrations (ppb) per sector with TiO_2 (period A) and without TiO_2 (period B) for the right wall.

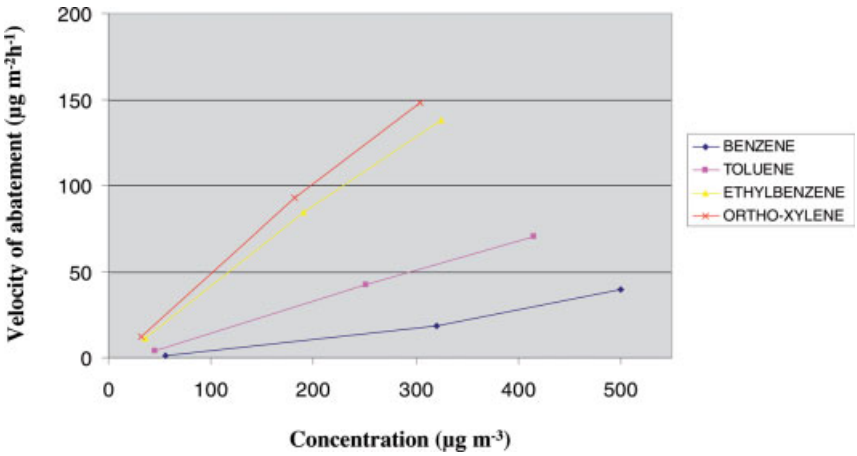


Figure 2.34 Steady-state concentration of aromatics in the vicinity of an Eco-cement wall coating with a change in surrounding air velocity.

2.6 Conclusions

Photo-oxidation studies on paint films show a clear demarcation between nanoparticle and pigmentary grade titanium dioxide, with the former being more active. Model system studies based on 2-propanol oxidation and hydroxyl analysis go some way to predicting pigment activities but precise correlations do not always exist in the real world.

The use of nanoparticle anatase in conjunction with pigmentary rutile grades is also a viable option for the development of self-cleaning paint surfaces. For

antibacterial surfaces nanoparticles are effective, whereas pigmentary grades are ineffective. Highly effective photocatalytic grades of nanoparticles can also be prepared through control of the preparation and subsequent drying operations.

The antibacterial activity of the nanoparticle pigments was inversely proportional to particle size and relates to their intrinsic ability to generate active carriers giving rise to active surface species. Pigments calcined at higher temperatures, consequently with fewer structural defects, are more active, because defects act as recombination centers for the electrons and holes. Hence the antibacterial efficiency of TiO_2 is not determined by surface area but by the ability to generate active carriers, resulting eventually in the formation of effective chemical species such as peroxides (hydrogen peroxide). This is not surprising because of the size of bacteria relative to the pigments: the majority of the surface area offered by a pigment is sterically unavailable to the bacterial cells. In terms of future building developments, especially in the medical world, this would offer significant advantages for eliminating the potential of MRSA.

The paint coatings are also active to NO_x and VOCs, particularly once UV irradiated, with high levels of TiO_2 and $CaCO_3$ enhancing activity. This effect is associated with increased porosity of the paint system induced by both the titania and calcium carbonate particles. Unfortunately, higher levels of TiO_2 and $CaCO_3$ impart lower durability to the paint matrix. Higher levels of TiO_2 and $CaCO_3$ also reduce the translucency of the paint films, thus increasing absorptivity. On a positive note, higher levels of $CaCO_3$ would react with more HNO_3.

Photocatalytic cementitious materials offer significant advantages from an environmental point of view on all issues associated with long-term activity, durability, self-cleaning and depollution of NO_x and VOCs.

Acknowledgment

The authors would like to thank the PICADA consortium Global Engineering, Milan, Italy for the use of some of the data in this chapter.

References

1 J. Murphy, *Additives for Plastics Handbook*, 2nd edn., Elsevier, Amsterdam, 2001.
2 T. C. Patton, *Pigment Handbook*, Vol. 1, Wiley, New York, 1973.
3 D. R. Vesa, P. Judin, *Verfkroniek*, 1994, **11**, 17.
4 A. Gurav, T. Pluym, T. Xiong, *Aerosol Sci. Technol.* 1993, **19**, 411.
5 J. G. Balfour, *New Mater.* 1992, **1**, 21.
6 N. S. Allen, J. F. McKellar, *Photochemistry of Dyed and Pigmented Polymers*, Applied Science, London, 1980, p. 247.
7 N. S. Allen, M. Edge, A. Ortega, C. M. Liauw, J. Stratton, R. McIntyre, *Polym. Degrad. Stabil.* 2002, **78**, 467.
8 A. Fujishima, T. N. Rao, D. A Tryk, *Electrochim. Acta* 2000, **45**, 4683.
9 L. Gao, Q. Zhang, *Scr. Mater.* 2001, **44**, 1195.
10 W. Xu, S. Zhu, X. Fu, *J. Phys. Chem. Solids* 1998, **59**, 1647.
11 M. R. Hoffmann, S. W. Martin, W. Choi, D. W. Bahnemann, *Chem. Rev.* 1995, **95**, 69.

12 M. Anpo, T. Shima, S. Kodama, Y. Kubokawa, *J. Phys. Chem*. 1987, **91**, 4305.

13 H. Jang, S. Kim, *Mater. Res. Bull*. 2001, **36**, 627.

14 X. Ye, Y. Jin, G. Lie, in Proceedings of ICETS 2000-ISAM, 11 October 2000, Beijing, Session 3, Vol. 1, 718–721.

15 G. A. Somorjai, *Chemistry in Two Dimensions: Surface*, Cornell University Press, Ithaca, NY, 1981, 551.

16 A. Mills, S. Morris, R. Davies, *J. Photochem. Photobiol. A: Chem*. 1993, **71**, 285.

17 D. Chatterjee, S. Dasgupta, *J. Photochem. Photobiol. C: Rev*. 2005, **6**, 186.

18 N. S. Allen, in N. S. Allen, (ed.), *Degradation and Stabilization of Polyolefins*, Elsevier Science, London, 1983, Chapter 8, 337.

19 N. S. Allen, M. Edge, *Fundamentals of Polymer Degradation and Stabilization*, Chapman and Hall, Chichester, 1992.

20 G. Kaempf, W. Papenroth, R. Holm, *J. Paint Technol*. 1974, **46**, 56.

21 N. S. Allen, J. F. McKellar, D. Wilson, *J. Photochem*. 1977, **7**, 319.

22 N. S. Allen, D. Bullen, J. F. McKellar, *J. Mater. Sci*. 1979, **14** 1941.

23 R. E. Day, *Polym. Degrad Stabil*. 1990, **29**, 73.

24 N. S. Allen, J. F. McKellar, D. G. M. Wood, *J. Polym. Sci., Polym. Chem. Ed*. 1975, **13**, 2319.

25 N. S. Allen, J. L. Gardette, J. Lemaire, *Dyes Pigments* 1982, **3**, 295.

26 A. Fujishima, T. N. Rao, D. A. Tryk, *J. Photochem. Photobiol., Rev. Ed*. 2000, **1**, 1.

27 M. T. Bryk, *Degradation of Filled Polymers*, Ellis Horwood, Chichester, 1991.

28 S. P. Pappas, W. Kuhhirt, *J. Paint Technol*. 1975, **47**, 42.

29 H. G. Voeltz, G. Kaempf, H. G. Fitsky, *Prog. Org. Coat*. 1972, **14** 1941.

30 R. C. Cundall, R. Rudham, M. S. Salim, *J. Chem. Soc., Faraday Trans. 1* 1976, **72**, 1642.

31 A. H. Boonstra, C. A. H. A. Mustaers, *J. Phys. Chem*. 1973, **79**, 1694.

32 M. Kaneko, I. Okura,(eds.), *Photocatalysis: Science and Technology*, Springer, Heidelberg.

33 N. S. Allen, M. Edge, G. Sandoval, J. Verran, J. Stratton, J. Maltby, *Photochem. Photobiol*. 2005, **81**, 279.

34 D. M. Blake, P. C. Maness, Z. Huang, E. J. Wolfrum, J. Huang, W. A. Jacoby, *Sep. Purif. Methods* 1999, **28**, 1.

35 Y. Kikuchi, Y. Sunada, T. Iyoda, K. Hashimoto, A. Fujishima, *J. Photochem. Photobiol. A: Chem*. 1997, **106**, 51.

36 K. Sunada, Y. Kikuchi, K. Hashimoto, A. Fujishima, *Environ. Sci. Technol*. 1998, **32**, 726.

3
Nanosized Photocatalysts in Environmental Remediation

Jess P. Wilcoxon and Billie L. Abrams

3.1
Introduction

3.1.1
Global Issues

Modern industrial economies have developed approaches to manufacturing, utilization and disposal of chemical and biochemical products which have inflicted considerable damage on our air and water environments. As such, advances in technology, medicine, mining, transportation, agricultural practices and military practices have not come without a price. Although the quality of human life has benefited in many ways from advances in these areas, the anthropogenic impact on the aquatic and terrestrial biosphere has been substantial, leading to pollution of the world's drinking water, soils and air. Unfortunately, the adverse anthropogenic effects on the environment are increasing [1]. This poses an undeniable threat to the ecosystem, biodiversity and ultimately human health and life.

Industry produces an estimated 300 million tons of synthetic compounds annually, a large percentage of which ends up as environmental pollutants [2]. As of 2001, approximately 100 000 metric tons of chemicals were released into surface waters and more than 720 000 metric tons were released into the atmosphere [3]. Table 3.1 shows the 2001 toxic release inventory (TRI) figures [3]. The numbers specifically for surface water and air pollution have not changed significantly for the 2004 TRI [4, 5]. However, as of the 2004 TRI, the total amount (including underground injection, landfills and wastewater) of toxic chemical released into the environment as a result of industrial practices stands at over 1.9 million metric tons [4, 5].

Accidental oil and gas spills on the order of 0.4 million tons have also resulted in significant damage to the aquatic ecosystem [2]. In the Niger Delta alone, more than 6800 spills have been documented (approximately one spill per day for the past 25 years); however the real number is thought to be much higher [6].

Nanotechnology. Volume 2: Environmental Aspects. Edited by Harald Krug

ISBN: 978-3-527-31735-6

Table 3.1 US EPA 2001 Toxic Release Inventory showing surface water discharge and total air emissions for all chemicals produced by industry. (Reprinted with permission from T. Ohe, T. Watanabe, K. Wakabayashi, Mutagens in surface waters: a review, *Mutation Research* **2004**, *567*, 109).

Industry type	Total water releases ($\times 10^3$ kg)	Total air emissions ($\times 10^3$ kg)
Chemical and allied products	26117.1	103348.6
Food and related products	25018.2	25463.3
Primary metal smelting and processing	20262.5	26132.9
Petroleum refining and related industries	7752.9	21849.6
Paper and allied products	7500.9	71283.5
Electric, gas, and sanitary services	1596.5	325492.4
Electronic and other electrical equipment	1332.2	5770.3
Fabricated metal products	790.8	18346.9
Photographic, medical and optical goods	646.1	3250.9
Coal mining and coal mine services	344.8	348.7
Tobacco products	241.7	1130.3
Metal mining (e.g., Fe, Cu, Pb, Zn, Au, Ag)	193.8	1294.8
Transportation equipment manufacture	89.9	30251.4
Textile mill products	79.6	2603.9
Stone, clay, glass, and concrete products	73.5	14181.8
Leather and leather products	56.6	547.7
Plastic and rubber products	32.2	34973.1
Solvent recovery operations (under RCRA)	10.7	442.0
Lumber and wood products	9.0	13825.1
Industrial and commercial machinery	8.2	3755.7
Petroleum bulk stations and terminals	5.1	9600.4
Chemical wholesalers	0.8	569.0
Furniture and fixtures	0.3	3548.9
Printing, publishing and related industries terminals	0.1	8750.2
Apparel	<0.1	155.7
No reported SIC code	483.2	1528.3
Miscellaneous manufacturing	16.6	3068.5
Total	100153.0	761763.6

Data obtained from: http://www.epa.gov/triexplorer/industry.htm and http://www.epa.gov/region5/defs/html/rcra.htm.

Historically, environmental protection regulations were lax, which encouraged disposal of many chemicals in landfills, with little documentation of the types and amounts of chemicals present. The worst of these sites became Superfund sites, requiring billions of dollars of expenditure to treat using existing technologies. These Superfund sites in the USA arose mainly as a result of industrial waste spills, abandoned mines, landfills and contamination of groundwater and soil due to abandonment of environmentally hazardous or controlled areas [7, 8]. As of the most recent US Environmental Protection Agency (EPA) release, there are more than 1600 Superfund sites across the USA [7]. The average cost of cleaning up one of these

sites is on the order of $25 million [8]. These Superfund sites are only a small sampling exemplifying the extent of pollution in the USA. The Department of Energy (DOE) along with other US government agencies are also responsible for billions of cubic meters of toxic contaminants affecting groundwater and soil [8]. Along these lines, the US military stock piles contain approximately 3×10^8 kg of munitions waste [9]. Generally these munitions wastes, such as RDX and HMX (hexahydro-1,3,4-trinitro-1,3,5-triazine and octahydro-1,3,5,7-tetratrinitro-1,3,5,7-tetrazocine, respectively), are difficult to break down and thus linger in soil and groundwater, posing significant health threats [9].

Agricultural activities also contribute to environmental contamination since they rely on nitrogen- and phosphorus-based compounds for crop fertilization. They also employ herbicides and pesticides to control weed and insect damage and to improve crop yields. The run off from these activities contaminates aquifers used for human water supply, damage coral reefs and contribute to algal blooms in costal areas and inland water bodies.

From the above discussion, it is evident that the impact of pollution by humans on the biosphere is extensive and perhaps overwhelming. However, as the scientific community assesses the problem(s) at hand, new approaches to remediation will emerge. This chapter outlines one possible approach to dealing with water and air pollution: photocatalysis, specifically using nanosized semiconductors. It is by no means an all-encompassing solution since there is no single approach that can address the problems noted above. However, it has a lot of potential in several specialized areas, especially where visible light illumination, a free energy source, can be utilized.

3.1.2 Scope

Since the topic of environmental remediation has been a focus in the scientific community for many decades, there are numerous reviews in this field. For a broader scope in the general field of environmental remediation, we refer the reader to some of these publications [2, 3, 10–12]. This is just a small sampling of the literature in this field and should not be regarded as a complete list.

Heterogeneous photocatalysis of pollutants is just one approach to environmental remediation and has also been reviewed by numerous authors [13–17]. TiO_2 has been the standard material for photocatalysis since its initial use in the photoelectrocatalytic generation of hydrogen reported by Fujishima and Honda in 1972 [18]. Since this time, the field of photocatalysis has grown significantly and over 2000 papers have been published [13, 17].

In recent years, with the avid interest in nanomaterials, there is also extensive literature addressing the effects of size on photocatalytic behavior [13, 14, 16, 17, 19–35]. There is good evidence showing that size effects play a role in photocatalysis. Accompanying a decrease in size is the increase in surface area with subsequent changes in surface chemistry, both of which are critical in photocatalysis. Most of the studies on nanosized photocatalysis for environmental remediation use TiO_2

as the photocatalyst material. However, different conclusions are often obtained because of differences in photocatalyst preparation, size determination and reaction rate methodologies used by various groups. There are large variations between findings where some groups claim enhancement in photocatalytic activity as a function of size and others claim the opposite.

With so many complete reviews in the field, it is difficult to present very new information to guide people in the field. However, we can present alternative perspectives that may aid in the thought processes needed to solve problems involving the search for new materials to enhance photocatalytic efficiency. Therefore, in this chapter, we focus on some important aspects relating the physical and chemical properties of nanoparticles to their photocatalytic behavior. Also, since most reviews in this field focus on TiO_2 as the photocatalyst, we will include detailed discussion of novel photocatalysts such as MoS_2 and other dichalcogenides, while discussing TiO_2 in a selective and critical manner.

Our technological focus concerns the application of nanosized photocatalysts in environmental remediation. Photocatalysis, the light-driven oxidation of organic and inorganic pollutants, can be an effective approach for the treatment of dilute, large-volume chemical contamination since visible sunlight is the most inexpensive and largest energy source available on the planet. Hence the limiting factor in this remediation method is the development of robust, inexpensive, environmentally benign photocatalysts.

This chapter emphasizes the key approaches to photocatalyst synthesis and characterization and the most salient research issues for the future. Following a review of the general field of environmental remediation, we give a historical background of the field with emphasis on the properties of titanium dioxide, since studies of the photophysical properties of this material have dominated the literature.

We then discuss selected photocatalysis studies and research findings associated with TiO_2. We next discuss research concerning layered semiconductors such as MoS_2 as photocatalysts. The ability to adjust the absorption onset and redox potential with size and surface chemistry in nanosized materials provide new opportunities in photocatalysis. Finally, we give examples of recent technical applications of TiO_2 semiconductors such as self-cleaning tiles and windows enabled by scientific research on TiO_2 photocatalysis.

3.2 General Field of Environmental Remediation

The best way to treat the pollution problem is to use conservation and preventive measures such as recycling to limit emissions into air and water sources. Steps are already being taken by industry to limit their emissions through optimization of manufacturing practices in addition to recycling and regeneration of chemicals, which also lead to economic savings in the long run [36]. Governments have also initiated legislation putting constraints on the permissible emissions. However, the fact remains that there is an abundance of toxic pollutants present in our water

systems, soil and air. As such, the challenge for environmental remediation is substantial.

Scientific and technical methods to mitigate environmental pollution rely on many approaches and vary for the cases of soil, water and air purification. The remediation approach chosen depends on the complexity and nature of the contaminated media and economic costs of the treatment. Often there are limited acceptable treatment approaches, as is the case with DDT contamination of sediments off the coast of California. Mixed wastes consisting of both radiological and chemical toxins are especially difficult and costly to separate and treat. The lack of documentation of the types and amounts of chemicals in many waste sites makes the remediation process especially costly.

The most common environmental remediation techniques traditionally used to treat large volumes of water intended for municipal water supplies include carbon adsorption, air stripping, oxidation through ozonation or chlorination and incineration, ultrafiltration and sedimentation [13, 34, 35]. These approaches are generally used for large-scale pollutant removal. The main drawbacks of these techniques are that most of them are transfer methods, where the pollutant is moved from one place to another or transformed from one phase to another. Activated carbon adsorption is a common and generally effective way to capture both airborne [such as volatile organic compounds (VOCs)] and waterborne contaminants. However, disposal of the saturated carbon is an issue [13].

On a smaller scale, water purifiers for home use utilize activated carbon absorbents to remove organic contaminants at the point-of-use. This point-of-use application is especially useful for rural locations lacking central waste treatment facilities. However, disposal of the concentrated activated carbon in a safe manner is still problematic.

Air stripping involves the removal of VOCs from wastewater, converting the pollutant from waterborne to airborne, necessitating the treatment of the resulting gaseous products. Outside Western countries, air stripping is often used to dilute airborne contaminants from industrial waste effluents. However, this approach again simply transfers the pollution problem from the liquid phase to the gas phase without destroying the chemical and is therefore banned in Western countries [13].

If this stripping process is accompanied by adsorption of the stripped pollutants and photocatalytic oxidation using a high surface area mesh containing a photocatalyst and a light source, the process is still very useful. Such an approach is being implemented commercially in Japan, as discussed in a later section.

Some of the other techniques such as incineration do lead to partial mineralization of the pollutants; however, they require high temperatures and again can result in unwanted gaseous by-products. Oxidation and chlorination are usually not capable of mineralizing all types of organic wastes completely. These techniques often lead to the formation of secondary pollutants, which can be just as toxic as the original pollutant.

Biological degradation processes in the presence of natural sunlight is an inexpensive, common route to degradation of certain noxious materials. The process is sometimes called the activated sludge process and relies primarily on bacteria [13]. It

is somewhat slow and efficient only at low toxin levels. It also requires that pH and temperature levels be controlled for the health of the bacteria. Also, many toxic chemicals, especially herbicides and pesticides, have low solubility in water and these pollutants are found primarily in sediments where sunlight does not penetrate. Therefore, although pesticides such as DDT have been banned in Western countries for decades, they are still ubiquitous in riparian and coastal environments.

Chlorination is a good process for killing viruses and bacteria; however, side chlorination reactions with organics present in the water may produce chlorinated species which are known to be carcinogens. In the presence of nitrates the chlorination efficiency decreases and nitrates are fairly common in many water sources.

Complete mineralization of organic pollutants is necessary and can be achieved by advanced oxidation processes (AOPs). This terminology is used to describe a group of photochemical techniques that lead to rapid and complete mineralization of a variety of organic compounds [13, 27, 37]. These techniques include UV ozonation, UV peroxidation (using H_2O_2) and heterogeneous photocatalysis. Use of ozone as an oxidant and disinfecting chemical is effective but can also produce unwanted by-products such as bromate ions, which are suspected carcinogens. Thus, current research in advanced oxidation techniques has emphasized combining photocatalysis with ozone treatment [38]. Generally, heterogeneous photocatalysis is preferable over the other two processes. It uses air (O_2) as opposed to O_3 or H_2O_2, which are relatively expensive reactants. Also, depending on which photocatalyst is employed, excitation wavelengths are in the near-UV (TiO_2 as photocatalyst) to visible range (sensitized or doped TiO_2, metal dichalcogenides as photocatalyst), whereas O_3 and H_2O_2 require short UV excitation wavelengths. However, all three methods are really only applicable to small-scale waste treatment sites where there is a low concentration of contaminant present. Large-scale waste remediation has yet to be demonstrated for these AOP techniques [13, 37].

Photo-oxidation of organic and biological pollutants can be implemented along with conventional methods using strong oxidants such as ozone and hydrogen peroxide. However, this is really only viable for high pollution levels. Typically, this treatment requires the concomitant use of short-wavelength UV lamps that increases the cost and restricts its application to point-of-use, small volume, water and air treatment. A strong motivation for the development of photocatalysts such as TiO_2 is to decrease economic costs by using at least a small portion of available solar light to photo-oxidize organic chemicals and convert them to harmless by-products in dilute volumes of water and air.

It has been shown that photocatalysts such as titania in combination with oxidants such as peroxide or ozone can also disinfect water by killing pathogens and oxidizing even stable contaminants. At present the costs of this combined advanced oxidation process are too great to be implemented on a large scale, but may be viable in point-of-use applications for smaller volumes in rural settings. Since titanium is the ninth most abundant metal [39] on Earth, the cost of producing titania photocatalysts is not likely to be the limiting factor for its application in environmental remediation.

Heterogeneous photocatalysis has a lot of potential in the field of environmental remediation, mainly due to the prospects of complete pollutant mineralization.

As such, it has been the focus of research for some time, with TiO_2 as the most studied photocatalyst. For economic viability, however, new photocatalysts which can use visible light must be developed. Even the UV efficiency of the best TiO_2 photocatalysts in water is only around 4%, which is too low to be economically competitive with existing approaches. The efficiency for gas-phase reactions, fortunately, is higher and therefore treatment of air contamination has been the first commercial application of TiO_2 photocatalysts.

Economic considerations in environmental remediation play a pivotal role in the decision to employ a particular technology. This point needs to be emphasized when comparing conventional treatment approaches with proposed methods such as photocatalytic oxidation using nanosized catalysts. However, these nanosized catalysts provide the possibility for accessing a much larger portion of the solar spectrum (i.e. the visible portion). This in itself would be very cost-effective, leading to a great step forward in the field of photocatalysis for environmental remediation.

No single remediation technology can be expected to address the diverse global problems outlined above. However, to be widely implemented any approach will have to possess economic advantages and be scaleable to deal with large volumes of contaminated solids, liquids and gasses. In the next 10–20 years it is possible that photocatalytic oxidation will play a significant role in environmental remediation if appropriate new materials such as nanosized photocatalysts can be developed and optimized for specific reactions. In the interim, as stated above, the best approach is resource conservation and recycling of materials so that waste generation is minimized. Societies and their governments should provide tax incentives to promote such approaches.

3.3 Photocatalysis

3.3.1 History and Background

A large surge in research and interest in the fields of photochemistry and photocatalysis was initiated by the oil crisis in the early 1970s, leading to a search for alternative energy sources. Of special interest was the generation of hydrogen from the photochemical splitting of water. Following the demonstration of photocatalytic water splitting on a TiO_2 electrode (in a photoelectrochemical cell) in 1972 by Fujishima and Honda, the field of photocatalysis and photoelectrochemistry flourished [18, 40, 41].

It is worth highlighting some of the initial achievements that participated in developing photocatalysis as a field. The photocatalytic properties of wide-bandgap metal oxide semiconductors such as ZnO and TiO_2 were first discussed in a tutorial article by Markham published in 1955 [42]. The first example of using light-driven catalysis in environmental remediation was that of Frank and Bard, who described the photo-reduction of cyanide ions in solution [43, 44]. The Ollis group reported the

first photo-oxidation of several types of organic toxins using TiO_2 in 1983 and this work initiated many other studies of this process [45, 46]. The first application of photocatalytic oxidation using TiO_2 to kill bacteria such as *Lactobacillus acidophilus*, *Saccharomyces cerevisiae* and *Escherichia coli* was reported in 1985 [47]. A group led by Fujishima demonstrated the first use of TiO_2 semiconductor powders to photo-oxidize HeLa tumor cells [48]. Eventually, extensive research on photocatalysis in the 1990s, particularly in Japan, led to the first report by Wang and co-workers of surface coatings of TiO_2 powders which had self-cleaning properties [49]. These surfaces were also superhydrophilic, which prevents water droplets from forming. This work would eventually lead to the commercial production of anti-fogging glass in Japan.

As the above historical timeline indicates, nearly all reviews of photocatalysis using semiconductors are dedicated to the properties of titania. [13, 17, 22, 28, 50]. TiO_2 is the most widely investigated material due to its ready availability, lack of toxicity and photostability. However, due to its wide optical gap of 3–3.2 eV, which precludes full use of the solar spectrum, research to develop alternative materials continues actively. Nevertheless, in Japan, commercial use of titania films in self-cleaning tiles and air filters has been shown to be economically viable. Self-cleaning construction materials such as tiles are sold by several Japanese companies and these developments are a subject of a recent review which we discuss in more detail at the end of this chapter [50].

The search for new materials or ways to improve visible light absorption of TiO_2 through sensitization and bandgap manipulation via doping continues. Improvements in the synthesis, characterization and processing of semiconductor nanoclusters is a major focus of this chapter. Recent developments in the last decade allow photocatalysis using new nanomaterials such as WS_2 and MoS_2, which in the bulk have near-IR bandgaps that are too small to drive many photo-oxidation processes. Nanoclusters of these materials have been demonstrated to have bandgaps which red shift into the visible region with decreasing size [51]. This permits the application of a wider range of semiconductors as candidates for viable photocatalytic processes in environmental remediation [52].

3.3.2 Definitions

Photocatalysis is the acceleration of a light-driven chemical reaction due to the presence of a chemical called a catalyst. By definition, the catalyst simply lowers that activation energy required for the process and the catalyst itself remains unchanged by the reaction [53]. Determining that a catalyst really remains unchanged after the reaction can be difficult and will be discussed in later sections. The specific photoreaction which is the subject of this chapter is the creation of electron–hole pairs in a semiconductor material. The electrons and holes interact with oxygen and water, resulting in the formation of free radicals such as dioxygen radical anions and hydroxyl groups. These radicals then begin a chain of "dark reactions", ultimately oxidizing chemicals in either the solid-, solution- or gas-phase environments, resulting in CO_2 and dilute mineral acids such as HCl. The complete reaction

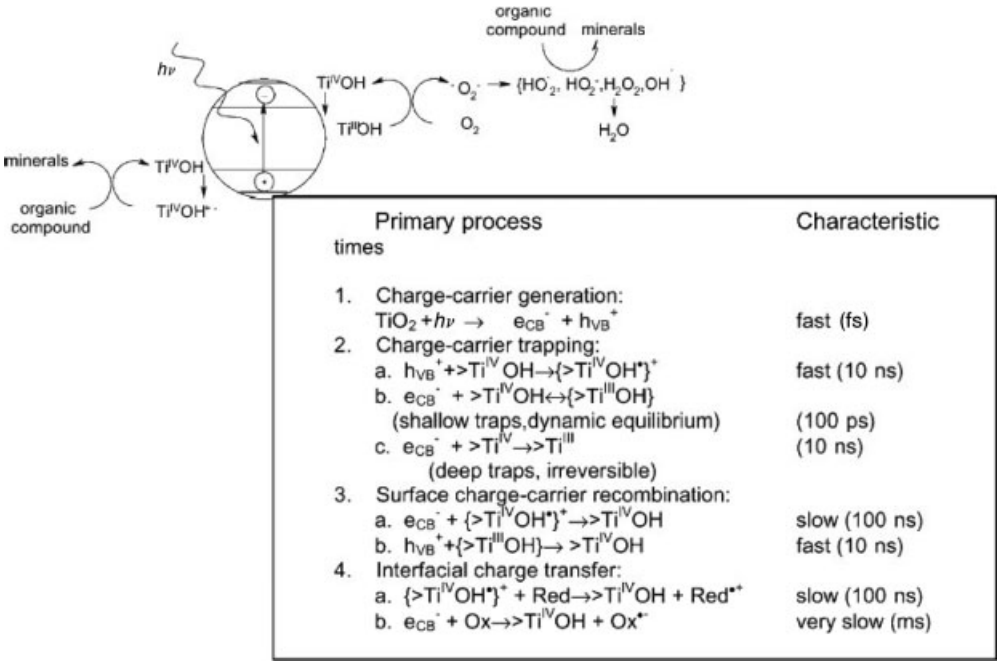


Figure 3.1 Light-activated photocatalysis in a TiO_2 particle. (Reprinted with permission from O. Carp, C. L. Huisman, A. Reller, *Progress in Solid State Chemistry* **2004**, *32*, 33 and A. Hagfeldt, M. Gratzel, *Chemical Reviews* **1995**, *95*, 49).

is often called total mineralization. A scheme for these processes using TiO_2 photocatalysts and the time scales associated with various steps is shown in Figure 3.1 [13] and Figure 3.2 [37]. Note the key role of hydroxyl radicals in both figures. Also, depending on the chemical, there are many possible rate-limiting steps in the photocatalysis reaction.

Laser photolysis and transient absorption experiments have allowed the time identification and time scales shown in Figure 3.1 to be determined [54]. Both trapped holes and electrons have distinct absorbance spectra in the visible region and the decay rate of these absorbance features following photoexcitation allows the carrier transfer kinetics to be followed. The selectivity and efficiency of these processes depends on many factors that we review in more detail subsequently.

3.3.3
Well-Known Example – Water Splitting Reaction

A light-driven electrochemical photocell to split water into hydrogen and oxygen was first described by Fujishima and Honda in 1972 and was a key result motivating interest in titania as a photocatalyst [18]. Although the oxidation and reduction potentials of TiO_2 are sufficient to drive this reaction, the kinetics are difficult,

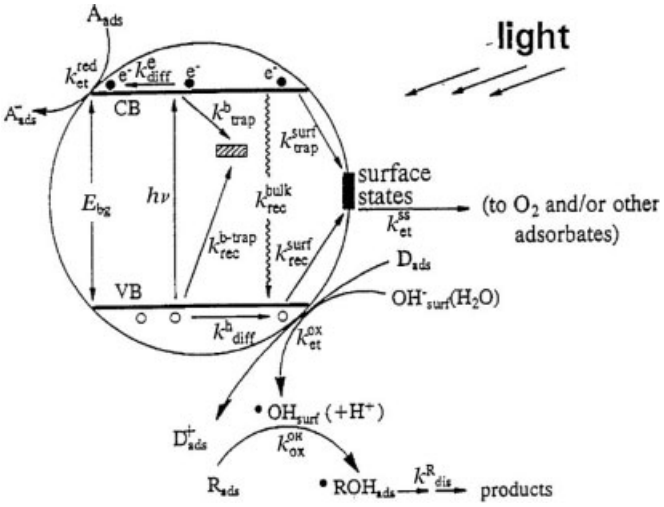


Figure 3.2 Photophysically and photocatalytically possible events on a heterogeneous photocatalyst. (Reprinted with permission from N. Serpone, *Solar Energy Materials and Solar Cells* **1995**, *38*, 369).

requiring four electrons to be transferred. Hence the efficiency of the process is low, estimated at less than 0.1%. This low UV light efficiency combined with the fact that TiO_2 absorbs only small amounts of the total solar spectrum means that most of the absorbed light simply generates heat. A lesson learned from these experiments is that although the thermodynamics for photocatalytic oxidation may be favored for a chosen chemical, competing pathways may limit the efficiency and these pathways must be eliminated by careful design of the nanocatalyst (see Section 3.4). It is unlikely that a single type of photocatalyst will suffice for all oxidations of interest.

Many later reports of direct, fully catalytic water cleavage are doubtful since chemicals present during the preparation of the particles (alcohols in the case of the TiO_2 synthesis), can act as electron donors, allowing hydrogen evolution to occur. However, the electrons are not furnished by the conduction band of the TiO_2 in this case. One must also be cautious regarding the contributions of reactor cell leaks to the observation of oxygen in these reactions. Mills and Le Hunte provide a good critique of these experiments [17].

3.4
Design Issues for Environmental Remediation Photocatalysts

3.4.1
Introduction

The factors affecting photocatalysis when using bulk semiconductors such as TiO_2 will also need to be considered when working with nanosized photocatalysts.

The important processes outlined in Figure 3.1 can be viewed as design parameters requiring optimization in order to have an efficient photocatalytic reaction.

3.4.2 Charge Separation

For the size range of nanoparticle photocatalysts discussed in this chapter, photogenerated charges rapidly diffuse to the particle surface and are trapped. Many TiO_2 surfaces, for example, have oxygen vacancies that function as electron traps. However, the electrons in these traps have lower energies than the initially photoexcited conduction band electrons and so are less likely to be transferred to oxidants. It is possible to modify a nanocluster surface to improve the electron transfer process as we discuss in the synthesis section of this chapter (Section 3.6). A longer electron lifetime improves the generation of hydroxyl radicals from surface Ti–OH species. Spatially separating the electrons from the holes by introduction of metal or semiconductor islands on the TiO_2 surface is a good approach for increasing the electron lifetime and the photocatalytic efficiency since the electron transfer is usually the rate-limiting step. This is why most reactions are conducted in oxygen-saturated water since oxygen is a good electron scavenger.

Several approaches to improving charge separation in titania photocatalysts have been suggested and implemented. Deposition of metal islands such as Ag or Pt on TiO_2 clusters has been shown to facilitate the electron charge-transfer process, which is kinetically the slowest redox event [55, 56]. The metal islands are thought to function as sinks for electrons, reducing the recombination of the photogenerated charges. The presence of Ag on the TiO_2 was shown to enhance dye photo-oxidation compared with the TiO_2 particles alone in a batch slurry-type reactor. Orlov and coworkers [55] used a flow-through reactor design with immobilized Au-coated TiO_2 particles to improve photoactivity for two common pollutants, methyl *tert*-butyl ether (MTBE) and 4-chlorophenol.

In general, not all dopants will enhance charge separation and the subsequent quantum yield of all reactions. For example, Fe^{3+}, which is one of the most common dopants, is capable of enhancing the photodegradation of certain pollutants (CCl_4, $CHCl_3$ [57, 58]) while negatively impacting the photo-oxidation of other pollutants (e.g. 4-nitrophenols [58, 59]). It has also been reported that there is an optimum dopant concentration which depends on TiO_2 particle size [58]. Particularly for Fe^{3+}, the necessary doping level was found to increase with decreasing TiO_2 particle size where sizes $\leq$11 nm showed enhanced behavior for photodegradation of $CHCl_3$ [58]. However, it is not clear that the particle size and size distribution were controlled, so specific conclusions based on size may be difficult.

Alternatively, coupling two semiconductors can be used to improve charge separation by transfer of one of the charges, say the electron, from the semiconductor with the higher potential to the lower one. This leaves the hole on the original semiconductor and can achieve better spatial charge separation. An example of this is TiO_2–CdS, where CdS can absorb visible light and transfer its electron to TiO_2 [60].

3.4.3
pH of Solution

The solution pH has an important effect on the adsorption of organic molecules on the catalyst surface, particularly if the chemical to be oxidized has a net charge or is very polar. For example, since the oxidation depends exponentially on the distance of the chemical from either the holes or the surface hydroxyl radicals, if the organic molecule has a positive charge (such as a quaternary ammonium salt), low pH values, where a metal oxide like TiO_2 will be positively charged, are less favorable than slightly alkaline values (e.g. $pH \approx 9$). The opposite will hold for anionic contaminants. For neutral organic substrates, the more important effect of pH is to increase the activity at higher pH values by producing a more hydroxylated surface. This sort of surface is more effective in trapping photogenerated holes. As an example, the pH corresponding to neutrality in Degussa P25 TiO_2 is around 6.2 [17], so for values less than this, the surface is charged and cationic chemicals are repelled, whereas for values greater than 6.2, anionic species are repelled. Despite these facts, most photocatalysis reactions using TiO_2 have a weak pH dependence.

3.4.4
Charge Transfer

The initial carrier generation process important in photocatalysis is the photogeneration of electrons and holes, which occurs on the femtosecond time scale [24], as shown in Figure 3.1. The carriers then diffuse to the cluster surface in less than 10 ns and are trapped. The holes can be trapped at Ti(IV)–OH^+ sites and the electrons at Ti(III)–OH sites in around 10 and 0.1 ns, respectively. Interfacial charge transfer of the Ti(IV)–OH^+ holes to adsorbed organic molecules and Ti(III)–OH to molecular oxygen then can occur on time scales of around 100 ns (holes) to milliseconds. The slow time for the latter means that methods to accelerate the transfer of electrons are important to minimize undesired surface trapped carrier recombination.

3.4.5
Presence of Simple and Complex Salts

Metal salts, especially simple ones such as NaCl, are present in most wastewater and many natural aquifers. As such, they affect both photocatalytic activity and selectivity. Most common anions found in water systems such as chloride, nitrate, phosphate and sulfate decrease the catalytic oxidation of organic compounds [13, 61–63]. The presence of ions during water purification using photocatalysis is an important research topic since some present and future sources of water will require desalination and this source of salt water also contains organic pollutants from off-shore dumping, etc. If brackish water must first be completely desalinated to avoid poisoning of the photocatalyst, the cost of environmental remediation of organic contaminants will not be economically competitive with carbon adsorption

approaches that do not have this requirement. Hence the ability to formulate a photocatalyst which is somewhat salt tolerant is important.

Certain anions, such as sulfate and phosphate ions, can form reactive species such as H_2PO_4, which are good oxidants, thus improving the photo-oxidation rate. However, their other effect is to adsorb on the photocatalyst surface, which can block active catalytic sites. This adsorption is strong enough that washing with an alkaline solution is necessary to restore the photocatalytic activity. For example, phosphate adsorption at low concentrations of only 1 mM has been reported to reduce the photo-oxidation of organics such as ethanol and aniline by around 50% [61].

The effect of cations on the photoactivity is more varied and dependent on the type of organic molecule considered in addition to the metal ion type and concentration [13, 31, 64, 65]. At high concentrations, their presence is mostly unfavorable since they may undergo photo-reduction by the photocatalyst itself and deposit on the surface, blocking active sites. By maintaining the reaction media at low pH, the positively charged photocatalyst surface will repel cations, lessening their negative impact.

Photo-reduction, however, can be useful in improving activity in the case of metals such as Pt, Ag or Pd, provided that full coverage of the catalyst surface does not occur. Instead, metal islands are formed, which function as electron storage sinks and facilitate charge separation and transfer. This topic is discussed in more detail in the section on synthesis and photocatalyst surface modification. Certain types of metal ions such as Cu(II) and Fe(II) can enhance photocatalysis by limiting carrier recombination through trapping either electrons or holes. The facile ability to change oxidation state is critical to this function. However, many metals, including Fe, are prone to hydroxide formation and deposition of these metal hydroxides on the photocatalyst surface is almost always detrimental. This means that an optimal metal ion concentration in the range 100–500 ppm (0.01–0.05%) should be determined and used for a chosen photocatalytic reaction.

There is a close connection between the pH of the solution and whether an ion such as Cu(II) accelerates photocatalysis or not. For example, in the photo-oxidation of aliphatic acids by TiO_2 in the presence of Cu(II), it was reported that for pH values <4 Cu(II) forms complexes which are active intermediates, trapping holes, whereas copper diacetate complexes form at higher pH values and poison the photocatalyst [66, 67].

In our work on the photo-oxidation of pentachlorophenol, we demonstrated that nanosized TiO_2 and SnO_2 synthesized by ambient temperature hydrolysis, followed by dialysis to remove unwanted by-products, was quenched by the addition of simple salts such as NaCl at concentrations of only 10 mM [68]. These low concentrations of NaCl have a poisoning effect on the catalyst, although it is mild in this case. Similar observations have been made in field tests of TiO_2 where deionization of the aqueous waste stream has been found to be necessary [69]. This is an important consideration when estimating costs of remediation using photocatalysts, since desalination is a costly process. In other studies by Barbeni *et al.*, no inhibition of the PCP photocatalysis by TiO_2 in the presence of NaCl at 1 mM occurred, so the presence of low level salts is acceptable [70].

3.4.6
Effect of Surfactants

In our work, we investigated not only the effect of simple salts on the rate of photo-oxidation of PCP, but also that of complex, surface-active salts/surfactants such as dodecyltrimethylammmonium chloride (DTAC), a cationic quaternary ammonium salt with an organic, hydrophobic tail group. Using this surfactant, we could isolate the effects of the common anion chloride from that of its ion pair, NaCl in one case and dodecyltrimethylammonium chloride for the other. To our surprise, this complex cationic surfactant significantly accelerated the rate of photo-oxidation of PCP by Degussa P25 titania in addition to our nanosized photocatalysts [68]. Two explanations of these experiments are reasonable:

1. Binding of PCP at photocatalytic surface sites by the presence of a bulky surfactant such as DTAC is not the rate-determining factor in the photo-oxidation of PCP by TiO_2. Instead, this surfactant either aids hole or electron transfer to PCP or electron-accepting species such as oxygen. Possibly free hydroxyl radicals are formed, which can diffuse to the PCP so that direct hole transfer to bound PCP is not required.
2. The sodium cation in NaCl is mainly responsible for the strong quenching of the photo-oxidation. This contradicts other studies cited above, however.

The significant enhancement (note the logarithmic scale on the vertical axis) due to the presence of a surfactant is shown in Figure 3.3. Note that at similar concentrations of NaCl the rate of PCP photo-oxidation is significantly slower. A common argument against using surfactant-stabilized nanoclusters as catalysts is that the surfactant will block access to the cluster surface and thus poison the catalyst. Our study shows this is not the case for DTAC. We also discovered that substitution of a bromide counterion for the chloride in DTAC does not increase the photo-oxidation of PCP as greatly as for DTAC. It is also worth noting that the surfactant peaks in the

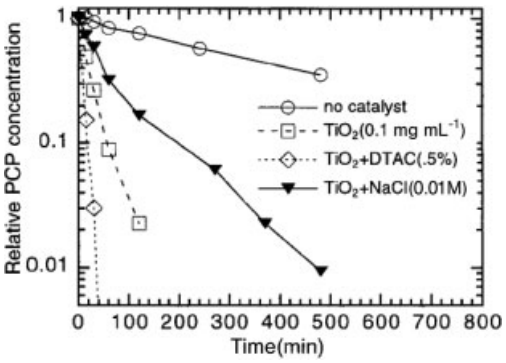


Figure 3.3 Relative PCP concentration vs. irradiation time, showing the effect of surfactant type on photocatalytic destruction of PCP. (Reprinted with permission from J. P. Wilcoxon, *Journal of Physical Chemistry B* **2000**, *104*, 7334).

chromatograms [high-performance liquid chromatography (HPLC) was used to monitor reactant and product species throughout all photocatalysis experiments; see Section 3.6.2) did not decrease in area during the photo-oxidation of PCP, indicating that both DTAC and DTAB are resistant to photo-oxidation.

In studies of the reduction of nitrate ions by nanosized CdS which was surface stabilized by charged organic thiols, Korgel and Monbouquette also found significant photocatalytic activity despite the strong surface binding of thiols to CdS [71]. The best way to rationalize this result is to remember that the surfactant used to stabilize a nanocluster in solution is in dynamic equilibrium with the surface. By this we mean a free stabilizer is constantly being exchanged with bound surfactant, thus allowing other chemicals transient access to the photocatalyst surface. So, binding of other chemicals is not statically blocked by the surfactant. In fact, the surfactant might increase the local concentration of organic chemical near the photocatalyst surface since its hydrophobic tail groups have much more favorable interactions with non-polar organic chemicals than does the continuous phase, water. An enhancement of the local concentration of pollutants will increase the reaction rate.

3.4.7
Effect of Solvent and Dissolved Oxygen

Most photocatalysis studies have been performed in aqueous solution. An observation common to photo-oxidation of most organics in water is that dissolved oxygen plays a vital role in the oxidation process by forming first dioxygen radical anions by abstraction of the electron formed upon photoexcitation of the semiconductor and then peroxides by reaction with the water [72]. Many researchers deliberately aerate their photocatalyst slurry suspensions in order to optimize the photo-oxidation process and claims have been made numerous times that nearly total quenching of the photo-oxidation process occurs when inert gas purging is employed [70]. Thus, the use of aprotic organic solvents should cause a severe quenching of the photo-oxidation due to both the lack of OH radicals and reduced oxygen levels because of decreased oxygen solubility.

Molecular oxygen has a strong affinity for electrons and its presence should reduce undesired carrier recombination, thus increasing the effectiveness of the photocatalyst [73]. However, there is evidence that at high concentrations of oxygen the photocatalytic activity is reduced, possibly due to changes to the TiO_2 surface such as hydroxylation, which could interfere with the adsorption of the organic on catalytic sites [74].

If water and dissolved oxygen play critical roles in the photocatalysis process using TiO_2, then carrying out these reactions in a polar, but aprotic solvent such as acetonitrile (ACN) would be expected to quench the process and could alter the photo-oxidation pathway. However, complete quenching is not observed in anaerobic catalytic photo-oxidation of pentachlorophenol by Degussa TiO_2 in the aprotic solvent ACN (filled triangles, Figure 3.4). Uncatalyzed photo-oxidation in ACN, which occurs by a different mechanism, is quenched, however (open diamonds, Figure 3.4) [68]. In Figure 3.4, it can be observed that although the rate of catalytic photo-oxidation is 2–5

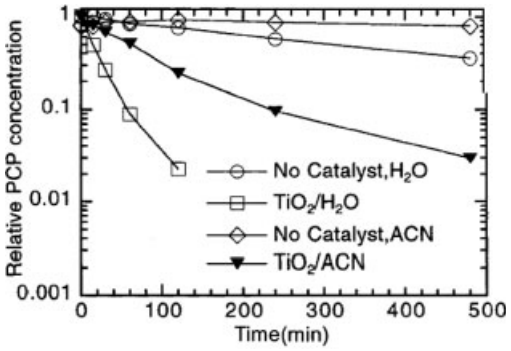


Figure 3.4 Relative PCP concentration (10 ppm at $t = 0$) vs. irradiation time using a xenon arc lamp. Significant photocatalysis occurs in oxygen-free solutions of acetonitrile (ACN). (Reprinted with permission from J. P. Wilcoxon, *Journal of Physical Chemistry B* **2000**, *104*, 7334).

times slower in nitrogen = purged ACN than water, photocatalysis does occur. Also, the mechanism for PCP photo-oxidation in water and in ACN is similar, as verified by the presence of common elution peaks for the main photo-oxidation intermediate in the HPLC traces for both solvents.

3.4.8
Light Intensity

For low light intensity, up to around 25 mW cm^{-2} at 365 nm (depending on reactor geometry and nanoparticle concentration), the rate of electron–hole transfer is fast enough to allow each photocatalyst particle to absorb photons and create electron–hole pairs. Hence the rate is first order or proportional to the light intensity. At higher intensities, the existence of an electron–hole pair on a particle prevents the photogeneration of another one on the same particle and full utilization of the incident photons cannot be realized. The photocatalysis rate is found to be proportional to the square root of the intensity in this regime [73]. The practical aspect of this observation is that slurry-type batch reactors are best operated at low light intensity such as found in natural sunlight, which provides 2–3 mW cm^{-2} in the absorbance region of TiO_2. It also indicates that schemes for concentrating sunlight to higher fluxes to increase the rate of photo-oxidation will be of limited utility.

3.5
Potential for Nanomaterials in Environmental Remediation

3.5.1
Introduction

Nanosized materials for photocatalysis and photo-oxidation have evolved from conventional bulk metals and semiconductors to colloidal (large cluster, ~10–100 nm

in size) materials and, most recently, to nanosized materials or small clusters (1–10 nm). We distinguish these classes of materials by the way in which physical properties depend on surface area and the significant changes in electronic and photocatalytic behavior with decreasing size. In the nanosize regime, quantum confinement and cluster surface chemistry dominate materials properties.

Nanoparticles of a given material can exhibit very different behavior to their bulk counterpart. For example, the potential for oxidation and reduction could become stronger with decreasing size, so more types of organic or inorganic materials can be photo-oxidized or photo-reduced at faster rates. Since the specific surface area is so much greater, more economical use of material reduces costs. Also, the large surface to volume ratio means that subtle changes to the surface due to addition of other atoms or molecules can lead to dramatic alterations of physical properties such as substrate binding. A number of fields, including magnetism, luminescence and renewable/alternative energy to sensors/taggants as well as photocatalysis, will benefit from capitalizing on the surface–size relationship. The unifying theme in these disparate fields is controlled alteration and possible enhancement of physical properties due to quantum size and cluster interface effects.

Semiconductor surfaces in small clusters can be considered defective relative to perfect micron-sized crystals of the same material. This can be a major advantage in photocatalysis since either electron or hole traps at the surface will determine the recombination time, in many instances reducing the recombination rate and increasing the probability of carrier transfer to an adsorbed organic molecule. The effect is that the efficiency for catalytic photo-oxidation can increase with decreased size in addition to being strongly influenced by substitutional ions such as iron [58].

In the last 20–30 years, scientific interest in the properties of nanosized materials and the need for new materials for photocatalysis applications such as environmental remediation and renewable energy has motivated research in both fields. As mentioned above, several reviews relating nanoparticles and photocatalysis have already appeared. The effects of size quantization using colloidal materials was reviewed in 1988 and 1989 by Henglein [75, 76]. A large motivation for new developments in colloidal science was the investigation of particle size effects on optical properties and photocatalysis. Complementing Henglein's viewpoint, other reviews provided a detailed overview of colloidal semiconductors and their photochemical properties [23]. The photo-redox reactions in nanocrystalline systems were subsequently reviewed by Hagfeldt and Gratzel [24]. This early research by Henglein and Hagfeldt and Gratzel (along with others) emphasized the use of optical techniques to learn about the size-dependent behavior of semiconductors. Techniques such as transient absorbance to monitor electrons and holes and their recombination kinetics guided future research in photocatalysis. Several reviews (1995–98), by Howe [16], Hoffmann *et al.* [26] and Linsebigler *et al.* [40], analyzed the effect of carrier confinement and size effects in TiO_2. Other reviews have focused more specifically on the use of nanomaterials in environmental and energy applications [30, 32, 35]. Beydoun *et al.* presented a broad overview of the nanoparticles in photocatalysis [21]. They reported many instances where nanoparticles, especially TiO_2, enhance the photocatalytic process. It is difficult to separate the effects of carrier

confinement due to decreased particle size from surface chemistry changes which occur in the same size range. We will give examples of both types of size-dependent changes in this chapter.

3.5.2 Nanomaterials and Advantages in Photocatalysis

3.5.2.1 Semiconductor Nanoclusters

As in the case of their bulk counterparts, nanosized semiconductors have a range of forbidden energy states whose most important effect for photocatalysis is to allow the creation of valence band holes and conduction band electrons using light. Fast recombination of these photoexcited holes and electrons is an undesirable effect for photocatalysis. Hence nanosized semiconductors with strong photoluminescence are poor candidates as photocatalysts. Fortunately, the diffusion time for electrons or holes in nanoclusters is so fast that the normal bulk recombination process is not important. The most important effect is trapping of the carriers at the surface. This trapping can be influenced by cation or anion vacancies [e.g. Ti(IV) or O^{2-}] at the surface which are deliberately or accidentally introduced during the synthesis. It is also possible to use surface-active molecules such as surfactants, certain ions and the general solvent environment to change the carrier lifetime at the cluster surface and improve the oxidation process for adsorbed organic chemicals. Deposition of metal islands such as Pt that improve charge separation by serving as electron traps is another avenue for enhancing charge separation. Each of these synthetic strategies will be illustrated in more detail later with examples from the literature. Many of these effects are simply due to the different geometry and chemical bonding in a cluster surface at small dimensions (1–3 nm) and would occur even in the absence of quantum confinement of the electron–hole pair.

3.5.2.2 Quantum Confinement

The quantum size effect was first discussed by Frohlich in 1937 [77, 78]. The concept of quantum confinement that he described evolved from a series of observations where light interactions with certain materials varied as a function of form. Frohlich discussed quantum size effects in the context of observed differences in light-scattering behavior between small particles and corresponding thin films of the same material [77, 78]. The quantum confinement model is similar to the particle-in-a-box model where an electron is confined in a finite volume in space and the number and energy levels of the possible states is determined by the confining potential. In an aromatic molecule, for example, this electron confinement or delocalization is determined by the degree of bond conjugation setting a length scale over which the electron is confined by the electrostatic potential. In such a system only a discrete number of energy states are possible and as the physical size or delocalization of the electron increases so does the number of states, eventually becoming so closely spaced as to be essentially continuous.

In a semiconductor, the states which make up the valence and conduction bands are also so closely spaced as to appear continuous. However, as the semiconductor

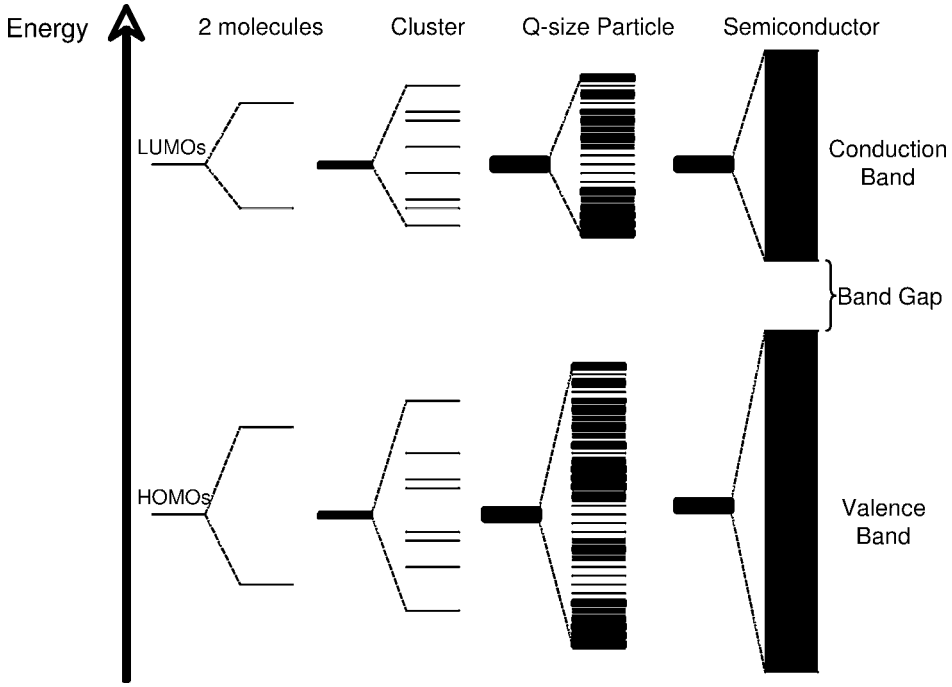


Figure 3.5 Molecular orbital schematic showing the development of bands as a function of increasing particle size. (Reprinted with permission from D. W. Bahnemann, *Israel Journal of Chemistry* **1993**, *33*, 115).

size decreases to a length scale comparable to the confinement potential of the electron–hole pair or exciton, (the Bohr radius), the particle size strongly influences the exciton energy, increasing the energy as the size decreases. Thus, the onset of light absorption shifts to the blue with decreasing size. The continuous bands become discrete states (as in a molecule) and discrete absorption bands emerge (Figure 3.5). Since such a nanosized semiconductor lacks the long-range translational symmetry of its bulk counterpart, the distinction between light-driven direct (momentum conserving) and indirect transitions (requiring phonon assistance) between the highest energy states in the valence band and the lowest energy states in the conduction band is lost. This means that optical transitions with a low probability or oscillator strength in an indirect semiconductor such as silicon or MoS_2 are likely and the extinction coefficient is strong in nanoclusters of these materials. As the semiconductor size approaches molecular dimensions, it is most appropriate to model the valence band as consisting of the highest occupied molecular orbital(s) (HOMO) and the conduction band as the lowest unoccupied molecular orbital(s) (LUMO). In order to model these states properly, chemical bonding at the surface must be understood and included. Ligands bound to surface atoms should also be included. This is nearly impossible to do properly, so quantum confinement

predictions of energy shifts with size are qualitative, rarely agreeing with experiments for sizes less than 4–5 nm.

Discussions of quantum confinement have been the subject of several reviews [13, 15, 17, 34, 75, 76, 79–81]. Many reviews have considered the effects of quantum confinement on electrical, optical and photocatalytic properties. In particular, Brus modeled the shift in redox potential with decreasing particle size for CdS and InSb [82]. Brus's model is based on a description of bulk state behavior in the limit of small crystallite size. It uses approximations from band theory of perfect lattices and assumes a crystal structure matching that of the bulk material. As a result, it could not account for surface states. In general, this model predicts fairly mild quantum effects as a function of size, especially when the crystallite sizes are larger than 5 nm.

Some of the predictions are based on the relationship between the particle size and the effective mass of the exciton [i.e. the effective mass model (EMM)]. There is some controversy as to whether the effective mass model should be used to make quantitative estimates of particle sizes based on absorption spectra shifts and the use of the same exciton effective mass as that for the bulk counterpart [79]. The actual value for the exciton effective mass is itself dependent on size and shape.

Both cluster size and shape influence the energy shifts observed in nanoparticles. A rod-like cluster will have different confinement potentials for the longitudinal and transverse directions, for example. Also, the confining potentials for electrons and holes are typically different, as reflected in the different effective mass of the electrons and holes or curvatures of the potential surfaces for the valence and conduction bands. In most materials, electrons are much more mobile than holes and the conduction band will shift more strongly than the valence band with decreasing size. As noted by Wise [83], the particle size at which confinement effects become prominent is greatly influenced by whether or not the charge carrier mobilities are similar. Certain materials such as PbS and PbSe satisfy this requirement, but most others such as CdS and TiO_2 do not. Chemically, this is due to the unequal sharing of electrons in polar semiconductors such as metal oxides and sulfides.

The blue shift of the bandgap absorption onset impacts the design of nanosized photocatalysts in several ways and the length scale at which such effects are first observed is materials dependent. This increase in the bandgap as a function of decreasing size allows the possible use of a wider range of materials as photocatalysts. Many of materials are not catalytically active in their bulk form. For example, semiconductors such as MoS_2 and WS_2, which have near-IR absorption onsets, can be shifted into the visible region by decreasing the cluster size. Both valence and conduction bands shift in energy, the valence band becoming more positive and thus the holes more strongly oxidizing. The photoexcited electrons in the conduction band are shifted to more negative potentials, also improving their transfer to such species as molecular oxygen. In general, the greater the effective mass of the holes compared with the electrons, the larger is the shift in the conduction band energy with decreasing size. The amount of the shift can be estimated by measuring the rate of hole or electron transfer to fluorescent hole or electron acceptor molecules as a function of cluster size [84].

The interplay between quantum size effects and surface chemistry changes with decreasing particle size can be optimized in order to enhance the activity of a particular catalytic reaction. However, in order to do this, there must be fairly good synthetic control over particle size and monodispersity. This size control is not available through all synthesis techniques and seems to be difficult in the case of TiO_2. As a result, many conclusions relating photocatalytic activity to size may seem contradictory. More often than not, other factors such as photocatalyst environment or synthetic protocol dominate the changes in the observed reactivity.

3.5.2.3 **Surface Chemistry**

In addition to increasing the strength of the confining potential and shifting the light absorption onset, decreasing particle size results in a larger fraction of atoms which are in surface sites with bonding differing from the interior atoms. For small clusters (1.5–2 nm), ~70–80% of all the atoms reside on the surface [85]. Sometimes cluster surface structures are considered defective. However, these defects may be useful for photocatalysis. For example, a metal oxide defect structure (e.g. TiO_2) with oxygen vacancies can enhance both the adsorption of water on the surface and the dissociation rate of water into hydroxyl groups and protons. This water dissociation process requires the presence of paired acid–base sites that are situated in close proximity. Surface sites with acid character such as titanium cations initially bind water molecules, whereas neighboring sites with basic characteristics such as Ti–O–Ti structures can accept a proton from the water molecule. In nanosized materials, the structural arrangement of interior atoms that is simply the phase structure as determined by diffraction methods, is less significant than the surface chemistry. For example, in the case of titania, the surface hydroxyl groups play a critical role in initiating the oxidation process by both influencing the adsorption of organic chemicals and aiding the dissociation of adsorbed water into hydroxyl radicals and hydrogen ions. Free hydroxyl radicals are very good at attacking and oxidizing a wide range of organic groups and, with a potential of around 2.8 V, are more effective than any radical except fluorine. Even surface-bound hydroxyl radicals with a potential of 1.5 V can be effective oxidants [86].

Synthetic changes of the surface properties of nanosized catalysts can also be used to modify redox potentials and substrate binding, independent of quantum confinement effects. Kamat and Meisel have taken advantage of this effect through alterations of the interface of nanosized TiO_2 particles deposited on Au [27]. The intimate contact between the metal and the surface of the nanosized TiO_2 was reported to be crucial in improving charge transfer. Additional examples of the effect of ions and metals on the surface of semiconductor photocatalysts will be given later.

3.5.2.4 **Other Unique Materials Properties**

Since most of the atoms in a nanocluster can reside at accessible surface positions, very small changes in the chemistry of the cluster surface can significantly alter its photocatalytic properties. For example, in a 50–60 atom cluster with a diameter of around 1.6 nm, the addition of a single foreign atom can change the interatomic spacing and thus the binding energy between the catalyst nanoparticle and an organic

substrate. The restructuring of a highly curved nanocluster surface due to other atoms, surfactants or ions also fundamentally alters the optical and binding properties compared with micron-sized powders, allowing a wider range of materials to be utilized as photocatalysts. An example emphasized in this chapter is nanosized metal dichalcogenides such as MoS_2.

3.5.2.5 Importance of Nanocluster Photostability

There are only certain types of materials which have suitable stability as nanosized clusters to be considered as useful photocatalysts. One group includes materials which are already oxidized, such as metal oxides. Even among this group, only titania is readily regenerated during the photocatalytic oxidation process, thus acting as a true catalyst. Other oxides such as ZnO are partially consumed upon use as photocatalysts [87].

Metal sulfides such as CdS or ZnS are not generally photostable in the presence of either water or oxygen. The reason for this is the polar nature of their bonds which make them subject to oxidation by the holes created upon photoexcitation. However, in certain specialized reactions such as the oxidation of H_2S, the presence of anions such as HS^- at the hole sites of the cluster surface prevents the holes from simply oxidizing the lattice. This allows typically unstable polar semiconductors such as CdS to be used as photocatalysts [60].

A class of photostable materials are metal dichalcogenides such as MoS_2 whose stability is due to their anisotropic structure. These materials have a two-dimensional layered structure with catalytically active metal edge sites located at the cluster surface protected by adjacent sulfurs (Figure 3.6). Photoexcitation occurs primarily within the metal d-bands and doesn't weaken the bonds responsible for the lattice. The resulting electronic structure makes them quite photostable in water.

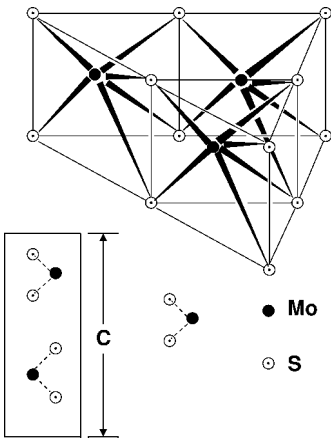


Figure 3.6 Schematic diagram of the 2D sandwich structure of MoS_2. (Redrawn with permission from J. P. Wilcoxon, P. P. Newcomer, G. A. Samara, *Journal of Applied Physics* **1997**, *81*, 7934).

3.6
Nanoparticle Synthesis and Characterization

3.6.1
Introduction

Research concerning nanocluster optical properties and photocatalysis has focused on a limited number of readily available materials such as TiO_2, ZnO or ZnS. Development of new synthetic methods should allow studies involving a much broader range of materials, increasing the efficiency and selectivity of nanosized photocatalysts. Furthermore, the ability to form alloys and later alter the cluster surface properties will give researchers further control of the photophysical properties.

Both solution-based and gas-phase syntheses of nanoparticles are used to prepare photocatalytic materials. Solution-based methods for the formation of metal oxide semiconductors rely primarily on the base- or acid-catalyzed hydrolysis of metal organic precursors such as titanium isopropoxide. These syntheses are generally executed by rapid mixing or co-injection of two solutions, one containing the metal organic in a polar organic and the second solution containing the acid catalyst. The resulting colloids are stabilized by charge in water at low pH. Some additional details about size and nanostructure control in these formation approaches are provided below in our discussion of TiO_2 synthesis.

Many commercial micro- and nanosized metal oxide powders are formed by proprietary methods which generally involve gas-phase hydrolysis of inexpensive compounds such as $TiCl_4$ (fumed silica powers are prepared in this manner from $SiCl_4$ also). A good example of a photocatalytically active material prepared by this method is Degussa type P25 TiO_2 (primary particle size $\sim$25 nm), which consists of roughly 80% anatase (the allegedly photoactive form) and 20% rutile phase. Because so much of the literature uses this material, we discuss some of its special properties in more detail below. It is generally more active and broadly effective for photo-oxidation than most other TiO_2 formulations [13].

Since Degussa P25 TiO_2 has both high activity and commercial availability, it has been employed in the majority of studies of photocatalytic activity described below. This material is formed by the high-temperature (>1200 °C) flame hydrolysis of $TiCl_4$. $TiCl_4$ is a very inexpensive but air-sensitive chemical precursor, so the presence of oxygen and hydrogen during the reaction produces both TiO_2 and HCl. The latter by-product is removed by treating the TiO_2 with steam. Nanosized Degussa TiO_2 has a BET surface area of $50\,m^2\,g^{-1}$ and consists of submicron- to micron-sized aggregates of 10–40 nm primary particles Another commonly used commercial TiO_2 photocatalyst, Sachtlebem Hombikat UV 100, consisting only of anatase, also has high photoactivity, which is believed to be due to a fast interfacial electron transfer rate resulting from its smaller particle size [54].

In general, gas-phase aggregation methods have a production cost advantage over solution-based methods since they use less expensive chemical precursors. However, the colloids or clusters formed in the gas phase cannot be dispersed as stable sols in

water. Degussa P25 powder, for example, consists of "popcorn-ball"-like 3–5 μm sized aggregates of roughly 20–30 nm diameter spheres and hence is not soluble in water, instead being used exclusively as a suspension or slurry in photocatalysis experiments. The slurry must be agitated (usually with a magnetic stirrer), under most conditions. Light does not penetrate into this opaque solution compared with a transparent solution of fully dispersed nanoclusters.

Solution-based methods for preparing the third class of semiconductors to be discussed, metal sulfides, are typically based on reaction of metal salts (e.g. $MoCl_4$) in non-aqueous solutions with a sulfur source (e.g. H_2S, NH_4S) under anaerobic conditions in the presence of a surfactant stabilizer [51]. A special method based on the use of inverse micelles as nanosized reactors to control the growth of semiconductor metal sulfide clusters is particularly useful for control of size, crystallinity and size dispersion, and we discuss certain aspects of this approach in more detail in a later section on MoS_2 and WS_2 photocatalysts.

Solution-based semiconductor clusters generally have a net charge since it is almost impossible to match exactly the bulk stoichiometry in a small nanocluster. These clusters are typically synthesized using non-polar solvents such as hexane, allowing extraction of the charged clusters into polar organic phases such as acetonitrile or tetrahydrofuran. These solvents are fully miscible with water. Once the extracted clusters have been dispersed in water, purification methods such as dialysis may be used to remove ions and organic by-products and also to change the surfactant used to stabilize the cluster solution in water.

The stability of a metal sulfide nanocluster in the presence of air and water depends critically on its nanostructure and only a certain class of layered dichalcogenides such as MoS_2 are stable against photo-oxidation in water. Accordingly, we focus on the synthesis of these clusters later in this chapter.

3.6.2 Characterization

Understanding the effect of photocatalyst size, structure, crystallinity and chemical composition requires a variety of characterization techniques from both surface science and analytical chemistry. Selection of appropriate characterization methods accelerates the pace of development of new photocatalysts by providing feedback to the synthetic chemist. Methods such as transmission electron microscopy (TEM) and selective area electron diffraction (SAD) can yield insights into the average particle size, size dispersion, nanocrystallinity and phase. Techniques such as X-ray photoelectron spectroscopy (XPS) and extended X-ray absorbance fine structure (EXAFS) are frequently used to determine oxidation state, elements and elemental ratios present in a photocatalyst. X-ray diffraction (XRD) provides information regarding average crystallite size and phase structure and small-angle X-ray scattering (SAXS) can also provide the average particle size and size dispersion. Since all these techniques require a high-vacuum environment for their application, the information that they provide needs to be complemented by analytical chemical methods allowing the photocatalyst to be studied in the liquid state, which is the environment where photocatalysts are typically used.

An important point about utilizing electron microscopy to study cluster size is that only a small portion of the sample (e.g. a few hundred particles) is analyzed. Therefore, if an unrepresentative portion of the TEM grid is chosen, conclusions regarding the average size and size dispersion for the entire sample are not warranted. Instead, SAXS is a more objective measurement since $\sim 10^{12}$ clusters are contributing to the SAXS signal. The drawback of SAXS measurements for the smallest photocatalysts (1–3 nm) is that the scattering is very weak and conclusions regarding size dispersion are more problematic. Since this size regime corresponds to the size of photocatalysts that are typically the most active, one must use analytical methods designed for the study of large molecules and polymers such as size-exclusion chromatography (SEC) to obtain precise size and size dispersion data for such nanoclusters.

For fully dispersed nanosized photocatalysts, it is possible to use HPLC to separate both chemicals and photocatalysts in complex mixtures and study each component by various analytical methods such as optical absorbance or photoluminescence. There are two mechanisms for separation. The first, SEC, depends on the degree of permeation of clusters into a porous chromatographic medium which is packed into a column of a given length. Clusters are injected into a flowing mobile phase such as toluene in which they are soluble and then transported through the porous medium. A good example of a hydrophobic porous medium suitable for SEC is microgels of cross-linked polystyrene. Chromatographic columns packed with microgel particles are designed to separate chemicals by size. The smallest chemicals can penetrate a larger fraction of these channels and therefore take a longer time to elute. Using molecules of a known size, one can calibrate the column so a given elution peak time can be used to obtain the hydrodynamic size of a nanocluster [88]. Figure 3.7 gives an example of the effect

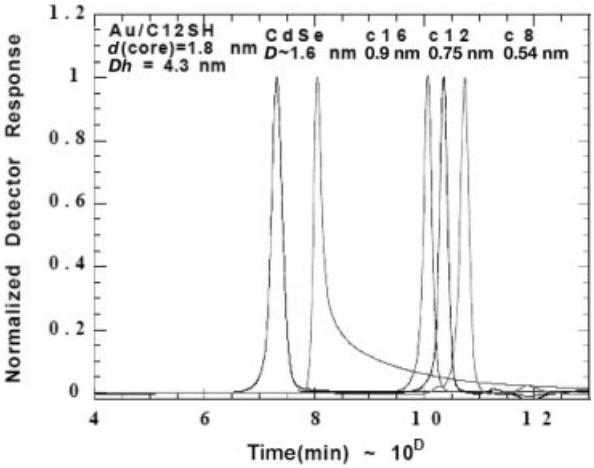


Figure 3.7 HPLC of monodisperse nanoclusters of Au and CdSe compared with three alkane standards: octane, C8 (0.54 nm), dodecane, C12 (0.75 nm), and hexadecane, C16 (0.9 nm). (Reprinted with permission from J. P. Wilcoxon, P. Provencio, *Journal of Physical Chemistry B* **2005**, *109*, 13461).

of chemical size on the elution time for three calibration standards consisting of aliphatic hydrocarbons labeled C16 (hexadecane), C12 (dodecane) and C8 (octane). Larger metal (Au) and semiconductor (CdSe) nanoclusters are included in this chromatogram. The time axis is proportion to the hydrodynamic size, $D \sim \log t$ as shown. This hydrodynamic size includes surfactants which are on the surface of the nanoclusters.

The simple relation between elution time and hydrodynamic size breaks down when the chemical or nanocluster interacts chemically with the column material. Examples of materials which have strong chemical affinities for metal oxide clusters are silica- and alumina-based columns which may be modified with various types of organic moieties to make them more or less hydrophilic. If a nanocluster is somewhat hydrophilic, e.g. a metal oxide like TiO_2, and the column is also hydrophilic (e.g. also a metal oxide such as alumina), the cluster will interact or stick and then release constantly as it moves down the column. How strongly the molecule sticks to the column will influence its elution time, with the most hydrophilic clusters eluting at the longest time. This affinity chemistry can be used to separate and study clusters based on cluster surface chemistry. Since surface chemistry is so vital to photocatalytic activity and selectivity, this type of chemical affinity chromatography provides a useful complement to SEC.

Various types of detectors may be used in-line with the separation column to detect the elution time of chemicals. In Figure 3.7, the detector response has been normalized by its peak, but the total area under the elution peak is proportional to the amount of chemical and is a very useful piece of information. If the chemical or cluster absorbs light, an absorbance spectrometer which can detect both the complete spectrum or just a single wavelength can be used [89]. Since semiconducting clusters such as MoS_2 absorb in the visible region, monitoring the absorbance provides a signature for the elution of a cluster. Most organic pollutants absorb light only in the UV or near-UV range, so monitoring the light absorbance in this region allows the detection of these species. For non-absorbing chemicals, the change in the refraction of light when a chemical is present in the mobile phase can be used to detect the elution. Fluorescent molecules or nanoclusters (most semiconductor nanoclusters used as photocatalysts are at least weakly fluorescent and emit in the visible region) can be detected with an in-line fluorescence spectrometer. The entire absorbance spectrum corresponding to an elution peak can be used to distinguish different clusters or chemicals [90]. The width and degree of shape homogeneity of the spectra of the elution peak can also be used to gauge the size dispersion of the clusters.

Collection of an eluting peak allows further identification of the chemical or cluster by other analytical methods such as gas chromatography–mass spectrometry (GC–MS). Elemental composition can be quantified by using X-ray fluorescence, (XFS). The latter technique is non-destructive and can be used for either solutions or solid films. The collected solution can be dried out and the resulting powder subjected to gas adsorption measurements to determine the available area per unit mass. Comparisons of such data with the geometric area based on the particle size are useful for inferences regarding geometry.

Dynamic light scattering (DLS) measures the diffusion rate of particles in dilute solution, where by dilute we mean that the particles are separated from each other by many particle diameters. This is the case for most photocatalysis studies. DLS is very useful for monitoring particle photocatalyst size changes and aggregation which might occur as a result of the photocatalytic reaction. If aggregation of the colloids occurs, for example, less catalyst surface area will be available to substrates, lowering the photoactivity of the nanomaterial. It is important to establish that the nanosized photocatalyst average size remains unchanged as a result of the reaction.

Studies of TiO_2 synthesis in water–alcohol mixtures sometimes use DLS to measure the particle size in solution and to follow changes with time, since these clusters are not agglomerated and are fully dispersed. This is rarely done for most experiments using commercial gas-synthesized TiO_2 nanocluster photocatalysts since slurries or suspensions of these clusters in water are turbid. This means that light is multiply scattered, precluding DLS measurements of cluster diffusion. For example, the strongly turbid solutions formed by commercial powders of TiO_2 such as Degussa P25 are unfortunately unsuitable for DLS studies.

3.6.3 Detailed Examples of Nanocluster Synthesis and Photocatalysis

3.6.3.1 Semiconductor Nanoclusters

Photocatalysts based on semiconductor nanoclusters must have certain chemical, electronic and optical properties to be useful in environmental remediation. Other properties include low synthesis cost, photostability in water over a wide range of pH values and a light absorption range that includes a significant portion of the solar spectrum. In addition, the oxidation potential of the valence band hole must be sufficiently positive to form free radicals from adsorbed organics and/or create hydroxyl radicals from adsorbed water. Ideally, the reduction potential of the conduction band electrons must be negative enough to transfer electrons to dissolved oxygen or other electron acceptor molecules. No single material of a fixed size can satisfy all of these conditions, but as many reviews have noted, titania comes the closest [13, 17, 34]. It is photostable under near-UV illumination conditions, rendering it environmentally benign, and its redox potentials can drive photooxidation reactions for a variety of organic and biotoxins [13, 17]. It suffers primarily from too wide a bandgap, which requires that UV light excitation be used, corresponding to only 2–5% of the solar spectrum.

Many complete reviews focusing only on titania have appeared addressing its photophysical and photocatalytic properties [13, 14, 16, 17, 20, 22, 24, 26, 34, 40]. We shall selectively and critically review some of its properties based on these studies. We then discuss in more detail nanosized semiconductors which have received less attention, such as MoS_2 and WS_2. In our discussions we will emphasize the connection between synthesis, size, nanostructure and photocatalytic properties and also the characterization methods best suited to study this relationship.

3.6.3.2 TiO_2

Synthesis, Nanostructure and Electronic Properties The growth of a solid, colloidal material such as TiO_2 or SnO_2 from chemical precursors consisting of metal alkoxides can be catalyzed by acid hydrolysis. The hydrolysis is a stepwise process which initially produces TiO_2 molecules during the induction period. Over time, an increase in concentration of these molecules creates a condition of supersaturation, at which point nucleation occurs to form small agglomerates (typical sizes are 1–2 nm). It is generally known that a relatively short nucleation period removes the supersaturation condition and results in a narrow size distribution. Following the nucleation events, growth continues until all the precursor is exhausted.

Over the years, protocols have been empirically developed to synthesize TiO_2 colloids in the nanosize regime. Unfortunately, reviews and papers discussing the use of nano-TiO_2 as a photocatalyst usually give very little or no specific information regarding the many details of the synthesis, particularly the important issue of colloid size control. Research suggests that particle size is important for the photocatalytic activity of TiO_2 [58]. Hence it is worthwhile to give some synthesis examples from our laboratory discussed in a previous paper [68].

Our method is based on slow injection of a solution of either titanium or tin isopropoxide in 2-propanol into a rapidly stirred acidic solution (pH = 1.5). The ratio of 2-propanol to water was fixed at 1:10. Under these conditions, we observed that smaller sizes are favored by higher precursor concentrations and faster rates of injection for colloid sizes between 10 and 100 nm. A plausible rationalization of this observation is that higher titanium isopropoxide concentrations will produce larger numbers of critically sized nuclei of TiO_2, with less available precursor left to add to these nuclei, so growth will both be faster and end more quickly.

A systematic increase in final colloid size occurs upon increasing the pH of the acidic solution. However, this size increase is accompanied by a loss of long-term stability against aggregation. A critique of this explanation of the observed smaller sizes at higher concentrations is that it ignores the possible sintering and/or exchange of atoms between clusters which may occur after all the precursor is exhausted. This can result in changes in the colloid size distribution and even the average size. Our conclusions regarding size control differ from those of other workers who synthesized larger (400–500 nm) TiO_2 colloids by a similar process and found no systematic dependence of final colloid size on initial precursor concentration [91]. However, these experiments did not use acid-catalyzed hydrolysis and so the nucleation process was slower for an equivalent precursor concentration and resulted in colloids 400–500 nm in size. The small specific surface area of such colloids makes them unsuitable for efficient photocatalysis.

Nanostructure, Surface Structure and Photocatalytic Activity Although the TiO_2 anatase phase is favored thermodynamically, the solution growth process that we described can produce a mixed anatase–rutile phase and also amorphous material. Degussa P25 photocatalyst, for example, has a 70 : 30 anatase:rutile composition. Both anatase and rutile phases have tetragonal lattices with a rock salt-like structure.

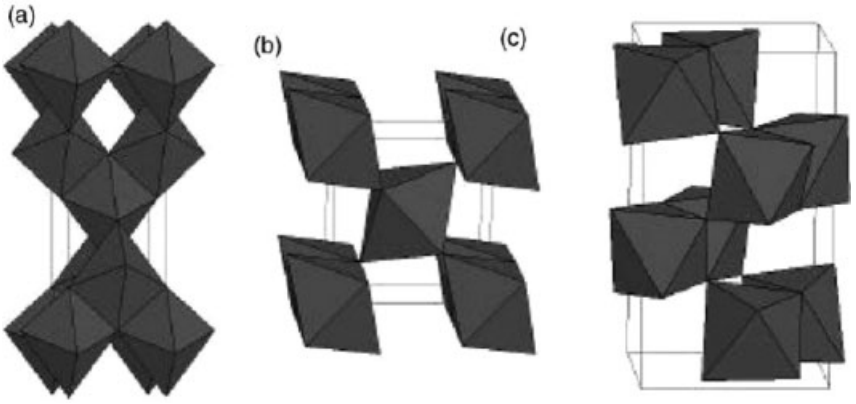


Figure 3.8 Crystal structure of (a) anatase, rutile (b) and (c) brookite showing the connections between TiO_6^{2-} octahedral subunits which distinguish the crystal structure of TiO_2 (Reprinted with permission from O. Carp, C. L. Huisman, A. Reller, *Progress in Solid State Chemistry* **2004**, *32*, 33).

If one views the structure as consisting of TiO_6^{2-} octahedral subunits, then these octahedra are connected by their edges in rutile and their vertexes in anatase as shown in Figure 3.8. Hence the latter has slightly less symmetry. Subsequent thermal treatments are generally used to maximize the anatase phase and improve crystallinity. Due to oxygen vacancies in the structure, the conductivity mechanism is n-type.

Many accounts in the literature assert that the anatase phase has greater photocatalytic activity than rutile [40, 92]. However, it is likely that the internal phase of nanocrystals of TiO_2 is less important to its photoactivity than the arrangement of the atoms at the surface, since these are the only atoms which interact directly with the substrate organic molecules. The number of hydroxyl groups present at the surface, which depends on the details of the synthesis, is also critical to the activity. It is worth noting that a mixed anatase–rutile–amorphous phase as found in Degussa P25 is, for most oxidation reactions, the most active known TiO_2 photocatalyst. This observation argues against an explanation of photoactivity solely in terms of internal phase. Instead, in this case improved charge separation has been invoked [93]. However, in our opinion, known impurities such as Fe(III) in Degussa P25 may actually be a better explanation.

When TiO_2 is used to photo-oxidize chemicals in water or air, water molecules rapidly adsorb on the surface. It has been estimated that there are about 5–15 hydroxyl groups for every square nanometer of TiO_2 surface [13, 94]. In addition to changing the adsorption characteristics of the catalyst surface, higher pH values increase the concentration of hydroxyl radicals, increasing the photo-oxidation rate of many organic chemicals. The adsorbed water molecules at the surface are also chemically important since they can form hydroxyl radicals upon reaction with photogenerated holes. These latter reactive species are key to the photo-oxidation of organic molecules in both the liquid and gas phase.

Substitutional Doping by Nitrogen The small amount (2–5%) of sunlight which is absorbed by TiO_2 limits its practical application in many circumstances. Thus, researchers have explored various synthetic strategies to increase the amount of light absorbed while still retaining the low cost and high stability of this material. A promising approach is to substitute nitrogen or sulfur atoms for oxygen in the TiO_2 lattice. Nitrogen doping has been shown to increase the absorbance onset of titania from 390 to almost 520 nm [95]. Nitrogen is introduced by exposure of the TiO_2 powder to ammonia gas at around 600 °C for several hours. This reduces the area as measured by the BET method by a factor of almost four. Perhaps related to this reduced surface area, it has been reported that the photocatalytic efficiency using these doped materials is only 14% of that found for the same photocatalytic reactions conducted at 351 nm, so most of the absorbed visible photons are not contributing to the desired photo-oxidation [96]. To overcome this deficiency, Morikawa *et al.* examined the effect of deposition of metals such as Cu, Ni, Pt, Zn and La on nitrogen-doped TiO_2 [97]. The ions were impregnated into the N-doped powders from aqueous solution, followed by evaporation of the water at 150 °C and calcinations at 300 °C for 2 h. They studied the oxidation of acetaldehyde using only light with wavelengths exceeding 410 nm. They reported that only Cu and Pt significantly enhanced the photocatalytic oxidation rate. They also found an optimal concentration of Cu of around 0.5 wt.%, at which the rate was enhanced by roughly a factor of two compared with the nitrogen-doped powder control. They hypothesized that the role of the Cu islands on the TiO_2 is to increase the carrier lifetime on the impregnated powders. It is worth noting that acetaldehyde was oxidized by TiO_2 doped with Cu at 0.5% continuously for 100 days without degradation of the catalyst, so this approach may work in practical photoreactors for indoor air purification.

Deposition of Metals An important factor for enhancing photocatalytic efficiency in small particles, noted previously in this chapter, is increasing the electron and hole lifetimes by improved charge separation. The deposition of noble metals on TiO_2 can reduce carrier recombination by serving as electron traps. The general synthetic approach used to deposit metals such as Pt on TiO_2 surfaces is photo-reduction of Pt salts such as chloroplatinic acid (H_2PtCl_6) in the presence of stirred slurries of the TiO_2 and near-UV light. The electrons promoted to the conduction band of the TiO_2 reduce the Pt salt which deposits on the TiO_2 surface. The amount of light exposure can be used to control the amount of Pt deposited [22].

As noted in a recent review by Choi, there are conflicting reports concerning the photocatalytic benefits of depositing Pt on the surface of TiO_2 [22]. The conflicting results can be traced to the different types of powders used, the loading level of the metal and the fact that different organic chemicals have been studied by each group. For example, Choi [22] found that the length of irradiation time used to deposit Pt on the TiO_2 affected not just the amount of metal deposited but also its oxidation state. Shorter photo-reduction times favored the presence of Pt(II) and Pt(IV) in addition to metallic Pt. Only Pt(0) was found to enhance the photocatalytic activity, which makes sense if the metal islands serve as electron storage sinks and reduce the recombination kinetics. Pt atoms in higher oxidation states may only serve to block access to

active sites and thus reduce the photo-oxidation of organic chemicals. One might also expect the physical size and adsorption mechanism to be strongly affected by the type of organic, which explains several of the conflicting reports concerning the efficacy of Pt deposition [69, 98].

An additional example of the use of metal island deposits in commercial anatase powders is the use of Ag/TiO_2 powders prepared by the photo-reduction of Ag ions by suspensions of TiO_2 [56]. The deposition of the Ag on TiO_2 was accompanied by increased absorption of the Ag/TiO_2 in the visible range (400–600 nm) where Ag clusters have significant absorbance. However, the absorbance was much broader than would be observed in isolated spherical clusters. This visible absorbance increased as the fraction of the Ag deposited increased from 0.5 to 2.0 wt.% relative to TiO_2. The catalytic photo-oxidation of two dyes was studied using a batch slurry reactor illuminated with UV light. Both dyes were photo-oxidized at substantially faster rates compared with the naked TiO_2 powder and, as in the Pt/TiO_2 studies described above, an optimal Ag loading of around 1.5% was found. The fast photo-oxidation of dyes by TiO_2 prevents the use of the dyes themselves to enhance the absorption of TiO_2 into the visible region. However, as discussed below, this dye sensitization has been proposed as a method of increasing the efficiency of TiO_2 photocatalysts.

The use of TiO_2 with deposited noble metals can be studied with either batch slurry reactors as described above and shown in Figure 3.9, in studies of TiO_2/Pt and TiO_2/Ag or using more practical flow reactors whose walls are coated with the catalyst material. A good example of the last reactor design was given by Orlov *et al.* [55]. In this work, the photo-oxidation of two significant types of environmental pollutants was studied using both Degussa P25 TiO_2 and the same material modified by deposition of gold particles.

The method for deposition of the Au nanoparticles on the TiO_2 was taken from that described by Haruta [99]. This method produces hemispherical gold islands with a large contact area along the perimeter between the TiO_2 support and the Au. This large contact area is considered critical in the activity for oxidation reactions since the TiO_2 will adsorb oxygen and water while the Au can serve as a reservoir of electrons. The synthesis method described by Haruta can produce small Au islands on any form

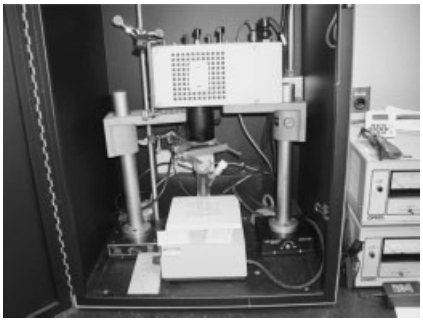


Figure 3.9 Photochemical reactor with overhead xenon lamp illumination, magnetic stirring, sampling sidearm.

of substrate, including powders, honeycombs and thin films. An ionic solution of a gold precursor such as $HAuCl_4$, when aged for about 1 h at a pH > 6 in the presence of a high surface area powder such as Degussa P25 TiO_2, forms deposits of $Au(OH)_3$ on the powder surface. Calcination can then be used to transform this coating into metallic Au islands. Control of the particle size is not critical to this method since only the perimeter Au atoms are believed to be catalytically active. In fact, Haruta showed that the rather large 4.6 nm diameter particles deposited by this method for pH > 6 are more active for the photocatalytic oxidation of CO than 2 nm Pt particles.

In the photocatalysis studies described by Orlov *et al*. [55], the Au islands on Degussa P25 TiO_2 were first formed at ~1 wt.% Au, followed by dip coating of the resulting slurry on the inner cylinder of a photochemical reactor to form a film. Heating the inner cylinder of the reactor to ~300 °C for 30 min created good film adhesion and a uniformly thick coating with good longevity. Axial illumination through the clear outer cylinder of the reactor permitted good light penetration, and air flow from the bottom of the reactor provided both mixing and ample dissolved oxygen. This fixed-bed reactor design could be used in the field by flowing contaminated water through the coaxial cylinders.

The reactor design and Au/TiO_2 catalysts was tested for two common pollutants, chlorophenols, which originate from agricultural run-off of insecticides and fungicides, and methyl *tert*-butyl ether (MTBE). MTBE has been used as a gasoline additive to improve wintertime combustion and is now banned in the USA due to leaks from storage tanks into the water system in the past. MTBE is carcinogenic at only a few parts per billion. Both compounds are relatively robust and novel treatment methods such as photocatalytic oxidation are being explored [100].

A significant finding of Orlov *et al*.'s work [55] was that the reaction rate for the mineralization of 4-chlorophenol was increased by 50% over that of Degussa P25 TiO_2 by use of the TiO_2 with Au deposited at 1 wt.%. A doubling of the reactivity was shown for the destruction of MTBE. The reactor design used was also shown to allow removal of either pollutant at water flow rates required to treat contaminated water pools. The temperatures required to form stable films of Au/TiO_2 on the reactor inner walls did lead to some loss of surface area and thus lower activity than a slurry reactor using the same Au/TiO_2 photocatalyst, but the activity was still greater than that of pure Degussa P25 TiO_2 powders. The long-term application and stability of the photocatalyst in real-world conditions still requires more testing. The additional cost of the Au would be acceptable provided that the catalyst stability can be shown to be exceptional.

Dye Sensitization [$Ru(pyr)_3$] Although several experiments indicate that organic dyes adsorbed on the surface of TiO_2 undergo photo-oxidation, certain reactions involving chlorinated hydrocarbons and aromatics can be extended to the visible region using dyes which are adsorbed on the surface of the TiO_2 particles. The basic idea is that the dye absorbs visible light and if the energy level of the photoexcited electron on the dye molecule is higher than that of the TiO_2 conduction band the electron will be transferred into the conduction band of the TiO_2 semiconductor. The oxidized dye is then capable of capturing another electron (i.e. oxidizing) from a pollutant and being regenerated. The

increased physical separation of electron and hole due to this transfer process can improve the efficiency of the photo-oxidation. Examples of dyes which are effective and reasonably stable for this process are ruthenium–bipyridine complexes [101, 102]. The cost of such complexes due to the expensive Ru may make such modifications impractical for large-scale systems, however.

An interesting method of extending the absorbance onset via electron transfer while avoiding the problems with degradation of the organic part of a dye is to deposit transition metal salts such as Pt, Rh or Au chloride on the surface of TiO_2. It has been demonstrated that photo-oxidation of chlorinated compounds such as 4-chlorophenol using only visible light at 455 nm is then possible [65, 103–105]. The metal chloride complex absorbs visible light and undergoes M−Cl bond breakage, leaving the metal in an oxidized state and the chlorine atom on the cluster surface. Then electron transfer from this chlorine atom to a chlorine atom in the organic compound also adsorbed on the TiO_2 occurs and regenerates the metal, making the reaction catalytic. There are less possibly adverse reaction pathways in this approach compared with the use of organic dyes as sensitizers.

Research Needs for Future Improvements The key to further improvements in TiO_2-based photocatalysts is to develop synthetic methods extending the light absorption into the visible region so that UV lamps and their costs can be eliminated. These new photocatalysts may have to be based on new nanosized materials as described in the next section. Also, not enough is known concerning the true longevity of TiO_2 when used under real conditions of salts and other inorganic metal ions in the presence of organic pollutants. Since any particulate photocatalyst must be immobilized to prevent contamination of the water and loss of the catalyst, high surface area flow reactors which allow light to reach all the photocatalyst surface are critical. Methods of inexpensively regenerating the active surfaces of such reactors will also have to be developed.

3.6.3.3 Alternative Photocatalytic Materials

Introduction Several alternatives to TiO_2 as potential photocatalytic materials have also been investigated in an attempt to access more of the visible portion of the solar spectrum and potentially to enhance the photocatalytic efficiency for different reactions. Some of these alternative materials include ZnO, SnO_2, CdS, MoS_2, WS_2 and, more recently, nitrides and oxynitrides such as TaN and TaON, respectively. Of the oxide materials (other than TiO_2), ZnO has been the most studied. Along with ZnO, we also review the newer nitrides, specifically TaN, as part of a small sampling of the materials that have been investigated. MoS_2 is outlined in detail in the latter part of Section 3.6.

ZnO ZnO is another wide-bandgap material which has been explored as a photocatalyst by several groups. One concern with this material is its photostability for certain reactions in aqueous solution [106]. For other reactions, especially photo-oxidation of dyes, ZnO may have a better efficiency than TiO_2 [107]. Since the cost and absorbance characteristics of the two photocatalysts are similar, ZnO in nanosize form may be a useful alternative photocatalyst.

ZnO can be made by acid-catalyzed hydrolysis of zinc acetate in ethanol. A white gel-like material is formed by using this synthesis. A photocatalytic film on glass is formed by dip coating on to a glass substrate and calcining the film at 500 °C [87]. Typical ZnO particle sizes from XRD linewidths are about 35 nm with a BET surface area of around 27 $m^2 g^{-1}$ (For comparison, Degussa P25 TiO_2 has a 25 nm particle size and a BET surface area of 50 $m^2 g^{-1}$). It was found that the ZnO fluorescence was quenched in the presence of chlorinated aromatics, which indicates the hole was being transferred to the aromatic. The most strongly absorbed compounds were chlorinated aromatics compared with chlorinated aliphatics. Photocatalysis studies confirmed that the more strongly adsorbed aromatic compounds were also photo-oxidized at a faster rate.

By monitoring the rate of disappearance of these compounds under 365 nm illumination, it was asserted that the ZnO efficiency was larger than films of Degussa P25 TiO_2 under the same conditions. However, this efficiency was obtained from the initial slope of the kinetic data showing the rate of disappearance versus irradiation time. Despite the rapid initial disappearance, mineralization of widely investigated compounds such as 4-chlorophenol was incomplete after 250 min of irradiation. Our photocatalytic studies of 4-chlorophenol using slurry reactors of Degussa P25 TiO_2 showed complete degradation of this compound within this irradiation time of 250 min [52]. It is possible that immobilization of the ZnO and TiO_2 on the glass slide used in these studies decreases the available surface area, making their photocatalytic reactor design the limiting factor. The authors did not explain how the total surface area and light flux were measured to allow a direct comparison of the two photocatalysts. Hence it is difficult to see how their conclusions regarding the superior efficiency of ZnO are supported.

ZnO fluorescence and its sensitivity to hole scavengers can be used to monitor the presence and the destruction of organic molecules as reported by Kamat's group [108]. The sensitivity to the presence of chlorinated aromatics is about 1 ppm using the fluorescence emission at typical concentrations of nanosized ZnO (1 mg mL^{-1}). The recovery of the fluorescence as the chlorinated aromatic is photo-oxidized allows the reaction kinetics to be monitored. This method also allows researchers to follow the state of the ZnO since any degradation or activity loss will lower the total fluorescence from the solution. An example of this behavior for 4-chlorocatechol is shown in Figure 3.10. Successive additions of 4-chlorocatechol do not cause loss of total fluorescence with time, indicating good ZnO stability.

Research suggests that ZnO probably does not provide any significant photocatalytic advantages compared with TiO_2, since it also fails to absorb a significant portion of sunlight. To bypass this shortcoming, other materials must be developed. An example of photocatalysts which absorb a significant fraction of visible light is given in the following section.

Ta_3N_5 and TaON Nanoparticles of nitrides and oxynitrides present two other potential classes of materials for photocatalysis. Specifically, Ta_3N_5 and TaON with bandgaps of ~2.07 and 2.4 eV, respectively, appear to be appropriate for visible light

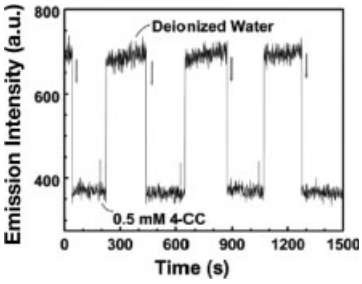


Figure 3.10 The fluorescence emission intensity from ZnO nanoparticle when exposed to aliquots of .5 mM 4-chlorocatechol (4-CC). The destruction of the 4-CC by the UV-driven photocatalysis requires about 200 s to restore the fluorescence. (Reprinted with permission from P. V. Kamat, R. Huehn, R. Nicolaescu, *J. Phys. Chem. B* **2002**, 106, 788).

photocatalysis and have been shown to be fairly active for the photocatalytic destruction of methylene blue (MB) [109]. Zhang and Gao synthesized nanosized Ta_3N_5 through high-temperature (600–1000 °C for 5–8 h) nitridation of Ta_2O_5 nanoparticles. The Ta_3N_5 crystallite size, which was determined by XRD, appeared to depend on the nitridation temperature, which was also critical in determining the extent of nitride formation (i.e. complete nitridation only occurred at higher temperature starting at 900 °C). Thus, smaller sized crystallites (~18 nm) nitrided at 700 °C often also contained the tantalum oxynitride phase. It therefore follows that with the higher temperatures needed for complete nitridation, that larger crystallite sizes on the order of 75 nm formed at 900 °C. However, there seems to be some variation in the resulting nitride formation based on temperature as a parameter since, according to the authors, pure phases of Ta_3N_5 also occurred at 700 °C but with resulting crystallite sizes of ~26 nm. Control over the synthesis parameters for this materials system seems to be difficult. Thus, the formation of a pure nitride phase also proved difficult.

In studying the photocatalytic degradation of MB, Zhang and Gao compared the reactivity of Ta_3N_5 nanoparticles with standards such as Degussa P25 TiO_2 and TiO_2 doped with N ($TiO_{2-x}N_x$) for visible light reactions using batch slurry-type reactors [109]. Compared with the $TiO_{2-x}N_x$ under UV/Vis illumination, the authors found a faster rate of photodegradation of MB by the 18 nm Ta_3N_5–TaON phased nanoparticles and also the 26 nm nanoparticles consisting of pure Ta_3N_5. The comparison with the $TiO_{2-x}N_x$ is perhaps slightly misleading since, as we discussed above, this form of TiO_2 has been shown to have decreased photocatalytic efficiency compared with pure TiO_2. The main role of the N doping is to extend the absorption into the visible region. A fairer comparison under the UV/Vis conditions would be with Degussa P25, which the authors did not show. The authors performed the comparison with Degussa P25 only under visible light conditions where it has minimal absorption and subsequently suppressed photoactivity. Accordingly, the Ta_3N_5 systems perform better than both the Degussa P25 and $TiO_{2-x}N_x$ under visible light illumination for the photo-oxidation of MB. The

authors also claimed that a size dependence to this photocatalytic activity exists where smaller Ta_3N_5 crystallites (~18 nm) are more active than larger ones (~75 nm). It is important to distinguish, as the authors did, between crystallite size and particle size. There may be some evidence of varying photoactivity as a function of crystallite size, but this cannot be translated into particle size since, as can be seen from their TEM images, there is a significant particle size distribution where large aggregates have formed. Also, as stated by the authors, the 18 nm sample was not a pure phase of TaN, but consisted of a Ta_3N_5–TaON mixture whereas the 75 nm samples were pure Ta_3N_5. It is therefore difficult to conclude that the photocatalytic activity of these materials is size dependent. It could very well be a case similar to TiO_2, where phase plays a very important role depending on the specific reactions of interest. Regardless of size dependence, these systems do show potential for visible light photocatalysis, but it is not clear that they are better than TiO_2, especially Degussa P25.

3.6.3.4 MoS_2 and Other Metal Dichalcogenides

High surface area TiO_2 powders in nanosized form such as Degussa P25 TiO_2, $d \sim 25$ nm, as mentioned above, are the most studied photocatalysts. However, TiO_2 has a significant disadvantage due to its large bandgap of ~3.2 eV. This means that it can only be excited by UV illumination of ~390 nm or shorter wavelength, thus requiring the use of UV lamps, increasing the cost of detoxification. If sunlight is used as the light source, TiO_2 absorbs only 3–7% of the solar spectrum, as illustrated in Figure 3.11 [52]. Also shown in Figure 3.11 is the absorption edge of bulk TiO_2 compared with three sizes of MoS_2: bulk, 8 nm and 4.5 nm. Research and analysis by Tributsch [110] has revealed that in order for photocatalysts to be useful in solar fuel production and chemical waste detoxification, the semiconductor material must meet the following requirements: (1) have a bandgap matched to the solar spectrum, (2) have valence and conduction band energy edges compatible with the desired redox potentials, (3) be resistant to photochemical degradation and (4) have a short carrier diffusion time leading to faster energy transfer for the surface compared with electron–hole recombination times [52]. All of these properties can be achieved through tailoring of the optical and electronic properties of MoS_2 nanoclusters based on the size. Prior to discussing MoS_2 nanoclusters, it is worthwhile reviewing the properties and uses of bulk MoS_2.

MoS_2 Bulk Properties and Historical Background MoS_2 and its structural isomorphs such as WS_2, MoS_2 and WSe_2 have excellent corrosion resistance, which has resulted in a variety of high-temperature catalytic and lubrication applications. This resistance to oxidation comes from the two-dimensional structure of these materials and the resultant electronic properties shown in Figures 3.12 and 3.13. In Figure 3.12, it is observed that the valence band is composed of Mo d_{z^2} states and the conduction band of Mo d_{xy} and $d_{x^2-y^2}$ states, so that excitation of an electron across the 1.75 eV gap doesn't weaken any chemical bonds.

Human use of Mo from MoS_2 has a long history. The mineral form of MoS_2, known as molybdenite, is a naturally occurring mineral and one of the most abundant

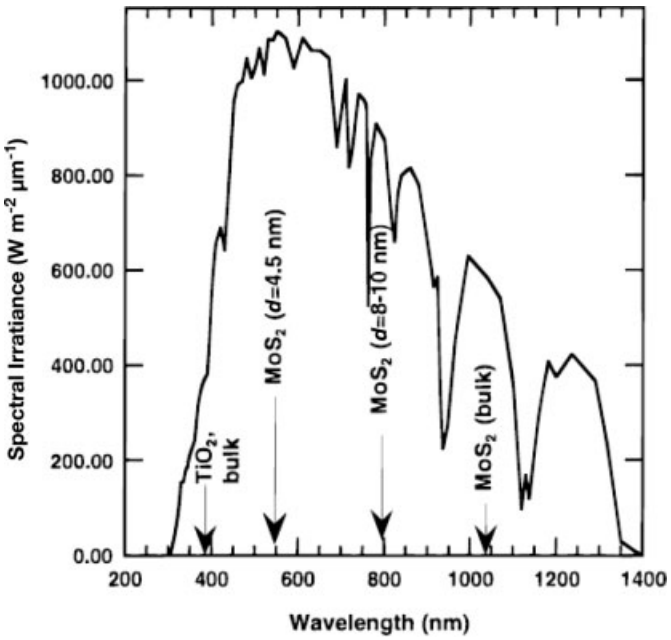


Figure 3.11 Solar spectral irradiance (AMD 1.5D) showing the radiation reaching the Earth's surface as a function of wavelength. Included in this spectrum are the absorption edges of bulk TiO_2 and MoS_2 in various forms. (Reprinted with permission from T. R. Thurston, J. P. Wilcoxon, *Journal of Physical Chemistry B* **1999**, *103*, 11).

forms found in the Earth's crust. This contrasts with most other metal sources such as Fe, Al and Ti, which occur naturally as oxides [111]. Due to its long history and the fact that molybdenite occurs naturally as a mineral in the Earth's crust, it is thought to have been used in very early times, such as in 14th century Japanese swords [111]. Initially, MoS_2 was thought to be the same as other ores such as lead or graphite and was named "molybdos", meaning "lead-like", by the ancient Greeks [111]. In 1778, Karl Scheele, a Swedish chemist, demonstrated that molybdenite was actually a sulfide mineral [112, 113] and that it contained molybdenum metal [111]. Extraction of the Mo was also later performed by Peter Jacob Hjelm in 1782 by reducing molybdenite with carbon [111]. Throughout the 19th century, interest in Mo was primarily non-commercial.

Commercial and technical applications of Mo were enabled by mining and extraction improvements, allowing the extraction of molybdenite in commercial quantities. The use of Mo in metal alloys steadily increased with time with the increased demands beginning around World War I. Most of molybdenite is MoS_2 and the rest consists of silicates and other rare metals such as Re [113]. To extract a purer form of MoS_2 from molybdenite, the mineral is subjected to a series of crushing, cleaning and purification steps [113]. The use of MoS_2 as a hydrotreating catalyst [114] for removal of heteroatoms from crude oil in fuel refining and as a high-temperature

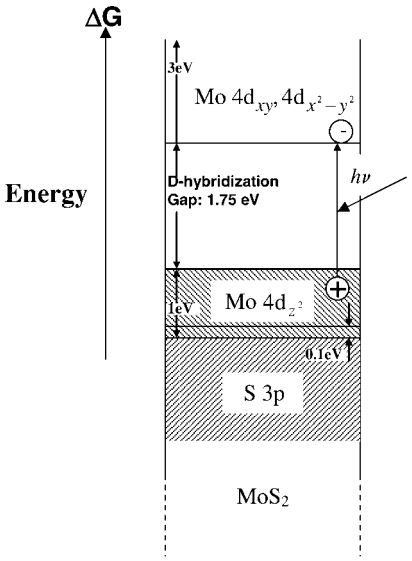


Figure 3.12 MoS_2 energy band diagram showing the exciton formation occurring between the d-states, accounting for stability against photocorrosion (i.e. no effect on the Mo–S bond upon excitation). (Redrawn with permission from H. Tributsch, *Zeitschrift für Naturforschung Teil A* **1977**, *32*, 972).

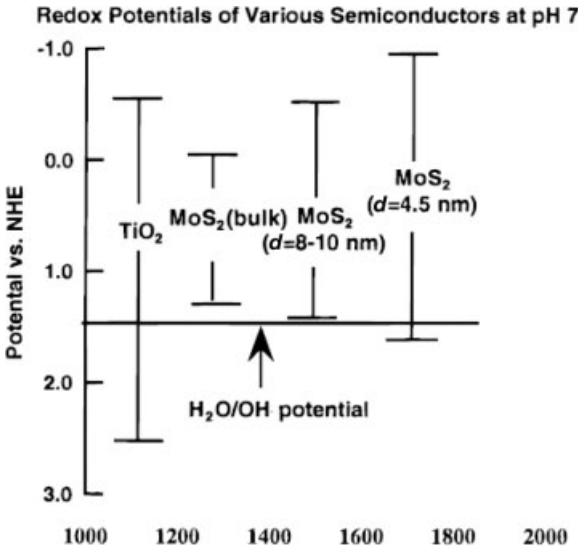


Figure 3.13 Conduction and valence band edge positions relative to the normal hydrogen electrode (NHE) for bulk TiO_2, bulk MoS_2, MoS_2 ($d = 8$–10 nm) and MoS_2 ($d = 4.5$ nm). (Reprinted with permission from T. R. Thurston, J. P. Wilcoxon, *Journal of Physical Chemistry B* **1999**, *103*, 11).

lubricant also increased rapidly during the 20th century. Research in Germany around the time of World War II identified formulations of MoS_2 on alumina supports capable of removing heteroatoms such as S, N and O from crude oil feedstocks, allowing subsequent refining operations using transition metal catalysts to be conducted without poisoning of these materials by sulfur. MoS_2 has been studied extensively both in the purified form and in its natural state. It was found to be easy to study using optical and electrochemical methods since it can be cleaved to very thin layers, making it suitable for optical and electronic studies [115].

Roscoe Dickinson and Linus Pauling elucidated the anisotropic crystal structure of MoS_2 in 1923 [116]. In this work, they noted that by 1904 Hintze had discovered that MoS_2 occurred as hexagonal crystals with complete basal cleavage planes. Dickinson and Pauling showed that Mo has six S atoms surrounding it in an equidistant manner at the corners of a triangular prism. They determined this through the use of Laue photographs and the theory of space groups [116]. Following this, they also showed that the S and Mo were stacked in S–Mo–S sandwiches along the *c*-axis and that there were two sandwiches per unit cell where $c = 12.3$ Å [116]. Later, Hultgren [117] and Pauling [118] showed that trigonal bonding in MoS_2 and WS_2 was due to d^4sp hybridization of atomic wavefunctions. The bonding within the MoS_2 sheet was determined to be covalent and the trilayer sheets were held together by weak van der Waals forces. Figure 3.6 shows a schematic of the MoS_2 sandwich structure. The most stable form of MoS_2 is the 2H-MoS_2 structure. Two other metastable polytypes also exist: 1T-MoS_2 and 3R-MoS_2[119]. The space group for the stable 2H-MoS_2 is $P6_3/mmc$ [120]. The quasi-2D, graphite-like structure where the trilayers consisting of S–Mo–S sandwiches held together by weak van der Waals forces allows the facile shearing reported by Dickinson and Pauling. MoS_2 has many applications in lubrication, especially in space applications and at high temperatures where its solid state form is an advantage [121, 122]. Other industrial applications include catalytic hydrodesulfurization (HDS) [114, 123]. It has also been proposed as a possible solar photoelectrochemical electrode for hydrogen generation from water [124].

MoS_2 is a semiconductor with an n-type conduction mechanism. Its optical absorbance properties are due to both direct and indirect transitions, which have been investigated by a variety of measurement techniques, including optical absorption and transmission [115, 125–127], reflectivity measurements [128, 129], electron energy loss measurements [130] and electron transport measurements [131]. For example, Goldberg *et al.* [127] measured the optical absorption coefficient, emphasizing the features below the first exciton [115] located at ∼680 nm. The lowest energy absorption at 1.24 eV (∼1000 nm) at room temperature was forbidden (i.e. indirect) and thus weak. [127]. Using photocurrent measurements, Kam and Parkinson obtained similar values of 1.23 eV for the indirect E_g of MoS_2 and 1.69–1.74 eV for the direct gap [132]. Roxlo *et al.* used photothermal deflection and transmission spectroscopy [133] to obtain a highly precise value of the indirect bandgap of 1.22 ± 0.01 eV. Using the augmented spherical wave method, Coehoorn and co-workers calculated detailed band structures for MoS_2, $MoSe_2$ and WSe_2 [120, 134].

It is possible to determine the valence and conduction band levels of bulk n-type MoS_2 using cyclic voltammetry. For example, Schneemeyer and Wrighton studied the oxidation reactions of biferrocene and *N,N,N′,N′*-tetramethyl-*p*-phenylenediamine using an MoS_2 electrode and found that the flat-band valence band potential was $+1.9$ V vs. SCE, demonstrating that the direct transition of around 1.7 eV controls the oxidizing power of the holes [135]. They noted that this potential should allow most of the oxidation reactions possible at TiO_2 to take place for MoS_2 also.

To test this hypothesis without the competing reaction of water oxidation, they studied the oxidation of Cl^- ion in an oxygen-free solvent, acetonitrile. They found that a sustained evolution of chlorine gas was indeed observed upon visible illumination of the MoS_2 electrode with no loss of photocurrent for over 8 h of continuous operation. During this experiment, many moles of electrons were generated without any visible (i.e. microscopic) evidence for oxidation of the MoS_2 electrode. Their results support the results predicted by Tributsch for the electrochemical photo-oxidation of water described in the next section [110].

Photocatalytic Properties of MoS_2 A detailed and systematic search for photocatalysts allowing the photo-oxidation of water using visible light was reported by Tributsch in 1977 [110, 124]. The principles he outlined for successful water photo-oxidation should also apply to photo-oxidation of organic molecules. For example, a suitable photocatalyst must be capable of absorbing visible light, have valence and conduction band potentials appropriate for the substrate molecule to be oxidized, have an efficient mechanism for electron–hole separation, be chemically stable (i.e. exciton creation must not weaken the chemical bonds in the structure) and be inexpensive and/or easy to synthesize [110].

Based on the first requirement, oxide compounds were determined to be unsuitable since they absorb mostly in the UV region (e.g. TiO_2, ZnO, SnO_2). Polar semiconductors such as CdS were also ruled out since they are unstable as the photogenerated electrons weaken the chemical bond. Transition metal dichalcogenides appeared to be the most suitable candidates [110]. Tributsch investigated a large number ($\sim$70) of metal dichalcogenides and concluded that the most favorable materials were WS_2, MoS_2 and TcS_2. Due to its lack of abundance and high cost, TcS_2 was deemed unsuitable. WS_2 was also ruled out since it was difficult to obtain as large crystals at that time. MoS_2 was readily available and easy to process, becoming the obvious final choice.

A detailed analysis of the MoS_2 band structure was also reported (Figure 3.12) [110]. Optical transitions occur between the d states, as shown in Figure 3.12. Since the formation of the exciton pair occurs via the metallic d states, light absorption has no adverse affect on the Mo–S chemical bonds. It is this phenomenon that accounts for the stability of MoS_2 against photocorrosion and accounts for the widespread use of MoS_2 as a stable, high-temperature solid-state lubricant and its applications as an electrode material with good corrosion resistance in water.

Tributsch demonstrated the oxidation of water with the formation of molecular oxygen and hydrogen using MoS_2 as an electrode [110]. However, the reaction needed

to be assisted by a redox catalyst such as tris(2,2′-bipyridine)ruthenium (II), since the rate of O_2 generation was low. This was probably due to the inherent limitations of MoS_2 in the bulk form and a kinetic bottleneck for transfer of the conduction band electrons to form hydrogen gas by reduction of water. Also, being an indirect bandgap semiconductor with a small bandgap (~1.22 eV) in the near-IR region, the potential may not be large enough to drive the necessary redox chemistry. This is important for other photocatalytic applications also, such as the photodestruction of toxic chemicals, since the valence and conduction band edges may not match the redox potentials sufficiently.

Motivated by Tributsch's calculations, other groups explored the use of bulk transition metal dichalcogenides in photoelectrochemical applications and in photocatalysis. The behavior of n-WSe_2 and $MoSe_2$ single crystals as photoanodes in regenerative photoelectrochemical cells [136] was compared with the behavior of WS_2 and MoS_2 [137]. The diselenides were reported to have conversion efficiencies of ~10% using a sodium iodide electrolyte. However, the conversion efficiencies for the disulfides were considerably lower. Other groups were subsequently able to increase the conversion efficiency to >14% using nanosized WSe_2 in a polyiodide solution [138]. A still higher conversion efficiency (17%) was achieved using WSe_2 through enhancement of the crystal quality using an etching technique developed by Tenne [138, 139].

Some work has been reported on the use of bulk dichalcogenide powders for photodestruction of chemical wastes. DiPaola *et al.* studied WS_2 and WO_3 and also mixtures of the two materials in the photocatalytic degradation of phenol [140]. They discovered that the mixed system had a much higher rate of phenol conversion than the individual pure bulk powders. It is unlikely that simply mixing two bulk powders would affect the recombination kinetics as their interfaces are not in physical contact, so this result is difficult to rationalize. Our work with bulk powders of MoS_2 and WS_2 [52, 68] contradicts the above report of photocatalytic activity for either phenol or pentachlorophenol. On the contrary, we found that slurries of these powders and also other layered materials such as PtS_2 inhibit the near-UV photo-oxidation process compared with solutions containing no catalyst. However, nanosized MoS_2 shows considerable photocatalytic activity.

Nanosized Photocatalysts of MoS_2 and WS_2 One key to the photocatalytic activity observed in nanosized MoS_2 is the effect of size on the optical and electronic properties of nanosized semiconductors. The confinement of holes and electrons to a reduced size particle results in a blue shift of the absorption onset compared with the bulk material. For example, as shown in Figure 3.11, the absorbance onset of MoS_2 blue shifts from the bulk value of ~1040 nm to ~550 nm for 4.5 nm sized clusters in solution. Just as significant for photocatalytic redox reactions, the increase in bandgap with decrease in cluster size causes the valence band to shift to more positive values and the conduction band to more negative values. Both shifts enable photo-oxidation of organic chemicals to occur more readily. By measuring the rate of transfer of holes and electrons, one can estimate the energy levels as a function of cluster size, as shown in Figure 3.13, where the potentials are measured at

pH 7 versus NHE [52]. It should be noted that the pH of a solution of a metal sulfide has a different chemical effect on the cluster surface compared with a metal oxide such as TiO_2, where low pH tends to make the surface positive and change its substrate binding properties. For example, the formation of $Ti–OH_2^+$ structures, which is important at typical pH values of around 1 used in many photocatalytic studies, is not possible for MoS_2.

An advantage of MoS_2 nanoclusters over bulk MoS_2 is the increased surface to volume ratio (S/V), allowing for more efficient surfaces traps per volume. In addition to lowering the cost of the photocatalyst by reducing the amount of material required, the high S/V ratio increases the likelihood of surface trapping of the photogenerated carriers and so increases the lifetime of the charge carriers prior to recombination. This increases the probability of successful hole transfer to an organic pollutant. Also, since nanocluster solutions scatter a negligible amount of light, simplified analysis of the photocatalytic behavior of the system is possible. As in the case of TiO_2, the method of nanocluster synthesis is likely to have a significant effect on the cluster photocatalytic properties.

MoS_2 Nanocluster Synthesis Since photocatalytic applications of MoS_2 often require solution processing methods to form either films or disperse solutions, liquid-phase synthesis has some advantages compared with vacuum or gas-phase methods. Monodispersed solutions of MoS_2, $MoSe_2$, WS_2 and WSe_2 nanoclusters were first synthesized by Wilcoxon and co-workers using an inverse micelle approach [51, 52, 68, 141–144]. In this method, surfactant molecules are dissolved in a suitable non-polar organic such as toluene or octane. The surfactants are chemically bipolar with a water-loving or hydrophilic head group joined to a hydrophobic or water-hating tail group. Their bipolar nature causes the surfactants to associate and form droplet-like aggregates as shown schematically in Figure 3.14. The non-polar organic tail groups prefer to form an interface with the non-polar solvent, which allows the hydrophilic head groups to be shielded from the solvent and lowers the free energy of the solution. These aggregates are called inverse micelles since the curvature of the interface is the opposite to that of normal micelles that form in water with the hydrophilic head groups interfacing with a continuous water phase. The reader may be familiar with normal micelle aggregates as they are critical to allowing solubilization of hydrophobic entities such as oil and permitting detergent action, which gives soaps the ability to remove oil from clothes and skin.

Inverse micelles have a water-like interior volume which can dissolve ionic metal salts just as water does. However, the absence of water means that undesired chemical reactions such as hydrolysis to form metal hydroxides cannot occur. The dissolved metal ion pairs interact and are stabilized by the hydrophilic head groups instead of water, and these head groups are not reactive chemically. When this metal salt solution is then mixed with another inverse micelle solution containing an ionic sulfiding agent, such as a metal sulfide or H_2S, a controlled aggregation to form a nanocrystal of MoS_2 ensues [52]. The cluster growth process is relatively slow compared with the same reaction in a continuous, homogeneous medium since it requires diffusion and collisions between micelles to allow exchange of the growing

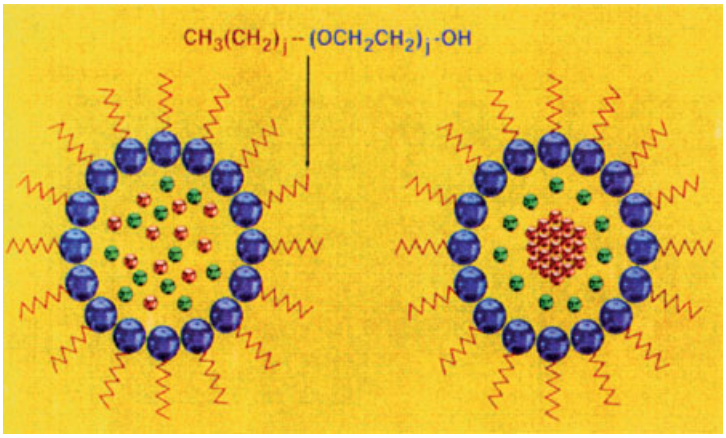


Figure 3.14 Schematic showing a two-dimensional drawing of a spherical inverse micelle with hydrophobic tail groups (red) forming an interface with a continuous oil medium (yellow). The hydrophilic head groups shown in blue provide a water-like environment which permits ionic metal salts such as $MoCl_4$ to dissolve and form ion pairs shown in red [Mo(IV) cations] and green (Cl^- anions) in the micelle interior. Reaction of the metal ions with a sulfur source such as H_2S results in the growth of a nanocrystal of MoS_2 as shown on the right. The surface of the nanocrystal is stabilized against aggregation by the surfactants which are later removed via an extraction process described in the text.

nanocrystals. The kinetics of this process are controlled by the size of the inverse micelle cage that is used to dissolve the Mo salt and also the initial salt concentration. The slow nanocrystal growth allows good ordering of the atoms even at room temperature and this is confirmed by HRTEM, showing atomic lattice planes and facets on the nanocrystals and also selected area electron diffraction (SAD). Figure 3.15 shows a SAD pattern of a 4.5 nm MoS_2 cluster, revealing the same hexagonal crystal structure as the bulk [51]. Figure 3.16 shows an HRTEM image of a MoS_2 cluster ~3 nm in size. This image reveals that the cluster is highly crystalline with no defects. The lack of defects in nanoclusters of this size is perhaps not surprising since any defect that may form has a very short distance to diffuse out to the nanocluster surface.

MoS_2 clusters synthesized in inverse micelles in non-polar oils such as decane can be extracted into immiscible polar organic solvents such as dry acetonitrile or methanol. In this process most of the ionic by-products and free surfactant is removed and the resulting solutions can be dissolved in water. Figure 3.17 shows a photograph of these solutions with each solution having a distinct color and absorbance onset determined by the average cluster size. Since acetonitrile can be obtained in high purity with no absorbance above 200 nm, detailed complete absorbance spectra on purified MoS_2 clusters are readily collected.

Optical absorption spectra of nanosized MoS_2 (4.5, 3 and 2.5 nm) compared with the bulk are shown in Figure 3.18 [51]. Curves 3, 4 and 5 correspond to solutions of 4.5 nm, 3 nm and 2.5 nm MoS_2, respectively (note the logarithmic scale of the absorbance axis). The absorption features in the bulk spectra can be traced to the

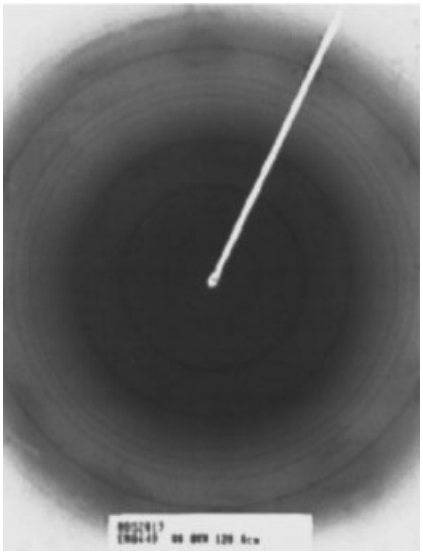


Figure 3.15 SAD pattern of 4.5 nm MoS_2 clusters showing a hexagonal structure. (Reprinted with permission from J. P. Wilcoxon, P. P. Newcomer, G. A. Samara, *Journal of Applied Physics* **1997**, *81*, 7934).

different transitions in the band structure as shown in Figure 3.19 [51]. There is a weak absorption onset at ~1040 nm (~1.2 eV) (Figure 3.18), which is attributed to the indirect bandgap occurring between the Γ point and the middle of the Brillouin zone between Γ and K [51]. The direct bandgap occurring at the K point is represented by the next absorption onset at ~700 nm (~1.8 eV). There are actually two peaks that correspond to this direct transition, labeled A_1 and B_1. The energy separation of these two peaks is most likely due to spin–orbit splitting of the valence band at the K point [51, 120, 126, 134]. There is another direct transition originating deep within the

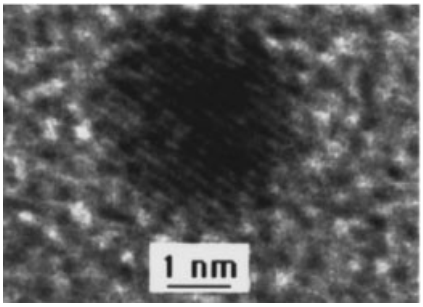


Figure 3.16 High-resolution TEM image of an ~3 nm cluster revealing its high crystallinity. (Reprinted with permission from J. P. Wilcoxon, P. P. Newcomer, G. A. Samara, *Journal of Applied Physics* **1997**, *81*, 7934).

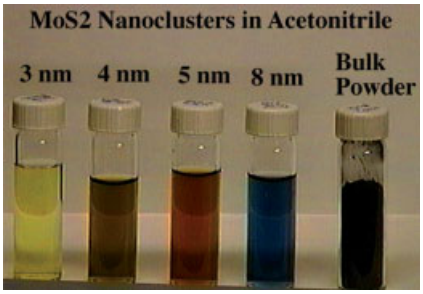


Figure 3.17 The effect of decreasing size on the color and absorption onset compared with a bulk powder is an illustration of the quantum confinement effect in MoS_2.

valence band, corresponding to the absorption onset at ~500 nm (~2.5 eV). This transition is labeled C and D.

As semiconductor clusters become molecular in size, the continuous bands due to the translational symmetry develop molecule-like structure and discrete bands. The valence band evolves into the highest occupied molecular orbital (HOMO) and the conduction band evolves to become lowest unoccupied orbital (LUMO). These bands are seen in the absorbance peaks of Figure 3.18. Photoexcitation leads to electron

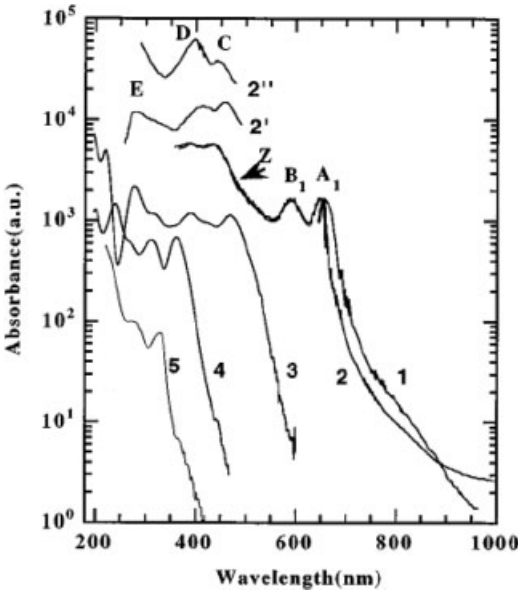


Figure 3.18 Optical absorption spectra of two bulk MoS_2 crystalline samples compared with three MoS_2 nanocluster samples. Curve 1 = synthetic crystal from ref. Goldberg; curve 2 = natural crystal from ref. Evans and Young; curves 2′ and 2″ = higher resolution of high-energy features (ref. Frindt and Yoffe); curve 3 = 4.5 nm MoS_2; curve 4 = 3 nm MoS_2; curve 5 = 2.5 nm MoS_2. (Reprinted with permission from J. P. Wilcoxon, P. P. Newcomer, G. A. Samara, *Journal of Applied Physics* **1997**, *81*, 7934).

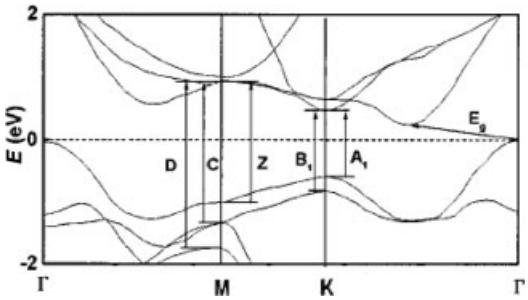


Figure 3.19 Band structure portion taken from the band calculations of Coehoorn *et al.* The transitions corresponding to the optical features in Figure 3.17 are labeled. (Reprinted with permission from J. P. Wilcoxon, P. P. Newcomer, G. A. Samara, *Journal of Applied Physics* **1997**, *81*, 7934).

promotion from the HOMO to the LUMO. Some weak luminescence is observed from these cluster solutions due to radiative recombination from the LUMO to the HOMO. This luminescence can be used to monitor the transfer rate of the electrons or holes to other adsorbed species and roughly estimate the size-dependent conduction band energy levels [84].

It is worth noting that other synthetic methods for the formation of MoS_2 nanoclusters result in clusters with metallic properties. For example, ultrahigh vacuum methods have been employed to deposit 2D Mo nanoclusters on gold surfaces followed by sulfidization of Mo metal islands [123, 145–147]. As a result of this synthetic route, the formation of triangular clusters through analysis using scanning tunneling microscopy (STM) [123, 145] was discovered (Figure 3.20). The

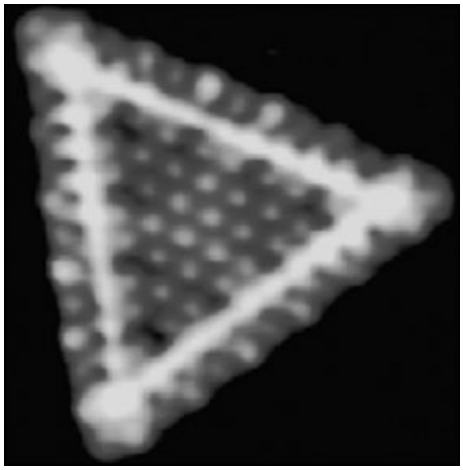


Figure 3.20 STM image of a triangular MoS_2 nanocluster with dimensions of 48 × 53 Å. (Reprinted with permission from M. V. Bollinger, K. W. Jacobsen, J. K. Norskov, *Physical Review B* **2003**, *67*, 085410).

STM image in Figure 3.20 shows a single layer of MoS_2. These triangular clusters were shown to have metallic conducting properties for the atoms at the edges. Since these clusters were grown on a support, it is not certain that free clusters would have the same geometry and electronic properties.

More recently, Lauritsen *et al*. showed that excess sulfur can be present at the edges depending on the cluster size [148]. They also showed that there is a tendency for the preferential formation of particular sizes, suggesting the formation of "magic sizes" that are thermodynamically favored. This is an important observation under these synthesis conditions since this phenomenon is observed in many other materials systems and under various synthetic routes. It also leads to better size control and, with this control, the authors were able to link atomic-scale structural analysis to cluster size using STM. The subsequent implications for enhanced catalytic activity, particularly for hydrodesulfurization (HDS), are thus promising.

Accordingly, Bertram *et al*. investigated both MoS_2 and WS_2 clusters formed in the gas phase and used mass and photoelectron spectroscopy to show that these clusters have planar platelet structures [149]. The platelet structures were shown to stack just as in the bulk to form larger clusters.

These MoS_2 clusters from Bertram *et al*.'s work were grown by evaporation of bulk MoS_2 using a pulsed electric arc. Both charged and neutral M_nS_m clusters grow within an inert seeding gas of helium. To study the effect of extra sulfur on the structure, they also vaporized Mo and W metal, allowing metal clusters to form of various sizes which were then exposed to controlled amounts of H_2S gas. The smallest platelets formed had a metallic character and were chemically inert. They began stacking just as in the bulk structure since multiples of fundamental platelet masses were observed. Mass analysis was consistent with extra sulfur atoms at edge sites. Simulations indicated these extra sulfur atoms stabilized the structures and a "magic sized" cluster of WS_2 is shown in Figure 3.21 taken from [149]. (The phrase "magic sized" derives from the larger abundance of certain masses of clusters observed due to extra stability for special numbers of metal atoms.) Clusters were stable over a wide range of M:S ratios. This was explained by the hypothesis that polysulfide chains could grow near the edges of the cluster. Their observation of $S : Mo > 2$ is consistent with our observations of excess sulfur in clusters grown in inverse micelles, purified by chromatography and analyzed using X-ray fluorescence (XRF) spectroscopy. The nature of the bonding at the catalytically active Mo edge sites which are protected by sulfur atoms is likely critical to the photocatalytic activity, as is known in the case of HDS catalysis, the major catalytic application of MoS_2 [148].

The small triangular cluster shown in Figure 3.21 has a nearly zero bandgap (i.e. is metallic) and only the edge W atoms are metallic, whereas the states near the HOMO (the Fermi level) are delocalized over the entire plane of W atoms. Furthermore, this metallic character is due to the excess of sulfur at the edges and, as the clusters are grown larger and approach the 2 : 1 S : W ratio of the bulk, a gap develops. The optical properties such as absorbance and luminescence of our larger 2.5, 3 and 4.5 nm MoS_2 and WS_2 clusters grown in inverse micelles are consistent with an energy gap

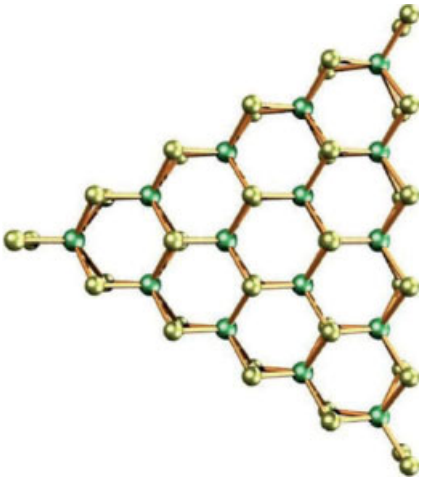


Figure 3.21 Structure of $W_{15}S_{42}$ platelet calculated by Bertram *et al.* (Reprinted with permission from N. Bertram, J. Cordes, Y. D. Kim, G. Gantefor, S. Gemming, G. Seifert, *Chemical Physics Letters* **2006**, *418*, 36).

between HOMO and LUMO states, indicating that they are semiconducting. Furthermore, the onset of the absorbance is consistent with an indirect optical transition even for the smallest 2.0 nm clusters studied by our group [144]. Indications are that clusters grown in free space differ in structure and electronic properties from those grown on substrates, possibly due to the significant interaction energies between the Au atoms in the substrate and the Mo atoms.

Quantum Size Effects in MoS_2 The effect of decreased cluster size and higher energy carrier confinement in both the in-plane and out-of-plane, transverse or *c*-axis direction of platelet stacking are important for the electronic properties of metal dichalcogenides. Studies of size-related phenomena were reported by Consadori and Frindt in 1970 through investigations of thin layers (13 Å, corresponding to a thickness of one unit cell) of WSe_2 [150]. They found that the optical absorption onset depended on sample thickness and showed a very small shift to higher energies with decreasing thickness. These effects were attributed to quantum size effects [150]. This was not the first observation of the effect of carrier confinement in small semiconductors; however, it may have been the first observation of quantum size effects in layered dichalcogenides. These effects were actually due to two-dimensional (2D) confinement in an infinitely large sheet and the shift in the energy of the A_1 exciton (E_{A_1}) was only 0.15 eV compared with infinitely thick samples.

As described below, compared with the shifts observed for metal dichalcogenide clusters in solution due to in-plane carrier confinement, the transverse out-of-plane confinement for thin layers is very weak. Lateral confinement of the carriers appears to be necessary in order to observe strong size effects. This result is consistent with the d–d band optical transitions which dominate the photoexcitation behavior of MoS_2. The energy shifts reported on thin WSe_2 reported by Consadori and

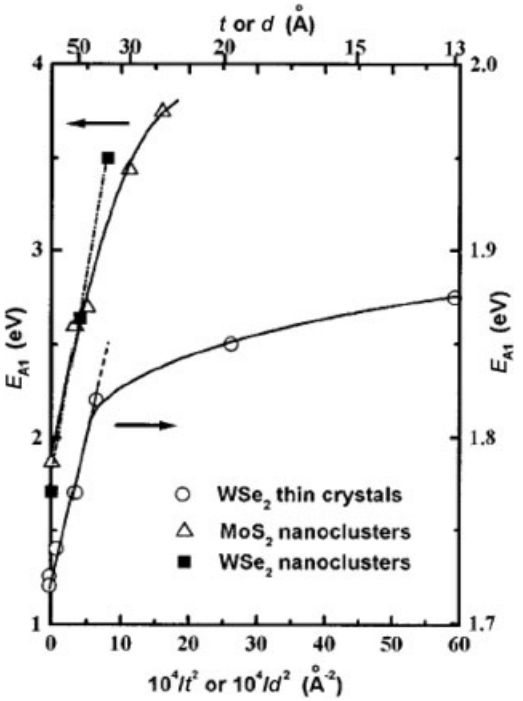


Figure 3.22 Influence of metal dichalcogenide layered structure dimensionality (2D vs. 3D) on the strength of quantum confinement of the A_1 exciton in WSe_2 and MoS_2. 2D confinement represents thin crystals and 3D confinement represents clusters. (Reprinted with permission from J. P. Wilcoxon, P. P. Newcomer, G. A. Samara, *Journal of Applied Physics* **1997**, *81*, 7934).

Frindt [150] are compared with the transverse confinement of MoS_2 and WSe_2 clusters demonstrated by Wilcoxon *et al.* [51] in Figure 3.22. In this figure, first exciton energy E_{A_1} is plotted as a function of $1/t^2$ for the 2D thickness study (right axis) and as a function of $1/d^2$ (where d = diameter) for the 3D cluster study (left axis). This $1/t^2$ dependence of ΔE_{A_1} is obeyed to $t \approx 40$ Å but then deviates significantly as the thickness decreases. The shift in E_{A_1} as sizes approach 40 Å in-plane carrier confinement is over an order of magnitude larger in the transverse direction. This deviation is even stronger at smaller sizes, which demonstrates the large difference between transverse and longitudinal confinement.

The absorption spectra of Figure 3.18 have structured features such as minima and maxima whose positions shift to the blue with decreasing cluster size. Figure 3.23 provides an examination of these peaks for MoS_2 clusters in dilute acetonitrile solution. The absorption edge shifts correspond to changes in the size-dependent bandgap and the density of energy states as the clusters become molecular in size. The effects of quantum confinement depend on the cluster dimension relative to the bulk excitonic Bohr radius (r_B). This dimension defines a cross-over from a strong to a

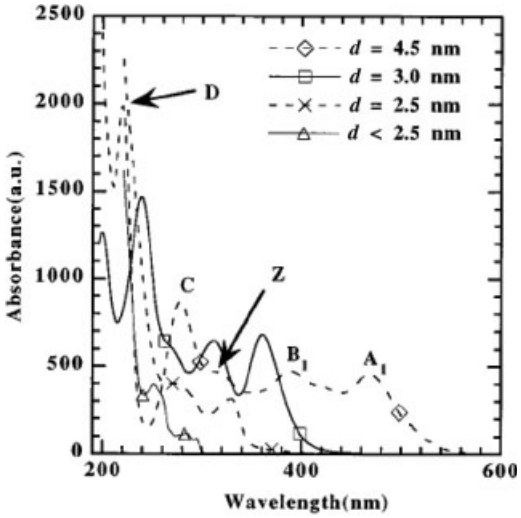


Figure 3.23 Optical absorption spectra of MoS_2 clusters ranging from <2.5 to 4.5 nm. The corresponding band structure transitions are shown, compare with Figure 3.17. (Reprinted with permission from J. P. Wilcoxon, P. P. Newcomer, G. A. Samara, *Journal of Applied Physics* **1997**, *81*, 7934).

weak regime. In 1938, Mott proposed a model of the hydrogen-like exciton for the $n = 1$ Rydberg states of the A_1 and B_1 excitonic peaks [151]. He thus determined that r_B for MoS_2 is 2 nm [51, 115, 151]. The 4.5 nm diameter MoS_2 clusters ($R = 2.25$ nm) are thus slightly larger than the bulk exciton, whereas 3.0 nm ($R = 1.5$ nm) and 2.5 nm ($R = 1.25$ nm) clusters should exhibit strong carrier confinement. Although $D = 4.5$ nm clusters are slightly larger than the r_B for MoS_2, they still exhibit properties vastly different from the bulk, which can be verified by the large blue shift in the absorption spectra shown in Figures 3.18 and 3.23.

The extent of carrier confinement manifests itself through the effect of particle size on spin–orbit splitting of the absorbance peaks. Decreasing cluster size is accompanied by an increase in the spin–orbit splitting of the A_1 and B_1 excitonic peaks. Effective mass theory (EMM) defines quantum confinement as the shift in the absorption edge or bandgap, $E_g(R)$, for a cluster with radius R, to be proportional to $1/2\,\mu R^2$, where μ is the reduced exciton mass [79]. When $E_g(R)$ is plotted against $1/R^2$, a straight-line plot should result with a slope proportional to $1/(2\,\mu)$ [51]. However, deviations from linearity for each excitonic peak occur as the cluster size decreases (Figure 3.24). The increase in spin–orbit splitting between the Wannier excitonic peaks, A_1 and B_1, as a function of decreasing cluster size also occurs [51]. The difference between peaks A_1 and B_1 is 0.67 eV for MoS_2 clusters of 2.5 nm [51] compared with ~0.20 eV for the bulk [120, 126, 134].

When $R/r_B > 4$, the quasi-particle nature of the excitons is predicted to be preserved [81]. The quasi-particle characteristics are lost, however, when $R/r_B < 2$ and the charge carriers are confined, leading to independent carrier behavior. The

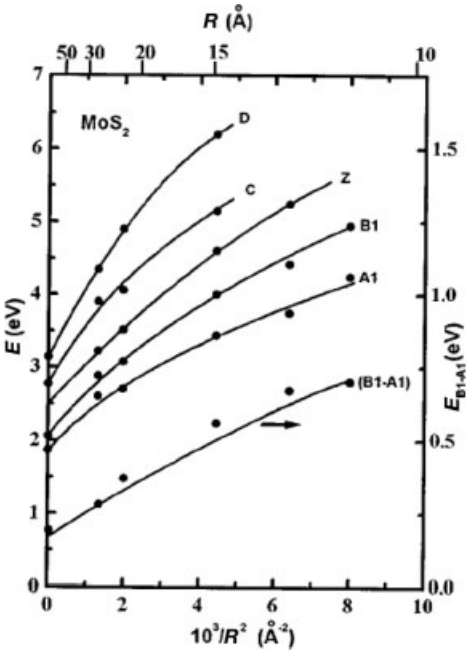


Figure 3.24 E(R) vs. $1/R^2$ showing deviations from linearity for each excitonic peak with decreasing MoS_2 cluster size. Increase in spin–orbit splitting is also represented by the right axis. (Reprinted with permission from J. P. Wilcoxon, P. P. Newcomer, G. A. Samara, *Journal of Applied Physics* **1997**, *81*, 7934).

presence of specific excitonic peaks in the absorption spectrum of MoS_2 is preserved for clusters sizes of $\leq$2.5 nm, while the significant blue shifts remain consistent with strong quantum confinement. This could be due to the 2D nature of the strong bonding in MoS_2 [51]. Since the hole and electron confinement radii differ, the true picture is more complicated. Bulk values for hole and electron masses may not apply to very small cluster sizes where the confining potentials are dependent on the cluster surface structure and chemical nature of the adsorbed ligands. For example, the presence of excess sulfur at the edge sites may increase or decrease the electron density on the Mo atoms and thus affect catalytic activity.

Photocatalysis Using MoS_2 Nanoclusters Nano-MoS_2 is suitable as a visible light photocatalyst due to the effects of strong quantum confinement on its optical properties. For example, the blue shift of the absorbance onset with decreasing cluster size increases the oxidation (valence band holes) and reduction (conduction band electrons) potentials (Figure 3.13). Photocatalysis experiments reported using either MoS_2 or WS_2 are much more limited than TiO_2 and date only to around 2000. Since photoexcitation can be driven with visible radiation, unlike most metal oxides, lower costs are potentially possible since a larger portion (visible) of the solar spectrum is accessible.

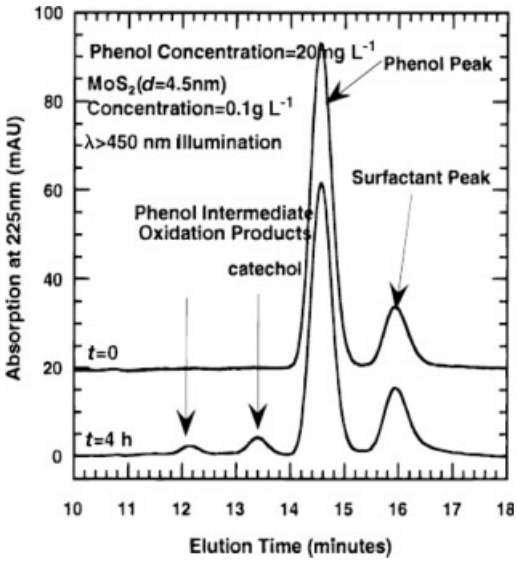


Figure 3.25 HPLC absorption trace showing phenol destruction with visible light ($\lambda > 455$ nm) using 4.5 nm MoS_2 clusters as the photocatalyst. (Reprinted with permission from T. R. Thurston, J. P. Wilcoxon, *Journal of Physical Chemistry B* **1999**, *103*, 11).

Thurston and Wilcoxon published some of the first reported photocatalysis using nanosized MoS_2 in the late 1990s [52]. They investigated the effect of MoS_2 cluster size on the efficiency of phenol photo-oxidation. A xenon lamp with appropriate band pass filters allowed comparisons with Degussa P25 TiO_2 photocatalyst [52]. A short-pass filter cutting off IR radiation above 1000 nm to minimize heating of the solution was used. HPLC using a reversed-phase octadecyl-terminated silica column and a mixture of water and acetonitrile as the mobile phase was utilized to analyze the phenol concentration as a function of illumination time. To study the MoS_2 cluster solutions, which were optically transparent, a long-pass filter limited the xenon arc lamp output to wavelengths greater than 455 nm. Figure 3.25 shows an absorption chromatogram for an experiment using 4.5 nm MoS_2 obtained using 455 nm light where the phenol elutes at ~14.6 min. After 8 h, the phenol peak area had decreased by ~25%. This was accompanied by the appearance of two phenol photo-oxidation products: catechol (13.4 min) and a possible isomer of catechol (12.2 min) [52]. Photo-oxidation using Degussa P25 TiO_2 in slurry suspension at 365 nm radiation showed similar behavior (Figure 3.26). As expected, no photo-oxidation of phenol using TiO_2 slurries illuminated with visible light occurred (Figure 3.27). The active sites on the MoS_2 clusters are possibly the empty d-orbitals accessible at the Mo metal edge sites. Phenol adsorption and hole transfer may occur at these locations and even the presence of a stabilizing cationic surfactant apparently does not prevent access to these sites. In fact, in later studies of pentachlorophenol oxidation using MoS_2 and TiO_2, certain cationic surfactants were shown to increase the rate of photo-oxidation.

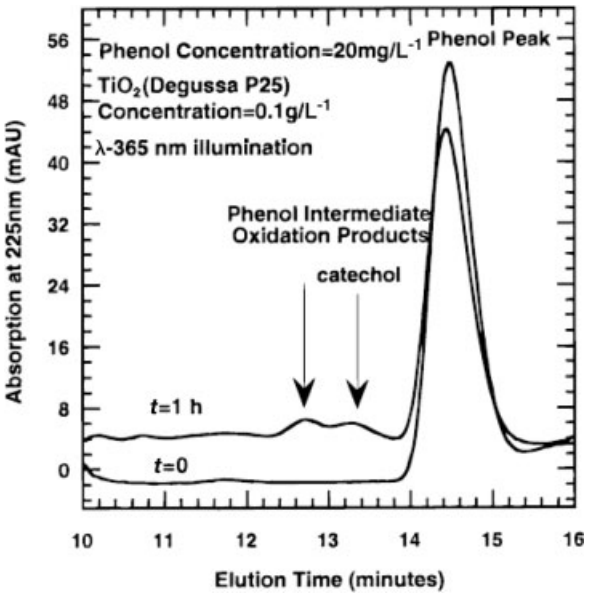


Figure 3.26 HPLC trace showing phenol destruction with 365 nm UV illumination catalyzed by Degussa P25 TiO_2. (Reprinted with permission from T. R. Thurston, J. P. Wilcoxon, *Journal of Physical Chemistry B* **1999**, *103*, 11).

Earlier we mentioned that there are technical advantages to using optically transparent solutions as compared with slurries to investigate photocatalysis. This advantage can only be realized with fully dispersed solutions of stabilized nanoclusters. Under these circumstances, analysis of the chemical state of the solution using chromatography combined with on-line optical absorbance yields the amounts of pollutant and oxidation by-products as a function of illumination time and also an elution peak corresponding to the nanosized photocatalyst. Unlike an aggregated powder such as Degussa P25 TiO_2, which is typically removed by filtration prior to chromatographic analysis, the state of the catalyst can also be determined. As an example, consider Figure 3.28, showing the absorbance spectrum collected at the elution peak of 4.5 nm MoS_2 clusters before and after photocatalysis [152]. Analysis of these spectra demonstrated that no decrease in the spectral area (proportional to the MoS_2 concentration) occurs with either visible or UV illumination. Importantly, the characteristic excitonic absorbance peaks in the MoS_2 spectra are present with no shifts in peak positions before and after the photocatalytic oxidation of phenol occurs. Since degradation would lead to a decrease in the area under the MoS_2 cluster peaks and a shift in their position, such measurements demonstrate that the phenol oxidation is catalytic. It is significantly more difficult to determine if any changes have occurred to a TiO_2 slurry catalyst.

By using larger, 8–10 nm, MoS_2 clusters, a much larger fraction of the incident light will be absorbed by the clusters. Unfortunately, the photo-oxidation was slower

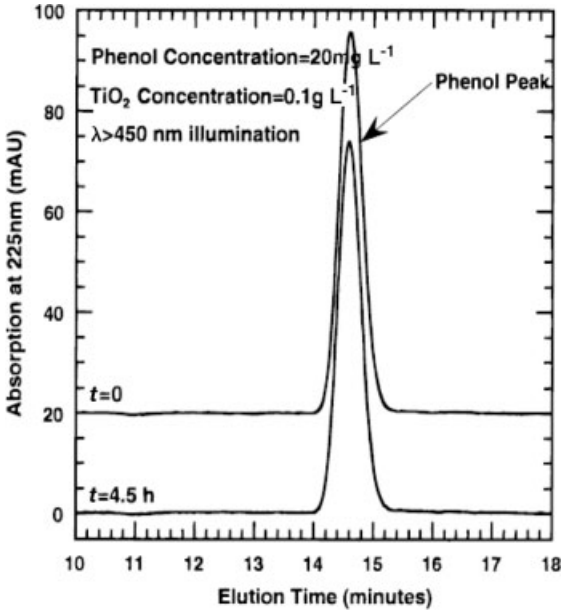


Figure 3.27 HPLC trace of attempted phenol destruction using Degussa P25 TiO_2 illuminated by visible light ($\lambda > 455$ nm). No decrease in phenol concentration is observed and no photo-oxidation products develop. (Reprinted with permission from T. R. Thurston, J. P. Wilcoxon, *Journal of Physical Chemistry B* **1999**, *103*, 11).

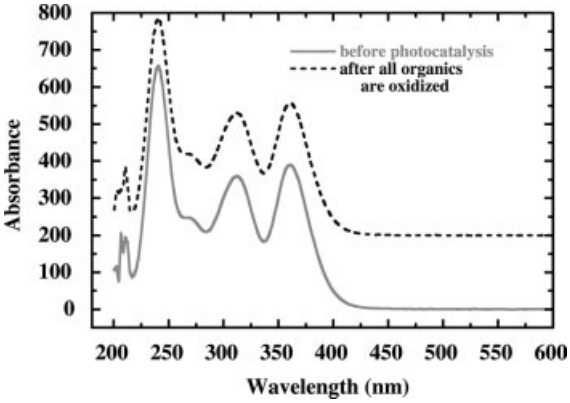


Figure 3.28 Optical absorption of MoS_2 before and after photocatalysis showing no spectral or intensity change. This suggests that nano-MoS_2 is acting catalytically with all its original properties preserved. (Reprinted with permission from J. P. Wilcoxon, T. R. Thurston, *Materials Research Society Symposium Proceedings* **1999**, *549*, 119).

in this case. Figure 3.29 shows a plot of the phenol concentration as a function of time under visible illumination for 4.5 nm MoS_2 clusters, 8–10 nm MoS_2 clusters and Degussa P25 TiO_2. These data were obtained from the calculated area under the phenol elution peaks of the chromatograms shown in Figures [25–27]. The catalytic nature of the 4.5 nm MoS_2 clusters for the phenol oxidation was verified by adding additional phenol after most of the phenol had been destroyed, repeating the reaction and observing the same reaction kinetics within the error bars indicated in Figure 3.29.

Using fully dispersed nanoclusters as photocatalysts is not practical in real environmental remediation, since it would be very difficult to remove the clusters from the purified water by filtration or other commonly used methods. Accordingly, experiments were undertaken to deposit the nanoclusters onto a TiO_2 powder which could be used as a slurry or eventually coated onto a high surface area support as has been done previously in flow reactors based upon TiO_2 catalysts. The deposited MoS_2 clusters then serve as sensitizers allowing visible light absorbance and transfer of the electron from the more negative MoS_2 conduction band to the TiO_2. Thus, both light absorbance and charge separation could be achieved. We deposited 8–10 nm MoS_2 onto several support materials: Degussa P25 TiO_2, SnO_2, WO_3 and ZnO [52]. P25 TiO_2 was the only support material that showed enhanced performance during

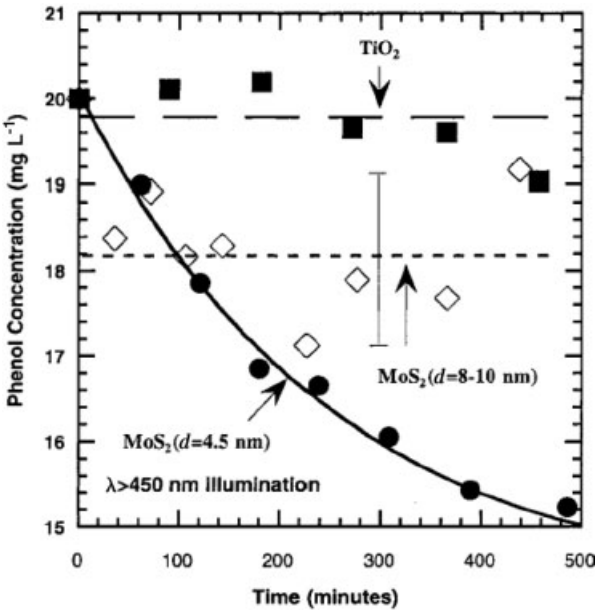


Figure 3.29 Visible light-induced photo-oxidation of phenol as a function of time as determined from HPLC. Order of phenol destruction: 4.5 nm MoS_2 > 8–10 nm MoS_2 > Degussa P25 TiO_2 (no activity). The error bar represents reproducibility. (Reprinted with permission from T. R. Thurston, J. P. Wilcoxon, *Journal of Physical Chemistry B* **1999**, *103*, 11).

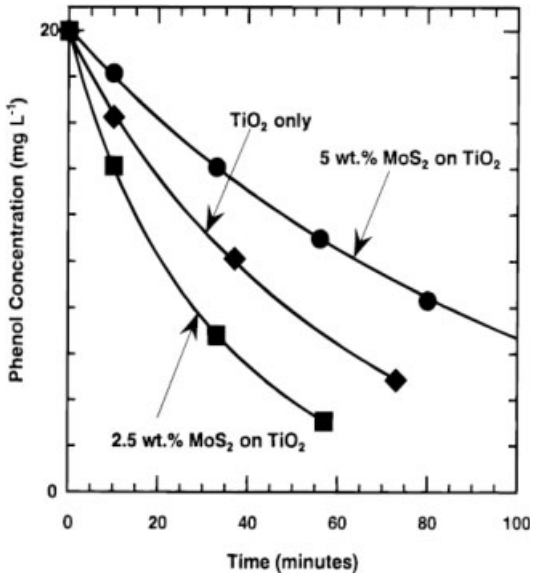


Figure 3.30 Phenol concentration as a function of time for Degussa P25 TiO_2 loaded with 8–10 nm MoS_2 at 0, 2.5 and 5 wt.%. Samples were irradiated with 365 nm light. (Reprinted with permission from T. R. Thurston, J. P. Wilcoxon, *Journal of Physical Chemistry B* **1999**, *103*, 11).

phenol destruction as a result of the MoS_2 deposition. A comparison of the phenol destruction rate under UV (365 nm) radiation for three different loadings of MoS_2 on TiO_2, 0, 2.5 and 5 wt.%, is shown in Figure 3.30. The 2.5 wt.% sample led to an increase in phenol photo-oxidation relative to TiO_2 alone whereas the 5 wt.% sample led to a decrease in the destruction rate. Hence an optimum loading is necessary for achieving the fastest phenol destruction rate for this system. This result is similar to that found for the deposition of metal clusters on TiO_2 described earlier. High concentrations of MoS_2 clusters may block critical phenol adsorption sites on the TiO_2, for example. Figure 3.31 shows a plot of the destruction rate of phenol as a function of MoS_2 loading on TiO_2. This figure reveals that once the optimum loading of 2.5 wt.% has been reached, there is a decrease in the phenol destruction rate.

Studies of the photocatalysis of pentachlorophenol (PCP) using 8–10, 4.5 and 3 nm MoS_2 clusters dispersed in water and also acetonitrile showed a strong size dependence of the rate of PCP destruction [68]. PCP is a very toxic chlorinated aromatic molecule. It belongs to the family of chlorinated phenols which are found widely in the environment due their widespread use as fungicides for wood preservation. Unfortunately, it has a very slow natural degradation rate. Also, direct photolysis of PCP has been reported to result in even more toxic by-products such as octachlorodibenzo-*p*-dioxin [17, 68]. Several studies have utilized TiO_2 as a photocatalyst and demonstrated that total mineralization of PCP to CO_2 and HCl is possible [26, 153, 154]. However, as discussed above, the limitation of TiO_2 is that it

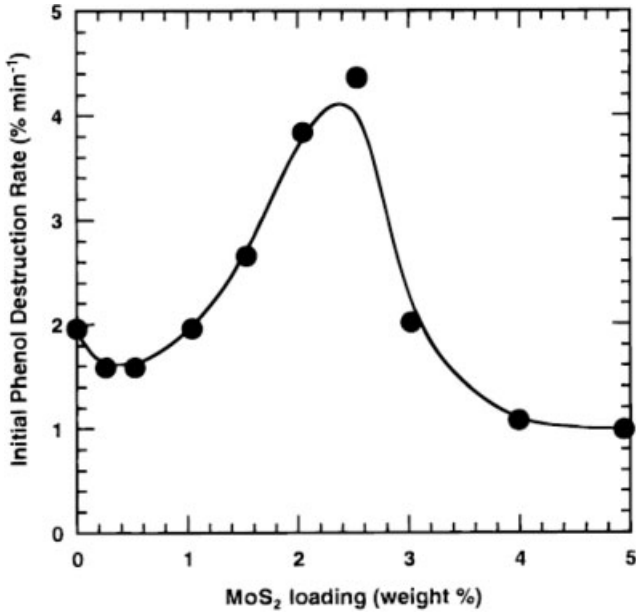


Figure 3.31 Initial phenol destruction rate dependence on 8–10 nm MoS_2 loading of Degussa P25 TiO_2 using 365 nm irradiation. (Reprinted with permission from T. R. Thurston, J. P. Wilcoxon, *Journal of Physical Chemistry B* **1999**, *103*, 11).

must be excited by UV light, which is only ~3% of the solar spectrum, probably requiring the use of lamps for a practical TiO_2 photoreactor.

It was necessary to use bulk CdS powder as a reference photocatalyst since TiO_2 is not active under visible illumination [68]. CdS has a similar absorbance onset (~525 nm) to MoS_2 nanoclusters [68]. The elution peak area from absorbance chromatograms was used to determine the PCP concentration as a function of illumination time. The decrease in PCP concentration as a function of time for 4.5 and 3 nm MoS_2 clusters, bulk CdS powder, bulk MoS_2 powder and bulk TiO_2 illuminated with visible light is shown in Figure 3.32 [68]. As might be expected, there was no change in PCP concentration using either bulk powders of MoS_2 or Degussa P25 TiO_2 under visible light radiation. Some decrease in PCP concentration was observed with bulk CdS powder. However, both sizes of MoS_2 clusters were more active than the CdS slurry powder.

The most interesting result of these studies (Figure 3.32) is the dramatic increase in PCP destruction rate for the 3 nm MoS_2 compared with the 4.5 nm MoS_2. In fact by 120 min, complete photo-oxidation of PCP occurs and there is no detectable PCP [68] (the sensitivity was 20 ppb). Although the 4.5 nm clusters absorb significantly more of the incident visible light, the most important size effect appears to be the more energetic electrons and holes created in the smaller 3 nm MoS_2 clusters. It is also possible that some of the activity increase is due to more favorable surface chemistry and binding properties of the 3 nm clusters compared with the 4.5 nm clusters. It is

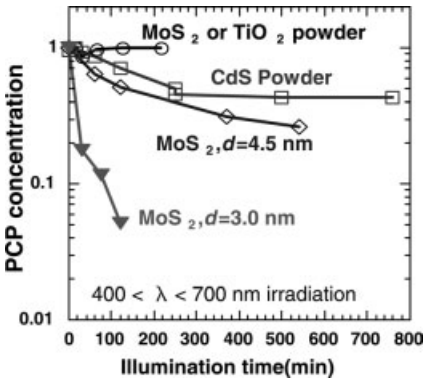


Figure 3.32 Semi-logarithmic plot of pentachlorophenol (PCP) concentration vs. visible light illumination time. Nanosized MoS_2 photocatalysts are compared with conventional semiconductor powders: MoS_2, TiO_2 and CdS. The smallest MoS_2 clusters (3 nm) with the most positive oxidation potential are the most active. (Reprinted with permission from J. P. Wilcoxon, *Journal of Physical Chemistry B* **2000**, *104*, 7334).

worth noting that this high rate of PCP oxidation also occurs in the presence of stabilizing surfactants and that PCP photo-oxidation using only visible light excitation of 3 nm MoS_2 clusters is higher than that observed for Degussa P25 TiO_2 using full lamp irradiation ($300\,nm < \lambda < 700\,nm$) [68].

The stability of the 3 nm MoS_2 clusters during photocatalysis reactions was also tested by Wilcoxon *et al.* [152]. The photo-oxidation of an alkyl chloride using visible light and 3 nm MoS_2 is shown in Figure 3.33. In this figure, the HPLC absorption chromatogram shows a peak at ~4.1 min (3 nm MoS_2), a peak at ~4.65 min (free surfactant) and a peak at ~5.2 min (organic impurity) at four different stages of the reaction: $t = 0$, 15, 30 and 60 min [152]. There is no evidence of the organic impurity left after 60 min. The surfactant that does not participate in cluster stabilization (i.e. free surfactant) present in the solution also photo-oxidizes as a function of continued irradiation and after 3 h is completely gone. However, the MoS_2 peak remains unchanged, demonstrating, again, the photocatalytic property of this material.

Electron transfer (ET) rates from MoS_2 clusters to electron acceptor molecules such as bipyridine (bpy) can be estimated by monitoring the change in the fluorescence decay rates as a result of ET [84]. For example, time-resolved fluorescence measurements showed a dramatic decrease in ET rates to bpy as the size of the clusters increased. The ET rates were fastest for 3 nm MoS_2 clusters, which is consistent with a more negative conduction band potential due to increased quantum confinement compared with 4.5 nm clusters. By using electron acceptor molecules with known redox potentials, one can also estimate how much the conduction band shifts with cluster size by such measurements. Similar experiments using hole acceptor molecules allow one to estimate the shift in valence band position with cluster size.

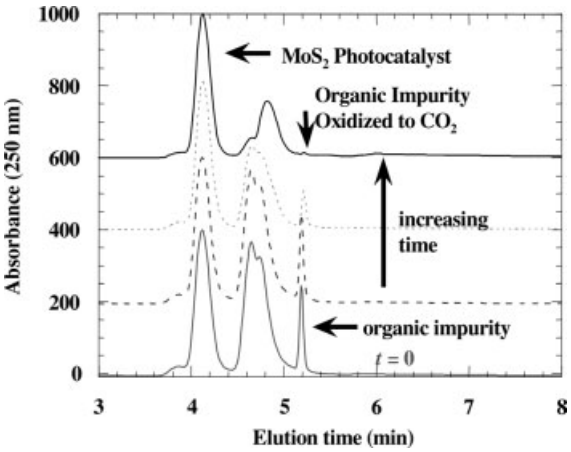


Figure 3.33 Liquid chromatographic analysis of a fully dispersed solution of MoS_2 containing an organic impurity. The absorbance at 250 nm vs. elution time is used to monitor the disappearance of the organic and the integrity of the MoS_2 nanocatalyst as a function of visible illumination time. (Reprinted with permission from J. P. Wilcoxon, T. R. Thurston, *Materials Research Society Symposium Proceedings* **1999**, *549*, 119).

ET to an acceptor in a photoredox reaction is preceded by electron–hole (e–h) separation. In order to have increased ET rates, there must be efficient e–h separation compared with e–h recombination. The small size of the MoS_2 clusters guarantees fast diffusion of the electron to the surface. This can be enhanced further through surface modification of the clusters. Wilcoxon *et al*. demonstrated that the deposition of gold (Au) cluster islands on the surface of 3 nm MoS_2 clusters suppressed e–h recombination with the possible outcome of enhancing e–h separation [51]. Evidence of this phenomenon is shown in Figure 3.34, where the decrease in photoluminescence (PL) as a function of the amount of Au on the cluster surface is demonstrated. As the Au content increases, the PL decreases. A decrease in PL represents a decrease in the e–h recombination rate relative to other non-radiative processes such as e–h transfer. This result implies that the photocatalytic properties of already active nanosized MoS_2 (4.5 nm or less) could be enhanced further with charge separation strategies that have proved useful for photocatalysts such as TiO_2.

One useful synthetic approach to improving the photocatalytic properties of nanosized MoS_2 is to form clusters of MoS_2 on the surface of another semiconductor such as TiO_2. By using MoS_2 as the light-absorbing component in such a coupled semiconductor system, visible light may be used to drive the photoredox reaction. This was first demonstrated, as noted above, in Thurston and Wilcoxon's work [52] on phenol photo-oxidation on MoS_2/TiO_2. The conduction band of bulk MoS_2 is not sufficiently negative to drive electron transfer from MoS_2 to TiO_2, but decreasing the MoS_2 cluster size will make the reduction potential more negative, suggesting that this strategy can be implemented using nanosized clusters of MoS_2 grown on TiO_2.

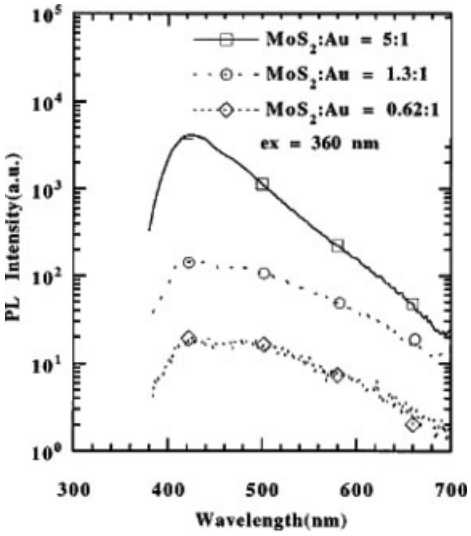


Figure 3.34 Influence of Au on the PL of 3 nm MoS_2. (Reprinted with permission from J. P. Wilcoxon, P. P. Newcomer, G. A. Samara, *Journal of Applied Physics* **1997**, *81*, 7934).

Ho *et al.* reported a new approach to making such a coupled semiconductor photocatalyst [155]. The approach mimics that used to deposit islands of Pt on TiO_2 by using the photogenerated electrons on TiO_2 to reduce a suitable Mo salt precursor and form islands of MoS_2. The metal precursor used was $(NH_4)_2MS_4$, where M = Mo, W. To maximize the transfer of photogenerated electrons from the Degussa P25 TiO_2 slurry, hydrazine was added to react with the holes and oxygen was removed by continuous nitrogen purging during illumination by a UV mercury lamp. The MS_2 islands that were formed were not nanocrystalline and needed to be calcined under nitrogen to create photoactive materials. XPS was used to determine the Ti:M ratios, which were 0.3 mol% (WS_2) and 0.6 mol% (MoS_2).

The absorbance of the coupled MS_2/TiO_2 systems was measured using diffuse reflection methods to minimize the effect of multiple scattering by the TiO_2. The measurements, shown in Figure 3.35, demonstrate that the growth of MS_2 clusters on the TiO_2 surface enhances the visible absorbance substantially. Their results suggest that MoS_2 clusters are more effective than WS_2 at enhancing the visible absorbance. Also, note that the absorbance onsets for both MoS_2 and WS_2 are blue shifted relative to the bulk.

Slurries of the MS_2/TiO_2 were tested for the photocatalytic oxidation of a dye, methylene blue, and 4-chlorophenol. In both experiments, a stirred batch reactor was used and oxygen was bubbled through the solution during the experiment. The concentrations of the organic molecules were monitored using HPLC absorbance measurements. The state of the photocatalyst following reaction was not investigated. For both methylene blue and 4-chlorophenol, significant photo-oxidation of the organic occurred using only visible light (400–700 nm). The results from their work

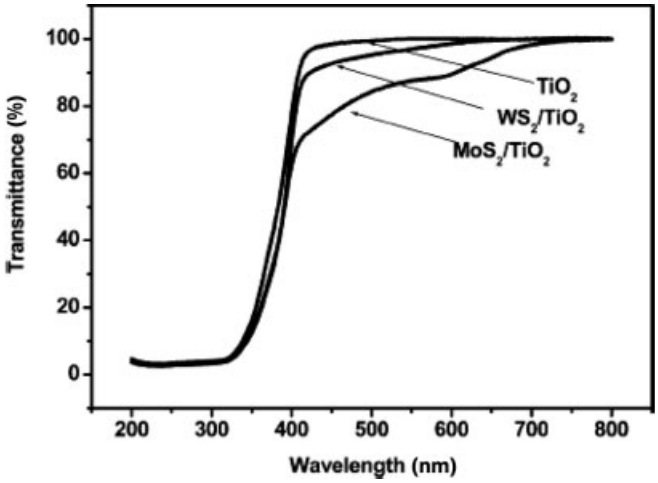


Figure 3.35 Diffuse reflectance spectra of TiO_2, 0.3 mol% of WS_2 on TiO_2 and 0.5 mol% of MoS_2 on TiO_2. (Reprinted with permission from W. K. Ho, J. C. Yu, J. Lin, J. G. Yu, P. S. Li, *Langmuir* **2004**, *20*, 5865).

are shown in Figure 3.36. Either WS_2/TiO_2 or MoS_2/TiO_2 seem equally effective for these photo-oxidation reactions.

They concluded that the effectiveness of the coupled system was indeed due to electron transfer from the MS_2 to the TiO_2 by measurements using electron spin resonance (ESR), where they observed a signal corresponding to Ti(III) radical.

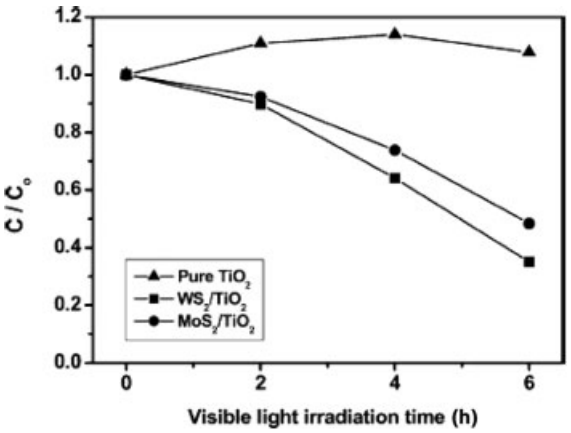


Figure 3.36 Change in concentration of 4-chlorophenol showing photodegradation using MoS_2 and WS_2 nanocluster-sensitized TiO_2 compared with pure TiO_2. Visible light irradiation was used, $\lambda > 400$ nm. (Reprinted with permission from W. K. Ho, J. C. Yu, J. Lin, J. G. Yu, P. S. Li, *Langmuir* **2004**, *20*, 5865).

3.7 Current and Future Technological Applications of Photocatalysts for Environmental Remediation

The photocatalytic reactions discussed thus far occur in water, but the same materials, especially TiO_2, have proved useful for photo-oxidation of VOCs in the air. In fact, this application represents the first commercial use of photocatalysts, which in Japan amounts to nearly US$300 million in revenue and about 2000 companies as of 2003 [50]. Air purification using TiO_2 photocatalysts is very dependent on reactor design, the most common configurations being annular plug flow reactors [156] and honeycomb reactors [157]. The oxidation process functions most efficiently at low concentrations and relatively low air flow rates. High flow rates may be limited by mass transport considerations. For flow rates of less than $20\,000\,ft^3\,min^{-1}$ may suggest that photocatalysis is more cost-effective than carbon adsorption or incineration of VOCs.

Because light penetrates long distances through air, the illumination of the photocatalyst in gas reactors is simplified compared with aqueous phase reactors, especially the most efficient opaque, slurry-type reactors. Also, higher reaction temperatures and pressures are possible, unlike in water purification where the boiling point of water is a limitation. In fact, a combination of air stripping of VOCs combined with gas-phase photocatalytic destruction is likely a viable approach to water purification.

An advantage of photocatalytic gas purification of VOCs is that relatively low levels of light are required. For outdoor applications, for example, ambient conditions of around 2–3 $mW\,cm^{-2}$ are available in the near-UV (UVA) region of the spectrum that can be absorbed by TiO_2. These levels are sufficient for many applications such as self-cleaning tiles and windows [50]. For indoor applications, fluorescent lighting of around 1 $\mu W\,cm^{-2}$ in the UVA region is available in most offices. The efficiency of the photocatalysis has actually been shown to be superior at these levels. For example, the quantum efficiency for photo-oxidation of 2-propanol by TiO_2 was ~28% and for the common indoor pollutant acetaldehyde it was nearly 100%, due to an autocatalytic process producing a free radical chain in the air [50]. Since gas-phase diffusion of both reactants and products takes place continuously, gas-phase photocatalytic processes can be self-cleaning, preserving the integrity of the catalytic surface. It also appears that free radical scavengers such as chlorine ions and electron scavengers such as oxygen are not as significant a problem in air as in water [158]. However, incomplete mineralization of some chemicals can lead to loss of efficiency with time due to build-up of intermediates at the catalyst surface. Water is not available in most indoor applications to wash these away, but the best outdoor designs take advantage of rain to preserve an active catalyst surface [50].

Gas-phase photoreactors may be incorporated into existing heating and air conditioning systems, where they are often utilized in combination with more traditional approaches such as HEPA filters and carbon adsorption. For example, Ao and Lee recently described experiments in which a commercial air cleaner was modified to incorporate TiO_2 on activated carbon filter illuminated using a 6 W UV

lamp (254 nm) to remove nitrous oxide and toluene at ppb levels [159]. The high efficiency of the combined activated carbon–TiO_2 photocatalyst was attributed to the ability of the carbon to adsorb and concentrate the NO pollutant, which then diffuses to the TiO_2 where it is photo-oxidized.

A recent review of the gas-phase photocatalysis by Fujishima and Zhang summarizes some of the most interesting results on gas-phase photo-oxidation using TiO_2 photocatalysts [50]. We will now examine some results from that review for both indoor and outdoor air cleaning.

3.7.1 Indoor Air Purification

The labor costs associated with keeping indoor surfaces free from bacteria and other contamination can be significant in the case of hospitals and nursing homes. With this application in mind, Fujishima's group was able to demonstrate that low UV light levels of around 1 mW cm^{-2} could destroy *E. coli* cells placed on a TiO_2-coated glass plate in <1 h. For comparison, they found that only 50% of the cells were dead following 4 h of illumination under the same conditions. This laboratory work was tested in several operating rooms in the form of antibacterial tiles with a TiO_2 coating. Tests showed that the bacterial levels were negligible after 1 h of illumination under the ambient fluorescent illumination. The bacterial levels in the surrounding air also decreased significantly [50].

The latter result is consistent with observations by Choi that OH radicals generated on the TiO_2 diffuse a significant distance through air and can oxidize soot many microns away from the active surface [22]. The evidence for this is shown in Figure 3.37, where soot was deposited on the surface of a glass substrate coated with TiO_2 photocatalyst. This SEM image shows that a gap develops at the interface between the TiO_2 and the soot. The width, *d*, of this gap was reported to increase continuously with UV illumination time, indicating that the active oxidants formed at the TiO_2 surface desorb and migrate across the glass to reach the soot.

Fujishima's group has investigated self-cleaning surfaces as possible interior wall materials for buildings and homes [160]. Build-up of indoor air pollution in modern, highly insulated homes and offices combined with out-gassing of chemicals such as formaldehyde and urethane used in building materials make such self-cleaning surfaces very attractive. They were able to demonstrate that many volatile organic compounds could be completely mineralized to CO_2 using weak (1 mW cm^{-2}) UV illumination.

Indoor air cleaners based on HEPA and activated carbon adsorption are commonly used to remove foul-smelling VOCs which exist in most indoor environments. They are readily adapted to use photocatalytic oxidation to extend filter life by the self-cleaning property of TiO_2. To maximize the surface area, a honeycomb-type filter which minimizes back-pressure is coated with TiO_2. A picture of such a filter taken from Fujishima and Zhang's review [50] is shown in Figure 3.38. TiO_2 nanoparticles can be embedded in activated carbon as described by Ao and Lee [159] and used in modified air conditioners or air cleaners. This approach permits the adsorbing

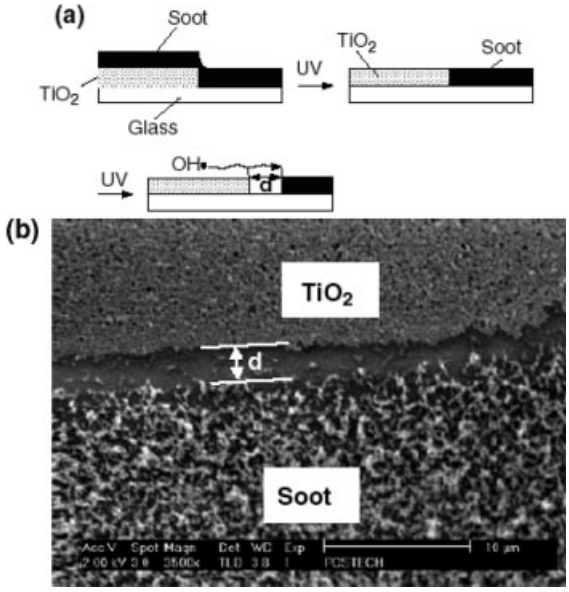


Figure 3.37 (a) Schematic diagram showing the photocatalytic degradation process of soot. Initially the soot is in contact with the TiO_2 catalyst. A gap, of width *d*, develops and grows as a function of irradiation time. (b) SEM of the gap, *d*, after 6 h of irradiation. (Reprinted with permission from W. Choi, *Catalysis Surveys from Asia* **2006**, *10*, 16).

carbon to regenerate itself by mineralization of the adsorbed pollutants. These types of filters can also kill airborne bacteria and viruses.

TiO_2 surfaces have another useful property that was accidentally discovered in the laboratories of TOTO Inc. in 1995. It was found that TiO_2 films having an appropriate

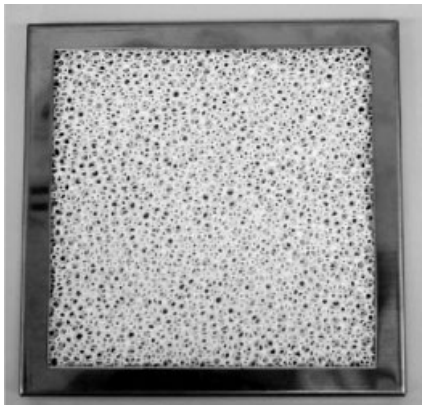


Figure 3.38 TiO_2-coated porous ceramic filters for use as air filters. (Reprinted with permission from A. Fujishima, X. T. Zhang, *Comptes Rendus Chimie* **2006**, *9*, 750).

fraction of SiO_2 incorporated and the exposed to UV light became superhydrophilic, having a zero contact angle with water (i.e. no droplet formation, perfect wetting) [50]. The restructuring of the surface which makes the surface superhydrophilic was subsequently elucidated [49]. This property eventually led to commercial windows coated with very thin, optically transparent TiO_2 films which were both self-cleaning and non-fogging. This is useful for both interior and exterior surfaces of windows.

3.7.2
Outdoor Air Purification

The application of TiO_2 photocatalytic films in exterior construction materials such as tiles is the largest commercial application of photocatalysts at the present time. These materials remain free of contamination by a combination of photocatalytic oxidation driven by sunlight and washing of the breakdown products by rain. The latter is made facile by the excellent wetting properties of the superhydrophilic surface of TiO_2 [50]. Fujishima and Zhang [50] estimated that self-cleaning tiles have been applied to over 5000 buildings in Japan as of 2003. The Japanese company TOTO Inc., which markets these tiles, estimates that the tiles need to be cleaned manually only every 20 years, compared with ordinary tiles which must be cleaned every 5 years. An added benefit is that the tiles can decompose pollutants such as nitrogen oxides from automobile pollution. The tiles are fairly effective, as shown in Figure 3.39 taken from Fujishima and Zhang's review [50], which

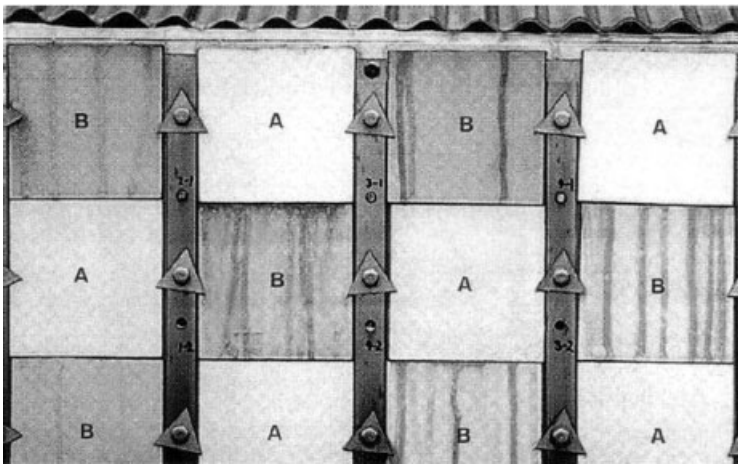


Figure 3.39 Exterior wall showing alternating self-cleaning (A) and ordinary (B) tiles. The TiO_2-coated tiles are obviously cleaner than the non-coated counterpart. (Reprinted with permission from R. Wang, K. Hashimoto, A. Fujishima, M. Chikuni, E. Kojima, A. Kitamura, M. Shimohigoshi, T. Watanabe, *Advanced Materials* **1998**, *10*, 135, and A. Fujishima, X. T. Zhang, *Comptes Rendus Chimie* **2006**, *9*, 750).

shows alternating tiles with and without TiO_2. Water from the roof runs over both types of tiles but the photocatalytic surfaces are much cleaner, as shown in this photograph.

Window glass (SiO_2) can also be made self-cleaning by embedding nanoparticles of TiO_2 in an SiO_2 matrix. An added benefit of this composite material is that water droplet formation does not occur on rainy days due to the superhydrophilic properties of TiO_2. This property is also very useful for the windshields of cars and motorcycles. With just a little rainfall a thin, uniform film forms which rapidly evaporates. Even in heavy rainfall droplet formation that causes light scattering is avoided since a uniform film of water forms on the glass. Since TiO_2 has a much higher refractive index than glass, with a continuous film on glass (i.e. SiO_2) the composition of the TiO_2/SiO_2 film can be carefully controlled to avoid excess refraction of light and which results in visual clarity [50].

Plastic tents made of materials such as PVC are also widely used in outdoor applications such as temporary buildings for exhibits or storage. Their flexible surfaces are difficult to clean, so Kangyo Co. in Japan coated PVC tents with TiO_2, leading to the formation of an inorganic/organic interfacial microstructure between the TiO_2 layer and the PVC to prevent photo-oxidation of the PVC. This approach is fairly effective, as shown in Figure 3.40 taken from Fujishima and Zhang's review [50], where a tent made partly of ordinary PVC and TiO_2-coated PVC was photographed by Taiyl Kangyo Co. after 2 months of exposure to air contamination [50].

Common construction materials such as cement can also be modified by the addition of TiO_2 photocatalysts [161]. Provided that the addition of the photocatalyst particles does not adversely affect the curing and final strength of the material, this modification is inexpensive and reduces maintenance costs by its self-cleaning ability. The preservation of ancient Greek statues against urban air pollutants is one novel example of the utility of these composite materials [13, 162].

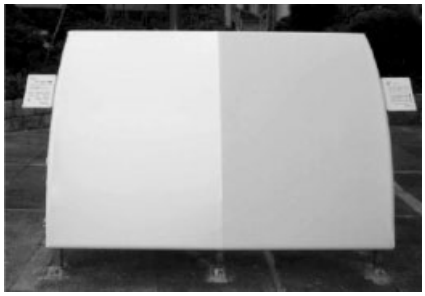


Figure 3.40 Photograph of a tent made by Taiyo Kangyo Co. where the left half was coated with TiO_2 and the right half was not coated. The left half is clearly cleaner after 2 months of outdoor exposure. (Reprinted with permission from A. Fujishima, X. T. Zhang, *Comptes Rendus Chimie* **2006**, *9*, 750).

3.8 Conclusion

Nanosized photocatalysts for environmental remediation can be a viable approach to photo-oxidation and removal of organic and inorganic pollutants in water and air. This process for photocatalytic oxidation of pollutants produces relatively benign products such as CO_2 and dilute mineral acids under ambient conditions. Potentially, it can be driven entirely by sunlight, minimizing its economic costs. Since very broad classes of chemicals can be destroyed, this approach is quite general. However, two main future research challenges were identified in our review, improvement of the photocatalyst efficiency and practical reactor designs for specific remediation problems.

In this chapter, we have emphasized the interaction of new synthetic methods for the production of nanosized semiconductors. Our review focused most on two photocatalytic materials. The first is TiO_2, which has been investigated as a photocatalyst for over 30 years and has an extensive literature with several excellent reviews. The other material, MoS_2, has been investigated for this application for only around 10 years but its structural and size-dependent optical and electronic properties suggest that it may prove as useful as TiO_2 in the future. It is an excellent example of how new nanomaterials can dramatically affect technical progress in environmental remediation.

The main limitation of TiO_2 as a photocatalyst has been identified as its wide bandgap, which limits its absorbance to 3–5% of the solar spectrum. Much effort has been made to improve its absorption efficiency through substitutional doping with elements such as nitrogen to extend its absorption range into the visible region. Unfortunately, a loss of efficiency accompanies the use of these longer wavelength photons and this has not yet been overcome. Other synthetic modifications to TiO_2 which were reviewed included deposition of metal islands and the use of sensitizers such as dyes or inorganic ions such as Fe(III) to extend its light absorption range. No single method was effective for all types of reactions. However, the use of metal island deposits on TiO_2 was shown to improve the efficiency of photo-oxidation for several types of difficult pollutants such as chlorinated aromatics. The key observation was that an optimum metal loading exists to optimize the photocatalytic activity. Mechanistically, it appears that atoms at the perimeter between metal islands and TiO_2 play a key role in the activity. Various experiments including transient absorption and electron spin resonance indicate that the role of the metal islands is to facilitate the transfer of electrons to species such as molecular oxygen, thus reducing the recombination of electron–hole pairs. This indicates how the proper choice of characterization methods can lead to insights into and better scientific understanding of the photocatalytic process.

It would be surprising if a single material such as TiO_2 could serve as a universal photocatalyst despite its low cost and photochemical stability. We demonstrated that layered metal dichalcogenides such as MoS_2 share its photochemical stability and low costs and have the additional feature of having size-tunable light absorption and redox potentials. The size-dependent electronic properties of nanosized semiconductors arise from carrier confinement in the small particle. We showed that such photocatalysts that normally absorb light in the near-IR region have a blue shift of their absorbance onset into the visible region and their redox potentials become favorable

enough to allow the complete mineralization of a toxic chlorinated organic such as pentachlorophenol using only visible light. The rate of photo-oxidation was shown to depend on photocatalyst size. A surprising observation was that the presence of a stabilizing surfactant on the cluster surface did not quench the photo-oxidation. In fact, addition of this surfactant to a common, very active TiO_2 photocatalyst also improved its photocatalytic activity. We recommend such surface modifications of nanosized semiconductors as a fruitful avenue for future research. Clusters in the size range 1–3 nm should be the best candidates for such surface tuning of their electronic states, since 60–90% of all their atoms reside in surface positions.

Synthetic strategies can be facilitated by feedback from traditional and novel characterization methods. As an example of the latter, we demonstrated how liquid chromatography could be used to follow the course of a photo-oxidation reaction and also to characterize the chemistry and size of the MoS_2 photocatalyst as a function of illumination time. Such *in situ* photocatalyst characterization is not possible with agglomerated slurry photocatalysts and is an advantage of fully dispersed nanoclusters as photocatalysts. The absorption spectra of nanosized photocatalysts is so sensitive to changes in the size or cluster surface chemistry that deactivation of the catalyst from agglomeration or surface changes can be rapidly monitored. We also discussed the use of techniques such as dynamic light scattering to follow photocatalyst agglomeration, a common process which reduces the available surface area and photoactivity.

We reviewed experiments in which two coupled nanosized semiconductors in physical contact can transfer photogenerated carriers from one to the other, leading to improved charge separation, reduced recombination and improved photocatalytic oxidation. Two example systems of MoS_2/TiO_2 were discussed. In these systems, the presence of nanosized MoS_2 on the TiO_2 surface allowed visible light to be used for the photo-oxidation. Such a scheme exploits the best characteristics of each material in addition to providing for sensitive tuning of the chemical adsorption properties. It is also a requirement for a practical reactor since nanosized photocatalysts must be immobilized to prevent them from acting as pollutants themselves. In the future we anticipate more studies of such composite systems as they allow scientists great latitude in photocatalyst design.

We described photocatalysis experiments using two types of reactor design: batch slurry reactors and fixed-bed flow reactors. Most basic research has used the slurry type since this design is not as sensitive to illumination geometry and exposes the maximum amount of catalyst surface to the pollutant. It is also the simplest design. However, such a design cannot be used in the field for large-scale treatment due to the costs and complexities associated with recovering the photocatalyst from the treated water. Instead, immobilization of the catalyst nanoparticles as a porous film on a high surface area support material is required. A honeycomb design is a common choice but ensuring complete illumination of the photocatalyst is difficult. The best geometry and light illumination for such a reactor and also optimal flow rates are specific to a given remediation problem and chemical. Since the kinetics of photo-oxidation will depend on chemical type and concentration, the ability to vary the flow rate and monitor the treated water are key aspects of reactor design. More research

into reactor design will be required before photocatalytic remediation can compete with current methods for water treatment. High surface area materials such as carbon aerogels are worthy of consideration as supports.

Even though remediation schemes using TiO_2 to treat liquid-phase pollutants are not currently economically viable, we presented recent technological applications of TiO_2 porous films to the continuous treatment of indoor and outdoor air pollution. The key observation for these gas-phase reactions was the high efficiency for gas-phase reactions at low light levels. This is fortunate since the light flux available in the near-UV region in natural sunlight can be used with TiO_2 films in a variety of outdoor applications, including self-cleaning tiles and glass. In indoor applications, there is sufficient near-UV light available due to fluorescent lighting that air cleaning and purification using TiO_2 impregnated on carbon is a useful method of removing noxious odors and killing bacteria. The combination of conventional carbon adsorption with the ability of TiO_2 photocatalysts to oxidize the adsorbed chemicals extends the life of the filters and invigorates and extends an older technology. Such applications should dominate the short-term uses of photocatalysts in environmental remediation.

As a final note, not only is environmental remediation a technical and scientific problem, it is also a social problem. The general population also needs to evaluate their habits and curb any contributions they may be making to the pollution problem. Environmental pollution is very much related to many other issues such as habitat destruction, overpopulation, lack of education, excessive consumerism, extreme poverty and corruption within governments and corporations. None of these problems can be solved in a completely isolated manner. Scientific and technical solutions exist in many of these areas and particularly, as outlined here in this chapter, for environmental remediation through photocatalysis. The use of nanomaterials as photocatalysts is certainly promising. However, we must be careful that we do not merely exchange one problem for another when using nanomaterials if there is a possibility that they may also become future pollutants. Evaluating the actual impact of new discoveries is the responsibility of all researchers if they wish to utilize their discoveries in any applied systems.

Acknowledgment

This work was supported by the Division of Materials Sciences, Office of Basic Energy Research, U.S. Department of Energy under contract DE-AC04-AL8500. The Center for Individual Nanoparticle Functionality (CINF) is supported by the Danish National Research Council.

References

1 S. A. Ostroumov, *Biological Effects of Surfactants*, Taylor & Francis Group, Boca Raton, FL, 2006.

2 R. P. Schwarzenbach, B. I. Escher, K. Fenner, T. B. Hofstetter, C. A. Johnson, U. von Gunten, B. Wehrli, *Science* 2006, **313**, 1072.

3 T. Ohe, T. Watanabe, K. Wakabayashi, *Mutation Research* 2004, **567**, 109.

4 US Environmental Protection Agency, *Resource Conservation and Recovery Act*, http://www.epa.gov/region5/defs/html/rcra.htm, 2007.

5 US Environmental Protection Agency, *Toxic Release Inventory – Industry Report*, http://www.epa.gov/triexplorer/industry.htm, 2006.

6 T. O'Neill, *National Geographic Magazine* 2007, February, 88.

7 US Environmental Protection Agency, *Superfund Information Systems*, http://cfpub.epa.gov/supercpad/cursites/srchsites.cfm, 2006.

8 W. X. Zhang, *Journal of Nanoparticle Research* 2003, **5**, 323.

9 Z. K. Liu, Y. L. He, F. Li, Y. H. Liu, *Environmental Science and Pollution Research* 2006, **13**, 328.

10 S. Babel, D. del Mundo Dacera, *Waste Management* 2006, **26**, 988.

11 J. Scullion, *Naturwissenschaften* 2006, **93**, 51.

12 A. S. Sheoran, *Minerals Engineering* 2006, **19**, 105.

13 O. Carp, C. L. Huisman, A. Reller, *Progress in Solid State Chemistry* 2004, **32**, 33.

14 M. A. Fox, M. T. Dulay, *Chemical Reviews* 1993, **93**, 341.

15 M. Gratzel, *Heterogeneous Photochemical Electron Transfer*, CRC Press, Boca Raton, FL, 1989.

16 R. F. Howe, *Developments in Chemical Engineering and Mineral Processing* 1998, **6**, 55.

17 A. Mills, S. Le Hunte, *Journal of Photochemistry and Photobiology A – Chemistry* 1997, **108**, 1.

18 A. Fujishima, K. Honda, *Nature* 1972, **238**, 37.

19 B. L. Abrams, J. P. Wilcoxon, *Critical Reviews in Solid State and Materials Sciences* 2005, **30**, 153.

20 D. W. Bahnemann, *Israel Journal of Chemistry* 1993, **33**, 115.

21 D. Beydoun, R. Amal, G. Low, S. McEvoy, *Journal of Nanoparticle Research* 1999, **1**, 439.

22 W. Choi, *Catalysis Surveys from Asia* 2006, **10**, 16.

23 M. Gratzel, in N. Serpone, E. Pelizzetti (eds.), *Photocatalysis: Fundamentals and Applications*, Wiley, New York, 1989, 123.

24 A. Hagfeldt, M. Gratzel, *Chemical Reviews* 1995, **95**, 49.

25 J. M. Herrmann, *Topics in Catalysis* 2006, **39**, 3.

26 M. R. Hoffmann, S. T. Martin, W. Y. Choi, D. W. Bahnemann, *Chemical Reviews* 1995, **95**, 69.

27 P. V. Kamat, D. Meisel, *Current Opinion in Colloid and Interface Science* 2002, **7**, 282.

28 P. V. Kamat, D. Meisel, *Comptes Rendus Chimie* 2003, **6**, 999.

29 O. Legrini, E. Oliveros, A. M. Braun, *Chemical Reviews* 1993, **93**, 671.

30 B. Levy, *Journal of Electroceramics* 1997, **1**, 239.

31 M. I. Litter, *Applied Catalysis B – Environmental* 1999, **23**, 89.

32 S. O. Obare, G. J. Meyer, *Journal of Environmental Science and Health – Part A* 2004, **39**, 2549.

33 N. Savage, M. S. Diallo, *Journal Of Nanoparticle Research* 2005, **7**, 331.

34 N. Serpone, R. F. Khairutdinov, Application of nanoparticles in the photocatalytic degradation of water pollutants. In: *Semiconductor Nanoclusters, Studies in Surface Science and Catalysis* (Eds. P.V. Kamat and D. Meisel), Elsevier Science B.V., 1996, 417.

35 W. A. Zeltner, M. A. Anderson, Nato Advanced Science Institute Series, Sub-Series 3, High Technology, NATO Advanced Research Workshop on Fine Particles Science and Technology – From Micro to Nanoparticles; July 15–21, 1995; Acquafredda di Maratea, Italy 1996, 12, 643.

36 Umicore, *Short History*, http://www.umicore.com/en/aboutUs/history/, 2005.

37 N. Serpone, *Solar Energy Materials and Solar Cells* 1995, **38**, 369.

38 T. E. Agustina, H. M. Ang, V. K. Vareek, *Journal of Photochemistry and Photobiology C – Photochemistry Reviews* 2005, **6**, 264.

39 Jefferson Laboratories, *The 10 Most Abundant Elements in the Earth's Crust*, http://education.jlab.org/glossary/abund_ele.html, 2007.

40 A. L. Linsebigler, G. Q. Lu, J. T. Yates, *Chemical Reviews* 1995, **95**, 735.

41 N. Serpone, E. Pelizzetti, in *Photocatalysis Fundamentals and Applications*, Wiley, New York, 1989, vii.

42 S. C. Markham, *Journal of Chemical Education* 1955, **32**, 540.

43 S. N. Frank, A. J. Bard, *Journal of Physical Chemistry* 1977, **81**, 1484.

44 S. N. Frank, A. J. Bard, *Journal of the American Chemical Society* 1977, **99**, 4667.

45 C. Y. Hsiao, C. L. Lee, D. F. Ollis, *Journal of Catalysis* 1983, **82**, 418.

46 A. L. Pruden, D. F. Ollis, *Journal of Catalysis* 1983, **82**, 404.

47 T. Matsunaga, R. Tomoda, T. Nakajima, H. Wake, *FEMS Microbiology Letters* 1985, **29**, 211.

48 A. Fujishima, R. X. Cai, J. Otsuki, K. Hashimoto, K. Itoh, T. Yamashita, Y. Kubota, *Electrochimica Acta* 1993, **38**, 153.

49 R. Wang, K. Hashimoto, A. Fujishima, M. Chikuni, E. Kojima, A. Kitamura, M. Shimohigoshi, T. Watanabe, *Nature* 1997, **388**, 431.

50 A. Fujishima, X. T. Zhang, *Comptes Rendus Chimie* 2006, **9**, 750.

51 J. P. Wilcoxon, P. P. Newcomer, G. A. Samara, *Journal of Applied Physics* 1997, **81**, 7934.

52 T. R. Thurston, J. P. Wilcoxon, *Journal of Physical Chemistry B* 1999, **103**, 11.

53 I. Chorkendorff, J. W. Niemantsverdriet, *Concepts of Modern Catalysis and Kinetics*, Wiley-VCH, Weinheim, 2003.

54 S. T. Martin, H. Herrmann, W. Y. Choi, M. R. Hoffmann, *Journal of the Chemical Society: Faraday Transactions* 1994, **90**, 3315.

55 A. Orlov, M. S. Chan, D. A. Jefferson, D. Zhou, R. J. Lynch, R. M. Lambert, *Environmental Technology* 2006, **27**, 747.

56 N. Sobana, M. Muruganadham, M. Swaminathan, *Journal of Molecular Catalysis A – Chemical* 2006, **258**, 124.

57 W. Y. Choi, A. Termin, M. R. Hoffmann, *Journal of Physical Chemistry* 1994, **98**, 13669.

58 Z. B. Zhang, C. C. Wang, R. Zakaria, J. Y. Ying, *Journal of Physical Chemistry B* 1998, **102**, 10871.

59 L. Palmisano, M. Schiavello, A. Sclafani, C. Martin, I. Martin, V. Rives, *Catalysis Letters* 1994, **24**, 303.

60 N. Serpone, E. Borgarello, M. Gratzel, *Journal of the Chemical Society: Chemical Communications* 1984, 342.

61 M. Abdullah, G. K. C. Low, R. W. Matthews, *Journal of Physical Chemistry* 1990, **94**, 6820.

62 M. Bekbolet, Z. Boyacioglu, B. Ozkaraova, *Water Science and Technology* 1998, **38**, 155.

63 O. J. Jung, *Bulletin of the Korean Chemical Society* 2001, **22**, 1183.

64 J. C. D'Oliveira, G. Alsayyed, P. Pichat, *Environmental Science and Technology* 1990, **24**, 990.

65 L. Zang, C. Lange, I. Abraham, S. Storck, W. F. Maier, H. Kisch, *Journal of Physical Chemistry B* 1998, **102**, 10765.

66 M. Bideau, B. Claudel, L. Faure, H. Kazouan, *Journal of Photochemistry and Photobiology A – Chemistry* 1991, **61**, 269.

67 M. Bideau, B. Claudel, L. Faure, H. Kazouan, *Journal of Photochemistry and Photobiology A – Chemistry* 1992, **67**, 337.

68 J. P. Wilcoxon, *Journal of Physical Chemistry B* 2000, **104**, 7334.

69 J. C. Crittenden, Y. Zhang, D. W. Hand, D. L. Perram, E. G. Marchand, *Water Environment Research* 1996, **68**, 270.

70 M. Barbeni, E. Pramauro, E. Pelizzetti, E. Borgarllo, N. Serpone, *Chemosphere* 1985, **14**, 195.

71 B. A. Korgel, H. G. Monbouquette, *Journal of Physical Chemistry B* 1997, **101**, 5010.

72 E. Pelizzetti, B.M., E. Pramauro, N. Serpone, B.E., M.A. Jamieson, H. Hidaka, *Chimica e l'Industria* 1985, **67**, 623.

73 D. F. Ollis, E. Pelizzetti, N. Serpone, *Environmental Science and Technology* 1991, **25**, 1522.

74 A. M. Braun, E. Oliveros, *Water Science and Technology* 1997, **35**, 17.

75 A. Henglein, *Topics in Current Chemistry* 1988, **143**, 113.

76 A. Henglein, *Chemical Reviews* 1989, **89**, 1861.

77 H. Frohlich, *Proceedings of the Royal Society of London Series A* 1939, **171**, 496.

78 H. Frohlich, *Proceedings of the Royal Society of London Series A* 1937, **160**, 230.

79 N. Serpone, D. Lawless, R. Khairutdinov, *Journal of Physical Chemistry* 1995, **99**, 16646.

80 A. L. Stroyuk, A. I. Kryukov, S. Y. Kuchmii, V. D. Pokhodenko, *Theoretical and Experimental Chemistry* 2005, **41**, 207.

81 A. D. Yoffe, *Advances in Physics* 1993, **42**, 173.

82 L. E. Brus, *Journal of Chemical Physics* 1983, **79**, 5566.

83 F. W. Wise, *Accounts of Chemical Research* 2000, **33**, 773.

84 F. Parsapour, D. F. Kelley, S. Craft, J. P. Wilcoxon, *Journal of Chemical Physics* 1996, **104**, 4978.

85 U. Kreibig, M. Vollmer, *Optical Properties of Metal Nanocluster*, Vol. 25, Berlin, Springer, 1995.

86 D. Lawless, N. Serpone, D. Meisel, *Journal of Physical Chemistry* 1991, **95**, 5166.

87 C. Hariharan, *Applied Catalysis A – General* 2006, **304**, 55.

88 J. P. Wilcoxon, J. E. Martin, P. Provencio, *Langmuir* 2000, **16**, 9912.

89 J. P. Wilcoxon, J. E. Martin, P. Provencio, *Journal of Chemical Physics* 2001, **115**, 998.

90 J. P. Wilcoxon, P. Provencio, *Journal of Physical Chemistry B* 2005, **109**, 13461.

91 J. H. Jean, T. A. Ring, *Langmuir* 1986, **2**, 251.

92 K. Tanaka, M. F. V. Capule, T. Hisanaga, *Chemical Physics Letters* 1991, **187**, 73.

93 H. Gerischer, A. Heller, *Journal of the Electrochemical Society* 1992, **139**, 113.

94 H. P. Boehm, M. Herrmann, *Zeitschrift für Anorganische und Allgemeine Chemie* 1967, **352**, 156.

95 R. Asahi, T. Morikawa, T. Ohwaki, K. Aoki, Y. Taga, *Science* 2001, **293**, 269.

96 T. Morikawa, R. Asahi, T. Ohwaki, K. Aoki, Y. Taga, *Japanese Journal of Applied Physics Part 2 – Letters* 2001, **40**, L561.

97 T. Morikawa, Y. Irokawa, T. Ohwaki, *Applied Catalysis A – General* 2006, **314**, 123.

98 J. Chen, D. F. Ollis, W. H. Rulkens, H. Bruning, *Water Environment Research* 1999, **33**, 661.

99 M. Haruta, *Catalysis Today* 1997, **36**, 153.

100 M. S. Chan, R. J. Lynch, *Environmental Chemistry Letters* 2003, **1**, 157.

101 P. V. Kamat, *Chemical Reviews* 1993, **93**, 267.

102 M. Yang, D. W. Thompson, G. J. Meyer, *Inorganic Chemistry* 2002, **41**, 1254.

103 H. Kisch, L. Zang, C. Lange, W. F. Maier, C. Antonius, D. Meissner, *Angewandte Chemie International Edition* 1998, **37**, 3034.

104 W. Macyk, H. Kisch, *Chemistry – A European Journal* 2001, **7**, 1862.

105 L. Zang, W. Macyk, C. Lange, W. F. Maier, C. Antonius, D. Meissner, H. Kisch, *Chemistry – A European Journal* 2000, **6**, 379.

106 V. Dijken, A. H. Janssen, M. H. P. Smitsmans, D. Vanmaekelbergh, K. Meijerink, *Chemistry of Materials* 1998, **10**, 3513.

107 S. Dindar, J. Icli, *Photochemistry and Photobiology A* 2001, **140**, 263.

108 P. V. Kamat, R. Huehn, R. Nicolaescu, *Journal of Physical Chemistry B* 2002, **106**, 788.

109 Q. Zhang, L. Gao, *Langmuir* 2004, **20**, 9821.

110 H. Tributsch, *Zeitschrift für Naturforschung Teil A* 1977, **32**, 972.

111 International Molybdenum Association, *About Molybdenum*, http://www.imoa.info/, 2003.

112 J. W. Blossom, *Molybdenum*, http://minerals.usgs.gov/minerals/pubs/

commodity/molybdenum/470798.pdf, 1998.

113 E. Graber, A. Klingsborg, P. M. Siegal, *Title?*, Wiley, New York, 1985, 774.

114 H. Topsøe, B. S. Clausen, F. E. Massoth, *Hydrotreating Catalysis Science and Technology*, Springer, Berlin, 1996.

115 R. F. Frindt, A. D. Yoffe, *Proceedings of the Royal Society of London Series A* 1963, **273**, 69.

116 R. G. Dickinson, L. Pauling, *Journal of the American Chemical Society* 1923, **45**, 1466.

117 R. Hultgren, *Physical Review* 1932, **40**, 891.

118 L. Pauling, *The Nature of the Chemical Bond and the Structure of Molecules and Crystals: an Introduction to Modern Structural Chemistry*, 3rd edn., Cornell University Press, Ithaca, NY, 1960.

119 E. Benavente, M. A. Santa Ana, F. Mendizabal, G. Gonzalez, *Coordination Chemistry Reviews* 2002, **224**, 87.

120 R. Coehoorn, C. Haas, J. Dijkstra, C. J. F. Flipse, R. A. de Groot, A. Wold, *Physical Review B:Condensed Matter* 1987, **35**, 6195.

121 N. Ohmae, in *Wear, 1st International Workshop on Microtribology (IWM), Proceedings of the 1st International Workshop on Microtribology (IWM), October 12–13 1992, Morioka, Japan, Vol. 168*, Switzerland, 1993, Morioka, Japan, 1993, 99.

122 L. Scandella, A. Schumacher, N. Kruse, R. Prins, E. Meyer, R. Luthi, L. Howald, H. J. Guntherodt, *Thin Solid Films* 1994, **240**, 101.

123 M. V. Bollinger, K. W. Jacobsen, J. K. Norskov, *Physical Review B* 2003, **67**, 085410.

124 H. Tributsch, J. C. Bennett, *Journal of Electroanalytical Chemistry* 1977, **81**, 97.

125 B. L. Evans, P. A. Young, *Proceedings of the Physical Society of London* 1967, **91**, 475.

126 B. L. Evans, P. A. Young, *Proceedings of the Royal Society of London Series A* 1965, **284**, 402.

127 A. M. Goldberg, A. R. Beal, F. A. Levy, E. A. Davis, *Philosophical Magazine* 1975, **32**, 367.

128 H. P. Hughes, W. Y. Liang, *Journal of Physics C: Solid State Physics* 1974, **7**, 1023.

129 V. V. Sobolev, V. I. Donetski, A. A. Opalovsk, V. E. Fedorov, E. U. Lobkov, A. P. Mazhara, *Soviet Physics Semiconductors – USSR* 1971, **5**, 909.

130 M. G. Bell, W. Y. Liang, *Advances in Physics* 1976, **25**, 53.

131 A. J. Grant, T. M. Griffiths, G. D. Pitt, A. D. Yoffe, *Journal of Physics C: Solid State Physics* 1975, **8**, L17.

132 K. K. Kam, B. A. Parkinson, *Journal of Physical Chemistry* 1982, **86**, 463.

133 C. B. Roxlo, R. R. Chianelli, H. W. Deckman, A. F. Ruppert, P. P. Wong, *Journal of Vacuum Science and Technology A* 1987, **5**, 555.

134 R. Coehoorn, C. Haas, R. A. de Groot, *Physical Review B: Condensed Matter* 1987, **35**, 6203.

135 L. F. Schneemeyer, M. S. Wrighton, *Journal of the American Chemical Society* 1979, **101**, 6496.

136 G. Kline, K. Kam, D. Canfield, B. A. Parkinson, *Solar Energy Materials* 1981, **4**, 301.

137 G. Kline, K. K. Kam, R. Ziegler, B. A. Parkinson, *Solar Energy Materials* 1982, **6**, 337.

138 R. Tenne, A. Wold, *Applied Physics Letters* 1985, **47**, 707.

139 G. Prasad, O. N. Srivastava, *Journal of Physics D – Applied Physics* 1988, **21**, 1028.

140 A. DiPaola, L. Palmisano, M. Derrigo, V. Augugliaro, *Journal of Physical Chemistry B* 1997, **101**, 876.

141 J. P. Wilcoxon,Method for the preparation of metal colloids in inverse micelles and product preferred by the method, *US Patent* 5 147 841, 1992.

142 J. P. Wilcoxon,Photo-oxidation method using MoS_2 nanocluster materials, *US Patent* 6 245 200, 2001.

143 J. P. Wilcoxon, A. Martino, R. L. Baughmann, E. Klavetter, A. P. Sylwester, *Materials Research Society Symposium Proceedings* 1993, **286**, 131.

144 J. P. Wilcoxon, G. A. Samara, *Physical Review B: Condensed Matter* 1995, **51**, 7299.

145 S. Helveg, J. V. Lauritsen, E. Laegsgaard, I. Stensgaard, J. K. Norskov, B. S. Clausen, H. Topsoe, F. Besenbacher, *Physical Review Letters* 2000, **84**, 951.

146 J. V. Lauritsen, M. Nyberg, J. K. Norskov, B. S. Clausen, H. Topsoe, E. Laegsgaard, F. Besenbacher, *Journal of Catalysis* 2004, **224**, 94.

147 J. V. Lauritsen, M. Nyberg, R. T. Vang, M. V. Bollinger, B. S. Clausen, H. Topsoe, K. W. Jacobsen, E. Laegsgaard, J. K. Norskov, F. Besenbacher, *Nanotechnology* 2003, **14**, 385.

148 J. V. Lauritsen, J. Kibsgaard, S. Helveg, H. Topsoe, B. S. Clausen, E. Laegsgaard, F. Besenbacher, *Nature Nanotechnology* 2007, **2**, 53.

149 N. Bertram, J. Cordes, Y. D. Kim, G. Gantefor, S. Gemming, G. Seifert, *Chemical Physics Letters* 2006, **418**, 36.

150 F. Consadori, R. F. Frindt, *Physical Review B: Condensed Matter* 1970, **2**, 4893.

151 N. F. Mott, *Proceedings of the Royal Society of London Series A* 1938, **167**, 384.

152 J. P. Wilcoxon, T. R. Thurston, in Materials Research Society Symposium Proceedings 1999, 548, 119.

153 G. Mills, M. R. Hoffmann, *Environmental Science and Technology* 1993, **27**, 1681.

154 N. Serpone, P. Maruthamuthu, P. Pichat, P.E., H. Hidaka, *Journal of Photochemistry and Photobiology A – Chemistry* 1995, 85, 247.

155 W. K. Ho, J. C. Yu, J. Lin, J. G. Yu, P. S. Li, *Langmuir* 2004, **20**, 5865.

156 L. A. Dibble, G. B. Raupp, *Environmental Science and Technology* 1992, **26**, 492.

157 M. L. Sauer, D. F. Ollis, *Journal of Catalysis* 1994, **149**, 81.

158 R. M. Alberici, W. F. Jardim, *Water Research* 1994, **28**, 1845.

159 C. H. Ao, S. C. Lee, *Chemical Engineering Science* 2005, **60**, 103.

160 T. Minabe, D. A. Tryk, P. Sawunyama, Y. Kikuchi, K. Hashimoto, A. Fujishima, *Journal of Photochemistry and Photobiology A – Chemistry* 2000, **137**, 53.

161 M. Lackhoff, X. Prieto, N. Nestle, F. Dehn, R. Niessner, *Applied Catalysis B – Environmental* 2003, **43**, 205.

162 I. Poulios, P. Spathis, A. Grigoriadou, K. Delidou, P. Tsoumparis, *Journal of Environmental Science and Health Part A* 1999, **34**, 1455.

4
Pollution Treatment, Remediation and Sensing

Abhilash Sugunan and Joydeep Dutta

4.1
Introduction

Environmental monitoring is becoming increasingly critical to protect the public and the environment from toxic contaminants and pathogens released into air, soil and water from toxic chemical wastes, spills, manufacturing waste and even underground storage tanks. The US Environmental Protection Agency (EPA) has imposed strict regulations on the maximum allowable concentrations of many environmental contaminants in air and water and is reported to have been monitoring over two million underground storage tanks containing hazardous (and often volatile) contaminants from as early as 1992 [1]. Nanotechnology has the potential to bring in solutions to minimize or eliminate the use of toxic materials and the generation of undesirable by-products, and also sensitively detect (and monitor) specific polluting agents well before any major environmental catastrophes occur. Research related to improved industrial processes and starting material requirements, development of new chemical and industrial procedures and materials to replace current hazardous constituents and processes, resulting in reductions in energy, materials and waste generation are being supplemented by the application of nanotechnology to control and predict the potential damage to the environment.

Futuristic examples of types of nanotechnology applications that may lead to reduction or elimination of pollutants of concern include atomic-level synthesis of new and improved catalysts for industrial processes; adding information into molecules (analogous to DNA) that senses toxic molecules; self-assembling molecules as the foundation for new chemicals and materials for toxic waste detection and prevention; and building molecules "just in time" in microscale reactors and on-line sensitive sensors for monitoring and catastrophe prediction and prevention. More contemporary possibilities include facile and accurate detection/monitoring of common airborne pollutants such as NO_x and CO [2], waterborne harmful agents such as pathogens [3] and metal ions [4, 5], amongst others. Monitoring hazardous materials with current methods is costly and time intensive and several limitations in sampling and testing with analytical

Nanotechnology. Volume 2: Environmental Aspects. Edited by Harald Krug

ISBN: 978-3-527-31735-6

Table 4.1 Field screening and monitoring technologies [7].

Most mature
Gas chromatography
X-ray fluorescence spectrometry
Photoionization devices
Flame ionization devices
Catalytic surface oxidation
Mass spectrometry
Infrared spectroscopy
Wet chemistry methods
Kits based on immunoassays and chemical reactions

techniques have been identified. The time and expense involved in the detection of environmental pollutants (i.e. sample acquisition, sample preparation and laboratory analysis) have led to renewed interest in finding newer solutions to analyze contamination in order to prevent, to seek remedial action for or to destroy the contaminants prior to pollution of the environment. Fast and cost-effective field-analytical technologies that can increase the number of analyses and drastically reduce the time required to perform them will help in the prevention of environmental catastrophes. Increasing the amount of analytical data tends to improve the accuracy of hazardous waste site characterization, leading to better management of the problems and the risk assessments can be improved by efficient clean-up procedures [6]. The different analytical techniques for contamination monitoring and testing that are widely used today are listed in Table 4.1 [7].

Environmental monitoring is a complex process involving hundreds of different substances that are deemed to pose a threat to the environment and can occur in the gaseous, liquid or solid phases with concentrations varying from a few percentage levels down to a few parts per trillion (ppt). Monitoring in both the external environment and at the point of discharge and sometimes in real time is necessary to prevent pollution, find remedial effects or decide when to destroy environmentally dangerous substances. There is a critical and growing need for more cost-effective and rapid techniques for the identification and quantification of pollutants in complex environmental matrices and for the conversion of contaminants into benign forms or their complete elimination, and nanotechnology has the promise to fill this need. Nanotechnology is being applied to bridge the need for accurate, inexpensive, sensitive and real-time, *in situ* analyses using robust sensors based on advantages delivered by the new techniques which can be remotely operated through satellite signals.

Although a number of chemical sensors are commercially available for field measurements of chemical species (e.g. portable gas chromatographs, surface-wave acoustic sensors, optical instruments), few are really suitable for continuous environmental and pollution control applications (Table 4.1). Detection of low concentrations for the monitoring of volatile organic compounds (VOCs) such as aromatic hydrocarbons (e.g. benzene, toluene, xylenes), halogenated hydrocarbons [e.g. trichloroethylene (TCE), carbon tetrachloride] and aliphatic hydrocarbons (e.g. hexane, octane) in air, groundwater and other saturated environments is

Table 4.2 Chemical classes of priority hazardous substances[a].

Compound class	%[b]
Volatile organic compounds (VOCs)	26.5
Inorganic elements/radionuclides	17.5
Phenols/phenoxy acids	10.5
Polycyclic aromatic hydrocabons (PAHs)	8.5
Halogenated pesticides/related compounds	8.5
Nitrosoamines/ethers/alcohols	7.5
Reactive intermediates	6.0
Miscellaneous	6.0
Benzidines/aromatic amines	4.0
Phthalates	3.0
Organophosphates/carbamates	2.0

[a]Compiled and published by the Agency of Toxic Substance and Disease Registry (PHS, Annual Report, 1990).
[b]Contribution to the US hazardous waste problem from a human exposure perspective.

urgently needed for the proper monitoring of the environment and prevention of further pollution. Volatile organic compounds from cigarette smoke, building materials, furnishings, cleaning compounds, dry cleaning agents, paints, glues, cosmetics, textiles and combustion sources are also a major source of indoor air pollution [8]. Nanotechnology has already been applied to remove some of these VOCs; it has already been reported that an ultraviolet (UV) illuminated titanium dioxide (TiO_2) catalytic surface can produce an overall reduction in air VOC levels [9]. Low-temperature activity of gold catalysts has been employed by Mintek in South Africa to construct a prototype air purification unit which removes carbon monoxide from the air at room temperature [10].

Over 700 chemical species have been identified at hazardous waste sites and the still unidentified compounds may number in the thousands [6]. All of the 600 compounds regulated under the Toxics Release Inventory (TRI) of these chemical species and numerous other agricultural and industrial compounds that are regulated under waste disposal and treatment regulations, however, pose similar risks to human health and ecosystems. The Agency for Toxic Substances and Disease Registry (ATSDR) in the USA has ranked 275 priority hazardous substances based on the frequency of occurrence at sites present on the National Priorities List, available toxicity data and the potential for direct or indirect human exposure [6]. The different chemical classes of hazardous substances of concern to human health are shown in Table 4.2

4.2
Treatment Technologies to Remove Environmental Pollutants

Cost-effective treatment of environmental pollutants requires the transformation of hazardous substances into benign forms and the subsequent development of

Table 4.3 Chemical processes that are the largest users of heterogeneous catalysts at present.

Reactions	Catalysts
CO, HC oxidation in car exhaust	Pt, Pd on alumina
NO_x reduction in car exhaust	Rh on alumina, V oxide
Cracking of crude oil	Zeolites
Reforming of crude oil	Co–Mo, Ni–Mo, W–Mo
Hydrocracking	Pt, Pt–Re and other bimetallics on alumina
Alkylation	Metals on zeolites or alumina
Steam reforming	Sulfuric acid, solid acids
Water-gas shift reaction	N on support, Fe–Cr, CuO, ZnO, aluminate
Methanation	Ni on support
Ammonia synthesis	Fe
Ethylene oxidation	Ag on support
Nitric acid from ammonia	Pt, Rh, Pd
Sulfuric acid	V oxide
Acrylonitrile from propylene	Bi, Mo oxides
Vinyl chloride from ethylene	Cu chloride
Hydrogenation of oils	Ni
Polyethylene	Cr, Cr oxide on silica

effective risk management strategies for the harmful effects of pollutants that are highly toxic, persistent and difficult to treat. Several new methodologies have been utilized to address new waste treatment approaches that are more effective in reducing contaminant levels that are commercially viable compared with the currently available techniques. Application of nanotechnology that results in improved waste treatment options might include removal of the finest contaminants from water (<300 nm) and air (<50 nm) and "smart materials" or "reactive surface coatings" with engineered specificity to a certain pollutant that destroy, transform or immobilize toxic compounds. Nanomaterials have been attracting increasing interest in the area of environmental remediation, mainly due to their enhanced surface and also other specific changes in their physical, chemical and biological properties that develop due to size effects. The development of novel materials with increased affinity, capacity and selectivity for heavy metals, which are a major pollutant source, has been actively studied because conventional technologies are often inadequate to reduce concentrations in wastewater to acceptable regulatory standards. Commercially available ion-exchange sorbents such as Duolite GT-73, Amberlite IRC-718, Dowex SBR-1 and Amberlite IRA 900X are limited in their ability to remove heavy metal contaminants and are often inadequate for most applications. Genetic and protein engineering have emerged as the latest tools for the construction of nanoscale materials that can be controlled precisely at the molecular level. With the advent of recombinant DNA techniques, it is now possible to create "artificial" protein polymers with fundamentally new molecular organization capabilities that are allowing targeted removal of toxic waste [11].

One of the major environmental pollution sources is automobile exhaust, consisting of harmful emission gases including NO_x, carbon monoxide and unburned

hydrocarbons (HCs), causing smog and acid rain. Most biological reactions that build the human body are catalytic, but application of catalysis in the manufacturing sector in our industrialized world started in the early 1800s and began to be used extensively (listed in Table 4.3) following the discovery of the platinum surface-catalyzed reaction of H_2 and O_2 in 1835 [12].

Since 1975, automobile manufacturers have taken a variety of steps to reduce the level of emission of these harmful emission gases which can be reduced by catalytic reactions in the catalytic converter via the following chemical reactions:

$$CO + O_2 \rightarrow CO_2 \quad \text{CO oxidation}$$

$$HC + O_2 \rightarrow CO_2 + H_2O \quad \text{HC oxidation}$$

$$NO_x + HC \rightarrow N_2 + H_2O + CO_2 \quad NO_x \text{ reduction by HC}$$

$$NO_x + CO \rightarrow N_2 + CO_2NO_x \quad \text{Reduction by CO}$$

The harmful pollutants are converted into relatively benign molecules such as CO_2, N_2 and H_2O through reactions that occur inside the automobile catalytic converters in the presence of catalysts, which consist of mixtures of platinum-group metals such as rhodium (Rh), platinum (Pt) and palladium (Pd). The future targets for the reductions of emission gases from automobile exhaust are very demanding and the requirement on NO_x has been proposed to be 0.05 g per mile, which is about one-quarter of the values that can be achieved through current catalytic converter technology. Transition metal carbides and oxycarbides are being considered as a replacement for the expensive Pt-group metals (Ru, Rh, Ir, Pd and Pt), since recent results that show strong similarities in the catalytic properties between transition metal carbides and the less abundant Pt-group metals. In addition to offering a very high surface-to-volume ratio, nanoparticles offer the flexibility of tailoring the structure and catalytic properties on the nanometer scale.

Nanocrystalline materials composed of crystallites in the 1–10 nm size range possess very high surface-to-volume ratios because of the fine grain size. These materials are characterized by a very high number of low coordination number atoms at edge and corner sites, which can provide a large number of catalytically active sites. For example, gold catalyst systems, consisting of gold nanoparticles on oxide supports, can be used for a wide variety of reactions [13, 14] and many of these have potential for applications in pollution control. Supported gold catalysts are active for the oxidation of methane and propane and the removal of NO_x has also been demonstrated. In exploratory work, gold on a transition metal oxide catalyst system has shown potential as a low-temperature three-way catalyst for automobile emission control [15, 16] with the "light-off" temperatures lowered for both hydrocarbons and carbon monoxide when fresh catalyst is used. A further automotive use for gold catalysts could be in the decomposition of ozone [17]. Consequently, the number of patents related to gold catalysis has shown an upward trend, with close to one-third of such granted patents involving pollution control (Table 4.4) for the period 1991–2001 [18].

Table 4.4 Comparative (%) number of patents granted in the field of catalytic gold nanoparticles[a].

Subject area	% (1991–2001)
Chemical processing	46
Pollution control	29
Catalyst manufacture	15
Fuel cells	10

[a]Adapted from Ref. [18].

In the area of pollution control, some patents [19, 20] have been filed claiming the use of gold catalysts in automotive emissions. Some promise for applications in motor vehicle emission devices most likely in the exhaust treatment of gasoline and diesel cars running at lower temperature ranges and for low light off applications, such as cold start conditions in gasoline engines, has been demonstrated [16]. The use of gold on a clay mineral containing magnesium silicate hydrate has been patented by Toyota for use with ozone to destroy odors [21].

Another promising technology, utilizing the enhanced surface properties of inorganic nanoparticles, involves photocatalytic degradation of organic pollutants in water. Figure 4.1 shows a schematic of this technique. It is based on irradiation of a semiconductor surface with light having energy greater than the semiconductor bandgap exciting electrons from their valence band to the conduction band. This generates electron–hole pairs on its surface. These electron–hole pairs are consumed by either recombination or surface trapping. However, the presence of organic molecules on the surface of the semiconductor material results in a catalytic redox reaction through interfacial charge transfer [22]. Semiconductor materials of choice are II–VI materials such as TiO_2 and ZnO. Noble metals such as gold and platinum on II–VI semiconductor nanoparticles act as a sink for photogenerated charge carriers and promote an interfacial charge-transfer process that leads to an increase in photocatalytic efficiency of metal oxide semiconductors. Under UV illumination, electrons accumulate on the metal surface (making electron–hole pair separation

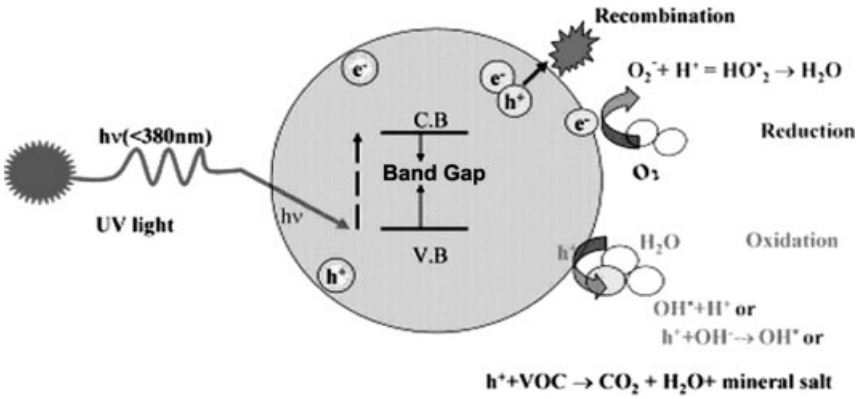


Figure 4.1 Schematic diagram of the photocatalysis on semiconductor surface.

possible, decreasing the recombination of surface charges) and the hole oxidizes absorbed species [23]. This technique has been demonstrated in degrading 4-chlorophenol [24] and chloroform [25] as models of harmful organic chemicals.

4.3 Remediation Technologies to Clean Up Environmental Pollutants Effectively

The likely first impact of nanotechnology research will be in remediation (clean-up of pollution) and end-of-pipe treatment technologies that include nanoscale materials. Substances of significant concern, both cancer causing and otherwise toxic, include heavy metals (e.g. mercury, lead) and organic compounds (e.g. benzene, chlorinated solvents, creosote, toluene). Substances such as DDT and chlordane that are no longer produced or used commercially, but persist in the environment, are also targeted for treatment.

Increased public concern for environmental clean-up has promoted the development of highly efficient photocatalysts that can participate in detoxification reactions. Environmental remediation by photocatalysts comes with several advantages: direct conversion of pollutants to non-toxic by-products without the necessity for any other associated disposal steps; use of oxygen as oxidant and elimination of expensive oxidizing chemicals; potential for using free and abundant solar energy; self-regeneration and recycling of the photocatalyst, etc. It is therefore no surprise that the research and development activities in this field have been very vigorous in recent years [26]. A significant amount of research on semiconductor-catalyzed photooxidation of organic chemicals has been carried out during the past 15 years [27]. The ability to catalyze the destruction of a wide variety of organic chemicals and complete oxidation of organics to CO_2 and dilute mineral acids in many cases, lack of inherent toxicity and resistance to photodegradation at low cost render this process highly suitable for environmental remediation [28].

Two major disadvantages of semiconductor-based photocatalysis techniques, however, are the lack of adequate fixed-bed reactor designs and the large bandgap of metal oxides (e.g. TiO_2 ~3.2 eV). For bulk TiO_2, this wavelength is in the near-UV region, ~390 nm, which means that only a tiny fraction (~3%) of the solar spectrum can be harvested. This high photon energy requirement can be overcome by doping the semiconductor material to produce tail-states in the energy bands, leading to visible light absorption and photocatalytic degradation of any adsorbed organic species on the nanoparticle [22]. Figure 4.2 shows the photocatalytic degradation of methylene blue, used as a test organic compound, with visible light by using doped ZnO as the oxide semiconductor. Using this approach, it becomes possible to utilize effectively 46% of the solar energy in the form of visible light, making photocatalytic degradation techniques more efficient than the conventional UV photon catalysis techniques.

Contamination of sediments and aqueous water systems by halogenated organic compounds presents a serious environmental threat due to their toxicity and resistance to biodegradation. These chemicals are widely employed as pesticides,

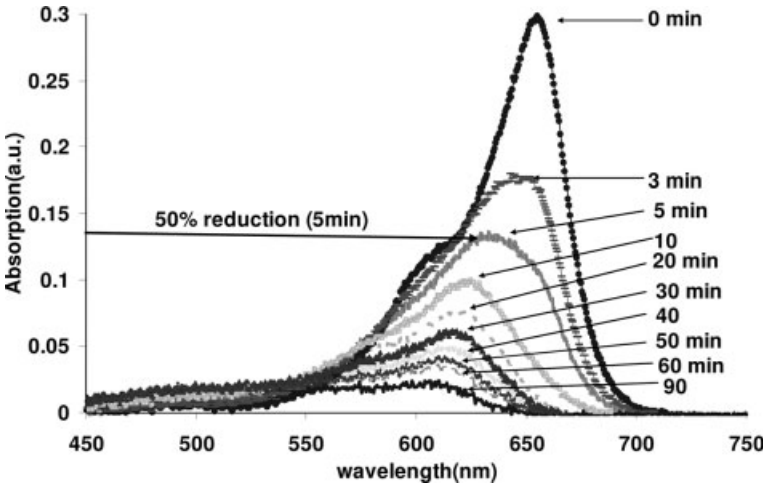


Figure 4.2 Visible light photocatalytic degradation of methylene blue with manganese-doped zinc oxide nanoparticles under 60 lux illumination (tungsten–halogen lamp).

insecticides and wood preservatives and are ubiquitous in the environment of both industrialized and agrarian nations. A subgroup of these chemicals, referred to as chlorinated aromatics, includes chlorinated benzenes, polychlorinated biphenyls (PCBs), pentachlorophenol (PCP) and insecticides such as DDT. Microbial degradation and naturally occurring hydrolysis of these compounds are very slow processes (e.g. for 4-chlorophenol at 9 °C the half–life is nearly 500 days). Some direct photodegradation also occurs, although the limited optical absorbance of chlorinated aromatics at wavelengths above 350 nm slows the process drastically. Sometimes this direct photolysis can actually lead to more toxic products; e.g. direct photolysis of PCP has been reported to lead to octachlorodibenzo-*p*-dioxin, an even more toxic species than its precursor [29]. It is clear that more effective methods of treatment of these chlorinated aromatics must be sought [30, 31]. To this end, a few groups have been investigating the photocatalytic oxidation of these compounds to form harmless CO_2 and HCl, a process referred to as total mineralization [32]. Photocatalysis is needed to reduce toxic pollutants in the atmosphere and water [33], VOCs in the atmosphere and also for reduction of NO_x [34] molecules (largely from vehicle exhausts) into N_2, N_2O, NO_2 and O_2 over semiconductor and zeolite catalysts at ambient temperature. Photocatalytic reaction between ammonia and nitric oxide has been investigated on a TiO_2 wafer under near-UV illumination [35]. Using a novel, integrated, nanobiotechnological approach comprising catalytic dechlorination using FeS nanoparticles followed by microbial degradation, it has been reported that complete removal of 5 mg L^{-1} of lindane (γ-hexachlorocyclohexane), which is an organochlorine pesticide and a persistent organic pollutant (POP), occurs from an aqueous solution in less than 10 h [36].

A variety of photocatalyst nanoparticles have been synthesized, such as oxides (TiO_2, ZnO, Fe_2O_3, WO_3, SnO_2, Ag_2O, V_2O_5, $SrTiO_3$), sulfides (ZnS, CdS, MoS_2,

Cu_xS, Ag_2S, PbS), selenides (CdSe, PbSe, HgSe), iodides (AgI) and modified systems such as coupled semiconductor systems (CdS/TiO_2, $CdSe/TiO_2$, SnO_2/TiO_2, ZnO/TiO_2, ZnO/CdS). Among them, TiO_2 nanoparticles and modified TiO_2 nanoparticles are the most extensively studied and considered to be the most efficient photocatalysts [37]. Other semiconductor nanoparticles generally have lower photocatalytic activity than TiO_2 and some have problems associated with stability, reactivity, etc. [26]. Fe_2O_3 easily undergoes photocathodic corrosion [38] and its active form α-Fe_2O_3 also has high selectivity for the reactant [39].

Many chlorinated aromatics and aliphatics are toxic, even at low concentrations, and exert a cumulative, deleterious effect on river basins and other streams that enter the environment from manufacturing operations and user applications. Reductive dechlorination of organics by various bulk metals (particularly Fe) in the aqueous phase has been well documented. Although nanoparticles possess several advantages (e.g. high surface area and surface energy), sustainability requires particle immobilization on a base membrane to avoid particle loss and agglomeration. Nanostructured metals immobilized in membrane phase leads to high reaction rates at room temperature, significant reduction of metal usage, minimizing the need for the recovery of non-chlorinated products (e.g. ethylene from TCE), leading to a subsequent improvement in water quality. Chlorinated organics and many pesticides and herbicides are toxic to aquatic life, even at low concentrations, and exert a cumulative, effect on receiving streams. The use of non-toxic, polypeptide-based membrane assemblies to create nano-sized metal domains has significant environmental importance [40]. Nanoscale bimetallic (Fe/Pd, 99.9% Fe) particles are considered as a new generation of remediation technology that could provide cost-effective remedial solutions to some of the most difficult waste dumping sites [41]. The complete reduction of aqueous perchlorate to chloride by nanoscale iron particles over a wide concentration range (1–200 mg L^{-1}) has been observed. The reaction is temperature sensitive, as evidenced by progressively increasing rate constant values of 0.013, 0.10 and 1.52 mg perchlorate per gram of iron per hour at temperatures of 25, 40 and 75 °C, respectively. The high activation energy of 79.02 ± 7.75 kJ mol^{-1} partially explains the stability of perchlorate in water. Iron nanoparticles may represent a feasible remediation alternative for perchlorate-contaminated groundwaters.

4.4 Sensors

The categorization of environmental sensors is based primarily on the physics involved and their operating mechanisms. For example, chromatography relies on the separation of complex mixtures by percolation through a selectively adsorbing medium, with subsequent detection of compounds of interest. Electrochemical sensors include sensors that detect signal changes (e.g. resistance) caused by an electric current being passed through electrodes that interact with chemicals. Mass sensors rely on disturbances and changes to the mass of the surface of the sensor

during interaction with chemicals. Optical sensors detect changes in visible light or other electromagnetic waves during interactions with chemicals. Within each of these categories, some sensors may exhibit characteristics that overlap with those of other categories. For example, some mass sensors may rely on electrical excitation or optical settings. Nevertheless, these four broad categories of sensors are sufficiently distinct for the purposes of this chapter. The following sections provide a summary and assessment of the sensors reviewed in each of the four categories.

4.4.1 Biosensors

Biosensors have been reported to detect compounds in several classes of concern, including phenols/phenoxy acids (e.g. phenol and catechol [42]), polyaromatic compounds (e.g. benzo[*a*]pyrene [43]), halogenated pesticides (e.g. triazines [44]), VOCs (e.g. benzene [45]) and inorganic substances (e.g. mercury [46]).

Biosensors typically consist of an enzyme (e.g. acetylcholinesterase, which binds with high affinity to organophosphate insecticides, and Japanese pine-comb fish luciferase, which has been used to measure Hg^{2+}), a receptor [e.g. γ-aminobutyric acid receptor (GABA), which binds to cyclodienes, pyrethroids, bicyclophosphates and orthocarboxylates; the muscarinic receptor, which binds organophosphate insecticides; and the aryl hydrocarbon receptor, which binds with high affinity to dioxins], antibody {e.g. antibodies bind to dozens of insecticides and herbicides and also environmental pollutants such as polychlorinated biphenyls (PCBs), dioxins, pentachlorophenol, benzo[*a*]pyrene and benzene, toluene and xylene (BTX)} or a microbe (e.g. cyanobacteria have been used to measure herbicides, *Trichosporon* cells have been used to measure biochemical oxygen demand and a genetically altered *Pseudomonas* has been used to measure naphthalene), which form the biological sensing element that is in intimate contact with a chemical or physical transducer (electrochemical, optical, mass or thermal). Selectivity and high binding affinities of biological macromolecules towards some environmental pollutants render them useful candidates as sensing elements for environmental biosensors [47]. For environmental monitoring, there are several general areas in which biosensors may have distinct advantages over current analytical methods. Recently, sensors using DNAzymes (segments of DNA with enzymatic activities) that can bind selectively to analytes of interest have been coupled with fluorophores or gold nanoparticles, to form sensitive and selective fluorescent or colorimetric sensors for a variety of analytes [48].

4.4.2 Electrochemical Sensors

Electrochemical sensors represent a key area where the use of nanotechnology (e.g. nanopowders), innovative materials and nano- and micro-fabrication techniques can give sensor products that offer significant enhancements with respect to sensitivity, selectivity, power consumption and reproducibility. The idea of using semiconductors as gas-sensitive devices dates back to the early 1950s when Brattain first reported

gas-sensitive effects on germanium [49]. The gas detection technique is primarily based on a change in the electrical resistance of the semiconducting metal oxide films [50]. The principal detection process is the change of the oxygen concentration at the surface of these metal oxides, caused by the adsorption and heterogeneous catalytic reaction of oxidizing and reducing gaseous species. The electrical conductivity depends on the gas atmosphere and on the temperature of the sensing material exposed to the test gas [51]. Crystallite size effects in resistive sensors are one of the most important factors affecting the sensing properties. The first evidence was obtained in 1982 by Ogawa *et al*. [52] from the measurements of the Hall parameters on SnO_2 nanoparticles. In that report, it was demonstrated that, when the grain size approaches about double the Debye length, the space-charge region can develop in the whole crystallite, which results in a higher sensitivity to adsorbed gas species. For example, the sensitivity of sensors based on tin oxide nanoparticles dramatically increases when the particle size is reduced to 6 nm [2]. Further, the surface reactivity of particles is known to increase rapidly with increase in the surface-to-bulk ratio because the strong curvature of the particle surface generates a higher density of defects which are the most reactive surface sites [12].

Selectivity has always been a major hurdle for solid-state gas sensors. A number of approaches have been developed to improve the selectivity of gas sensors, including doping metal impurities [53], using impedance measurement [54], modulating operating temperature [55] and surface coating [56], and there are some reports on the use of metal oxide-based gas sensors for air pollution monitoring [57].

4.4.3 Mass Sensors

Other strategies for sensing include nanomechanical sensors [3]. Cantilever sensors have also been used for detecting chemicals, such as volatile compounds [58], warfare pathogens [59], explosives [60] and glucose [61] and ionic species, such as calcium ions [62]. The key to using microcantilevers for the selective detection of molecules is the ability to functionalize one surface of the silicon microcantilever in such a way that a given target molecule will be preferentially bound to that surface upon its exposure. The bending and the changes in resonant frequency can be monitored by several techniques, with optical beam deflection, piezoresistivity, piezoelectricity, interferometry, capacitance and electron tunneling among the most important [63]. This strategy allows microcantilever sensors to measure extremely small changes due to molecular adsorption and, for that reason, they are extremely sensitive biosensors; with the cantilever technique, it is possible to detect surface stress as small as about $10^{-4}\,N\,m^{-1}$. Such measurement is also quantitative, related to the concentration of the analyte being detected [64]. Nonetheless, the factors and the phenomena responsible for the surface stress response during molecular recognition remain unclear. Electrostatic interaction between neighboring adsorbates, changes in surface hydrophobicity and conformational changes of the adsorbed molecules can all induce stresses, which may compete with each other and mean that the change in stress is not directly related to the receptor–ligand binding energy.

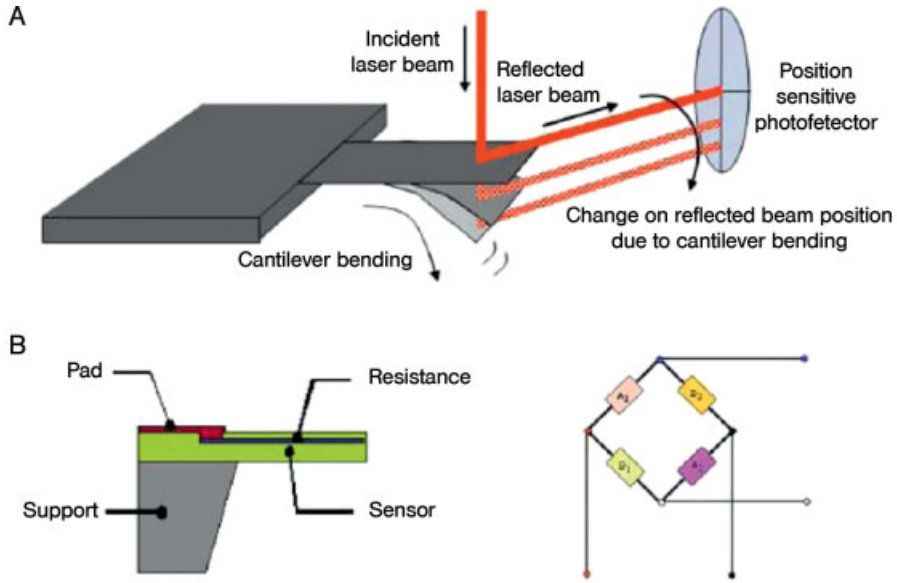


Figure 4.3 Mass sensors using microcantilevers: working mechanism. From Ref. [3].

Silicon, silicon nitride and silicon oxide cantilevers are available commercially with different shapes and sizes, analogous to AFM cantilevers, with typical lengths of 10–500 μm and ultra-thin cantilevers up to 12 nm thick. However, for specific applications (e.g. highly sensitive biosensors), cantilevers must be designed and fabricated to satisfy their requirements. Cantilever sensitivity depends critically on the spring constant: the lower the constant, the higher is the sensitivity for measurements in liquids based on the static method. Figure 4.3 shows a schematic of the working mechanism of a microcantilever sensor.

For accurate functioning of such sensors based on microcantilevers, the immobilization process should:

- avoid any change in the mechanical properties of the cantilever;
- be uniform, in order to generate a surface stress as large as possible; and
- allow accessibility by the target molecule.

Depending on the surface coating of the cantilevers, selectivity for various chemical and biological species can be achieved. Commercially available polymers have been used to coat cantilevers for differentiating between different VOCs in air. Baller *et al*. developed a Nanotechnology Olfactory SEnsor (NOSE) to characterize and to identify gaseous analytes [58]. In the field of explosives, Pinnaduwage and co-workers reported measurement of trinitrotoluene (TNT) in a small, localized explosion on an uncoated piezoresistive microcantilever [60, 65]. Heavy metal ions and ions in general have also been studied. Ji and co-workers used thiol-derivatized calixarene and crown ether macrocycle-functionalized cantilevers to detect Cs^{2+} ions in the range $10^{-11} - 10^{-7}$ M and K^{+} in the 10^{-4} M range [66]. Other functionalization

schemes have shown that cantilevers were able to detect, with great accuracy and selectivity, different ions, such as CrO_4^{2-} [67], Ca^{2+} [62] and Pb^{2+} [68].

Biological applications of cantilever based mass sensors include the detection of different pathogens, such as *Salmonella enterica* by Weeks *et al.* [59], *Vaccinia* virus by Gunter *et al.* [69] and fungal spores from *Aspergillus niger* by Nugaeva *et al.* [70]. Monitoring concentrations of specific pesticides plays an essential role in the environmental control field. An example of the application of a cantilever-based biosensor in this area was reported by Alvarez *et al.* [71] for the detection of the organochlorine insecticide dichlorodiphenyltrichloroethane (DDT). A synthetic hapten of the pesticide, conjugated with bovine serum albumin (BSA), was covalently immobilized on the gold-coated side of the cantilever; specific detection was then achieved by exposing the cantilever to a solution containing the specific monoclonal antibody to the DDT–hapten derivative. The specific binding of the antibodies on the sensitized side of the cantilever was measured with nanomolar sensitivity. Finally, a competitive assay was performed, with the cantilever exposed to a mixed solution of the monoclonal antibody and DDT and direct detection was achieved. With this detection strategy, DDT concentrations as low as 10 nM were detected, involving deflection signals in the 50 nm range. Many other applications have been described for the detection of pesticides and avidin– streptavidin [72].

4.4.4
Optical Sensors

For realizing sensitive chemical sensors and biosensors, optical methods employing optical fibers or integrated optics (IO) and, in the case of remote sensing, by connecting fiber pigtails, offer high sensitivity and fast responses. Contemporary methods for optical sensing of chemical and biological species at present are based mainly on interferometry, surface plasmon resonance (SPR) and luminescence. The relatively recent technique of luminescence quenching [73] is a new alternative. The photoluminescent (PL) properties of nanocrystalline (porous) silicon depend on the chemical nature of its surface; for example, metal ions can quench PL from porous Si [74], as can inorganic molecules [75–77]. The nanoscale size permits *in vivo* monitoring of processes within individual cells. Concentrations of toxic chemicals within carcinoma cells [78] have already been achieved. PEBBLE (probe encapsulated by biologically localized embedding) nanosensors have been prepared for the analytes oxygen [79, 80], potassium [81], zinc [82] and magnesium [83]. The wide variation of the optical properties of gold nanoparticles with particle size, interparticle distance and the dielectric properties of the surrounding media due to SPR [84–86] permits the construction of simple but sensitive colorimetric sensors for various analytes. Highly sensitive colorimetric sensors for biomolecules [87–89] and metal ions [4, 5], amongst others, have been devised using SPR of gold nanoparticles. In the example of sensing heavy metal ions, well-known metal ion chelators, such as chitosan, can be coated on the nanoparticle surfaces, such that the ligand changes its dielectric properties upon chelating the metal ions, resulting in an optical (colorimetric) signal (Figure 4.4).

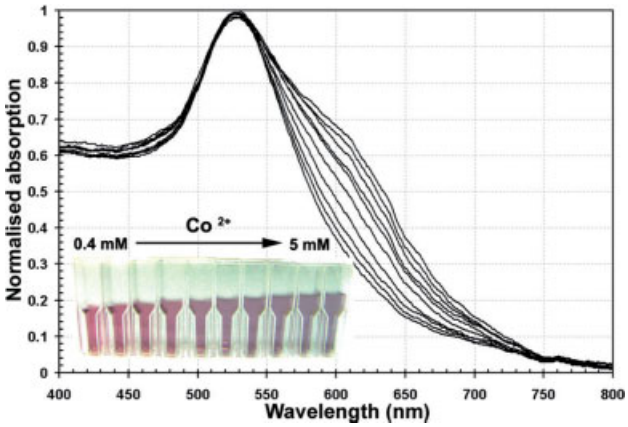


Figure 4.4 Colorimetric sensing of cobalt ions using gold nanoparticles.

4.4.5 Gas Sensors

Gas sensors for detecting air pollutants must be able to operate stably under deleterious conditions, including chemical and/or thermal attack. Therefore, solid-state gas sensors appear to be the most appropriate in terms of their practical robustness. The sensors used for detecting air pollutants are usually produced simply by coating a sensing (metal oxide) layer on a substrate with two electrodes. Typical materials are tin (IV) oxide (SnO_2), zinc oxide (ZnO), titanium dioxide (TiO_2) and tungsten oxide (WO_3), with typical operating temperatures of 200–400 °C [90]. When the active surface of a metal oxide (e.g. zinc oxide) grain is exposed to the ambient oxygen, the oxygen atoms are adsorbed on the surface as shown in Figure 4.5 and the adsorbed oxygen acts as an

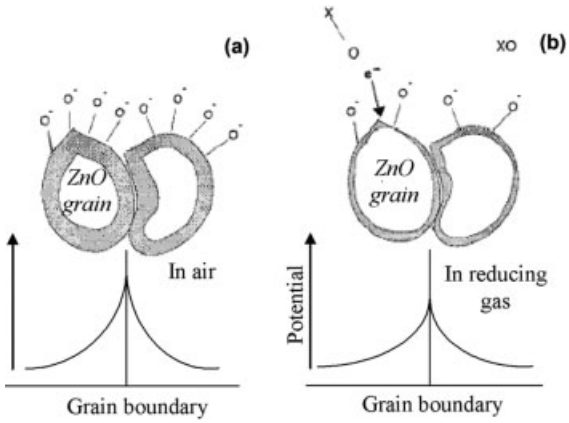


Figure 4.5 Surface effect on gas sensing including potential barrier at the grain boundary of ZnO nanoparticles: (a) in air and (b) in a reducing gas.

acceptor state. Being ionized, a depletion layer is formed on the surface of the grains and also in the neck regions, which raises the height of the potential barrier [91]. In the presence of a reducing gas, the adsorbed oxygen can easily react with the gas molecules and exit from the lattice, thus reducing the concentration of acceptors, which in turn lower the voltage barrier height. When the metal oxide sensor absorbs a reducing gas (CO, H_2), the depletion area at the surface will be decreased, leading to increased conductivity. On the other hand, if a metal oxide sensor absorbs an oxidizing gas (NO_2), the depletion zone at the surface will be increased, meaning decreasing conductivity. A change in conductivity/resistance is related to gas concentration. In the case of a ZnO sensor, conductivity decreases, which means resistance increases when the sensor absorbs NO_x, depending on the NO_x concentration [90].

A host of sensing materials are available to be used as the sensing layer for solid-state gas sensors so far reported, as shown in Table 4.5. Despite their simplicity and

Table 4.5 Materials for Solid-state sensors.

Sensing materials	Target gases
SnO_2	AsH_3, H_2S, NO_2, H_2, NH_3, SO_2, NO_x, CO, CH_3COOH, CCl_4, $C_2H_2O_4$, CH_4, CO_2
SnO_2, Fe_2O_3, La_2CuO_3, Ga_2O_3, ZnO, In_2O_3, V_2O_5, $Bi_2Sn_2O_7$, GaAs, Bi_2Sr_2	CO, CH_4, NO_x, H_2, C_2H_2, C_6H_6
SnO_2	NO_2, CO, NO_x, CH_4, O_2, ethanol, phenylarsine, C_6H_6, diethyl ether, H_2S, H_2, ammonia, *iso*-C_4H_{10}
SnO_2	H_2, NH_3, SO_2, CH_3COOH, C_2H_5OH, CH_4, AsH_3, H_2S, NO_2, NO_x, CCl_4, CO_2
ZnO, WO_3, TiO_2, Fe_2O_3, $CdIn_2O_4$, $NiTa_2O_6$, $CoTa_2O_6$, $CuTa_2O_6$, $BaTiO_3$(Ag), $SrFeO_3$, Cr_2O_3, In_2O_3, $BaSnO_3$, $Bi_2Sn_2O_7$, $Bi_6Fe_4Nb_6O_6$	H_2, CCl_2F_2, $CHClF_2$, CO, NH_3, H_2, CH_4, C_4H_{10}, N_2H_4, H_2S, SO_2, $(CH_3)_3N$, C_3H_8, C_2H_5OH, C_3H_8, Cl_2, NO_2
SnO_2, ZnO, Cu_2O	O_2
SnO_2	H_2, H_2S, NO, NO_x, C_2H_5OH, C_3H_6, diethyl ether, H_2NH_3, NO_2, C_3H_8, hydrocarbons, H_2Cl_2, Cl_2, $CHCl_3$, CCl_4, AsH_3, C_4H_{10}, *n*-C_4H_{10}
$YBa_2Lu_3O_7$-d, In_2O_3, $LnCoO_{3-x}$, Sm, Eu, La, $Bi_xMo_yO_z$, $Cr_{0.8}Ti_{0.2}O_3$, ZnO, ITO, $CdIn_2O_4$, In_2O_3, $Sn_{1-x}Fe_xO_y$	NO, O_3, CO, alcohols, ketones, NH_3, H_2, CH_4, C_4H_{10}, C_2H_5OH, NO_2, ethanol
SnO_2(Pd), Ag/SnO_2, SnO_2(Ag), SnO_2(ZrO$_2$), SnO_2(Al_2O_3), SnO_2(CuO), SnO_2(Pt), WO_3(Au), $Ti_{0.9}Cr_{0.1}O_2$, $IrTiO_2/In_2O_3$(MgO), Pt/In_2O_3(MgO), Ru/TiO_2, In/TiO_2, ZnO(Al_2O_3), SnO_2(La_2O_3), Er_2O_3/ZnO, Rh/WO_3	H_2S, CH_3SH, NH_3, dimethylamine, trimethylamine, capronaldehyde, 2-methylpyrazine

low production cost, solid-state gas sensors (SGS) usually exhibit drifts and variations in behavior. After the introduction of nanoparticles, sensitivity in gas sensors has improved. The use of nanoscale materials exposes a higher surface area of the sensing element to gas and hence the physicochemical reaction that proceeds at the surface is increased [92].

4.4.6 Novel Sensing Technologies and Devices for Pollutant and Microbial Detection

Protection of human health and ecosystems requires rapid, precise sensors capable of detecting pollutants up to the molecular level. Examples of research into sensors include the development of nanosensors for efficient and rapid *in situ* biochemical detection of pollutants and specific pathogens in the environment; sensors capable of continuous measurement over large areas, including those connected to nanochips for real-time continuous monitoring; and sensors that utilize lab-on-a-chip technology. Research may also involve sensors that can be used in monitoring or process control to detect or minimize pollutants and their impact on the environment.

4.4.6.1 Real-Time Chemical Composition Measurements of Fine and Ultrafine Airborne Particles

New technologies need to be developed that will permit the chemical composition of airborne nanoparticles (down to about 5 nm in diameter) to be determined. Knowledge of the chemical composition will provide a better understanding of the sources of these particles and how to control their formation in a manner that reduces their impact on human health. It is now well established that long-term exposure to fine particulate matter is a significant risk factor for cardiopulmonary and lung cancer mortality in humans [93]. In urban air, fine particulates typically exhibit a maximum in both number and mass in the 100–300 nm diameter range. Most of these particles are produced directly from combustion sources. However, a significant fraction of particles in this size range may also arise from growth and/or coagulation of much smaller particles. The mechanism of particle formation is difficult to assess without chemical composition measurements during these events. Technology already exists for chemical analysis of fine particles [94].

4.4.6.2 Ultrasensitive Detection of Pathogens in Water

The primary water quality problem in the developing world is waterborne diseases [95]. Table 4.6 lists some common waterborne pollutants. Conventional tests for the problem count indicator bacteria – *E. coli* (coliform counts) – and require a laboratory facility and trained personnel, with the only alternative to laboratory methods being the simple Colilert test kit [96]. The need for rapid, simple tests for fecal contamination of water is not confined to developing countries. In coastal areas of the USA, for example, monitoring beaches for the indicator species *E. coli* and *Enterococcus* by laboratory analysis can take more than 1 day. During that time, conditions can change and swimmers can be put at risk.

Table 4.6 Selected waterborne contaminants in developing countries [95].

Problem	Occurrence
Pathogens	Most significant risk to water quality
Metals, e.g. arsenic	Associated with certain geological conditions or with agricultural or industrial use
Pesticides	Localized areas of agricultural use; runoff; aerial transport
Algal toxins	An array of neuro-, hepato- and cytotoxins are produced by a range of cyanobacteria; typically found in water with elevated nutrient levels
Nitrates	Widespread; natural and agricultural sources
Fluoride	Localized areas, depending on geology
Organic compounds	Common sources are industry and transport; includes aromatic and aliphatic hydrocarbons, halogenated compounds and persistent organic pollutants

Current developments in pathogen detection in water rely on filtration culture methods and fluorescence-based methods (e.g. fluorescence probe methods and DNA microarray methods). These techniques, however, are not effective for *in situ*, rapid, quantitative measurements. With filtration culture methods, sample water is passed through a filter that is pretreated for visualization of the target pathogen. Growth of colonies on the filter indicates the presence of the target pathogen in the test water. Both the fluorescently labeled probe methods and the DNA microarray methods rely on detection using fluorescence spectroscopy, which is not quantitative. There is a need for rapid, quantitative and specific pathogen detection to ensure the safety of natural and man-made water supplies, including source, treated, distributed and recreational waters. Arrays of a highly piezoelectric (e.g. strontium-doped lead titanate) microcantilever smaller than 10 μm in length coupled to antibody proteins immobilized at the cantilever tip for *in situ* rapid, simultaneous multiple pathogen quantification in source water have been used [97].

4.4.6.3 Detection of Heavy Metals in Water

Arsenic is a well-known toxic chemical that the EPA and the World Health Organization (WHO) [98] list as a known carcinogen. Arsenic is found in a wide variety of chemical forms throughout the environment and can be readily transformed by microbes, changes in geochemical conditions and other environmental processes [99]. Although arsenic occurs naturally, it may also be found as a result of a variety of industrial applications [100], including leather and wood treatments [101] and pesticides [102]. Anthropogenic arsenic contamination results mainly from manufacturing metals and alloys, refining petroleum and burning fossil fuels and wastes. These industrial activities have created a strong legacy of arsenic pollution throughout the world. Unlike organic pollutants, arsenic cannot be transformed into a non-toxic material, but can only be transformed into a form that is less toxic when

exposed to living organisms in the environment. Because arsenic is a permanent part of the environment, there is a long-term need for regular monitoring at sites where arsenic-containing waste has been disposed of and at sites where it occurs naturally at elevated levels.

Fixed laboratory assays are generally required to measure arsenic accurately in an environmental sample to parts per billion (ppb) levels, defined as $\mu g\,L^{-1}$ for water or $\mu g\,kg^{-1}$ for solids. The preferred laboratory methods for the measurement of arsenic involve pretreatment, either with acidic extraction or acidic oxidation digestion of the environmental sample. Pretreatment transfers all of the arsenic in the sample into an arsenic acid solution, which is subsequently measured using any one of several accepted analytical methods, such as atomic fluorescence spectrometry (AFS) [103], graphite furnace atomic absorption spectrometry (GFAAS), hydride generation atomic absorption spectrometry (HGAAS), inductively coupled plasma atomic emission spectrometry (ICP-AES) and inductively coupled plasma mass spectrometry (ICP-MS) [104]. The instruments involved are bulky, expensive to operate and require fully equipped laboratories to maintain and operate. Field assays, in which lower sensitivities may be acceptable for purposes of sample screening or site surveys, strive for similar detection goals as fixed laboratory methods are relatively inexpensive and can produce a large number of screening results in a short time.

A considerable amount of research has been dedicated to developing an arsenic detection colorimetric solution that matches or exceeds the sensitivity of the Gutzeit method while improving safety, accuracy and reproducibility. One research group electrochemically reduced the arsenite ion to arsine gas. They found that the arsenic reduction by this electrochemical method compared favorably with reduction by sodium borohydride. They were able to achieve detection limits down to 50 ppb arsenite using this method. Gold, copper and iron(III) species were found to interfere with the sample reduction [105]. One such system reduces arsenic compounds to arsine gas, which then bleaches a dye in a solution containing detergents and metal particles. This system has been shown to be effective, with limits of detection for arsenic as low as 30 ppb [106].

Accurate, fast measurement of arsenic in the field still remains a technical challenge. Selective solid-state sensors for carcinogenic and toxic chromium(VI) and arsenic(V) in water based on redox quenching of the luminescence from nanostructured porous silicon and polysiloles has been undertaken [107]. Sensors based on silicon wafer and polymer technologies are readily adaptable to fabrication. The fluorescence quenching detection modality also is manufacturable. The essential electronics require a blue or UV LED as the excitation source and an inexpensive photodiode detector. Potential applications of such real-time solid-state sensors include remote sensing and industrial process control. The focus on chromium(VI) and arsenic(V) detection is dictated by the redox quenching mechanism that is being used, and also by the importance of chromium(VI) and arsenic(V) as regulated chemicals under the Safe Drinking Water Act. Chromium(VI) detectors need to be developed that can sense the analyte at concentrations at least as low as the 0.1 ppm action levels with at least 10% accuracy. For arsenic(V), the target range is 10–50 ppb at the same level of analytical accuracy.

4.5 Conclusions

Environmental protection and pollution issues are frequently discussed worldwide as topics that need to be addressed sooner rather than later. Nanotechnology can strive to provide and fundamentally restructure the technologies currently used in environmental detection, sensing, remediation and pollution removal. Some nanotechnology applications are near commercialization: nanosensors and nanoscale coatings to replace thicker, more wasteful polymer coatings that prevent corrosion, nanosensors for detection of aquatic toxins, nanoscale biopolymers for improved decontamination and recycling of heavy metals, nanostructured metals that break down hazardous organics at room temperature, smart particles for environmental monitoring and purification, nanoparticles as novel photocatalysts for environmental catalysts, etc. New advances have emerged in the use of nanotechnology in environmental protection, and these been discussed in this chapter. Strategies involved in detecting markers/tracers of "substances of interest" are a very important part of preventing, remediation or sensing pollution for prevention of major catastrophes. For example, quantum dots functionalized with an appropriate molecule could be tuned to detect hazardous materials selectively by simply looking for fluorescence signals from these detector dots. A host of applications, similar or very different from this, are actively being pursued worldwide, as discussed in this chapter. This review provides just a small beginning to the exciting new applications envisaged for the "small world" of nanotechnology for preventing environmental pollution.

Acknowledgments

We are grateful for support from the Swedish International Development Agency (SIDA) and the National Nanotechnology Center (NSTDA) of the Thai Ministry of Science and Technology (MOST).

References

1 US Environmental Protection Agency , *Report EPA/600/R-92/219*, EPA, Washington, DC, 1992.

2 M.-I. Baraton, L. Merhari, *J. Nanopart. Res.* 2004, **6**, 107.

3 L. G. Carrascosa, M. Moreno, M. Alvarez, L. M. Lechuga, *Trends Anal. Chem.* 2006, **25**, 196.

4 S. O. Obare, R. E. Hollowell, C. J. Murphy, *Langmuir* 2002, **18**, 10407.

5 A. Sugunan, C. Thanachayanont, J. Dutta, J. G. Hilborn, *Sci. Technol. Adv. Mater.* 2005, **6**, 335.

6 K. R. Rogers, L. R. Williams, *Trends Anal. Chem.* 1995, **14**, 289.

7 B. Manning, T. Maley, *Biosens. Bioelectron.* 1992, **7**, 391.

8 P. Schneider, G. Lorinci, I. L. Gebefugi, J. Heinrich, A. Kettrup, H. E. Wichmann, *J. Expos. Anal. Environ. Epidemiol.* 1999, **9**, 282.

9 L. Stevens, J. A. Lanning, L. G. Anderson, W. A. Jacoby, N. Chornet, *J. Air Waste Manage. Assoc.* 1998, **48**, 979.

10 D. T. Thompson, *Nanotoday* 2007, **2**, 40.

11 G. Prabhukumar, M. Matsumoto, A. Mulchandani, W. Chen, *Environ. Sci. Technol.* 2004, **38**, 3148.

12 G. A. Somorjai, *Introduction to Surface Chemistry and Catalysis*, Wiley, New York, 1994.

13 D. T. Thompson, *Chem. Br.* 2001, **37**, 43.

14 G. C. Bond, D. T. Thompson, *Catal. Rev. Sci. Eng.* 1999, **41**, 319.

15 D. T. Thompson, C. W. Corti, R. J. Holliday, presented at the ATT Congress, Paris, 2002, Paper 2002-01-2148, 2002.

16 J. R. Mellor, A. N. Palazov, B. S. Grigorova, J. F. Greyling, K. Reddy, M. P. Letsoala, J. H. Marsh, *Catal. Today* 2002, **72**, 145.

17 Z. Hao, D. Cheng, Y. Guo, Y. Liang, *Appl. Catal. B: Environ.* 2001, **33**, 217.

18 C. W. Corti, R. J. Holliday, D. T. Thompson, *Gold Bull.* 2002, **35**, 111.

19 L. A. Petrov, Bulgaria, *World Patent* WO 9 851 401, 1998.

20 P. Marecot, R. Emmanuel, *French Patent* 2 771 310, 1999.

21 Toyota Chuo Kenkyushu KK , *Japanese Patent* 9 150 033, 1997.

22 S. I. Shah, W. Li, C.-P. Huang, O. Jung, C. Ni, *Proc. Natl. Acad. Sci. USA* 2002, **99**, 6482.

23 P. V. Kamat, *J. Appl. Chem.* 2002, **74**, 1693.

24 S. T. Martin, C. L. Morrison, M. R. Hoffmann, *J. Phys. Chem.* 1994, **98**, 13695.

25 W. Chio, A. Termin, M. R. Hoffmann, *J. Phys. Chem.* 1994, **98**, 13669.

26 D. Beydoun, R. Amal, G. Low, S. McEvoy, *J. Nanopart. Res.* 1999, **1**, 439.

27 M. Schiavello, *Photoelectrochemistry, Photocatalysis and Photoreactors: Fundamentals and Developments*, Reidel, Dordrecht, 1985.

28 E. Pelizzetti, N. Serpone, *Homogeneous and Heterogeneous Photocatalysis*, Reidel, Dordrecht, 1986.

29 G. Mills, M. R. Hoffmann, *Environ. Sci. Technol.* 1993, **27**, 1681.

30 M. R. Hoffmann, S. T. Martin, W. Choi, D. W. Bahnemann, *Chem. Rev.* 1995, **95**, 69.

31 N. Serpone, R. Terzian, D. Lawless, P. Kennipohl, G. Sauve, *J. Photochem. Photobiol. A: Chem.* 1993, **73**, 11.

32 M. A. Fox, M. T. Dulay, *Chem. Rev.* 1993, **93**, 341.

33 K. Tanaka, T. Hisanaga, A. P. Rivera, in D. F. Ollis, H. Al-Ekabi (eds.), *Photocatalytic Purification and Treatment of Water and Air*, Elsevier, Amsterdam, 1993, 169.

34 L. A. Dobble, G. B. Raupp, *Catal. Lett.* 1990, **4**, 345.

35 K. T. Ranjit, B. Viswanathan, *J. Photochem. Photobiol. A: Chem.* 1997, **108**, 73.

36 K. M. Paknikar, V. Nagpal, A. V. Pethkar, J. M. Rajwade, *Sci. Technol. Adv. Mater.* 2005, **6**, 370.

37 M. A. Fox, M. T. Dulay, *Chem. Rev.* 1993, **93**, 54; A. L. Linsebigler, G. Lu, J. T. Yates, *Chem. Rev.* 1995, **95**, 735; R. F. Howe, *Dev. Chem. Eng. Miner. Process.* 1998, **6**, 55.

38 M. R. Hoffmann, S. T. Martin, W. Choi, D. W. Bahnemann, *Chem. Rev.* 1995, **95**, 69.

39 C. Kormann, D. W. Bahnemann, M. R. Hoffmann, *J. Photochem. Photobiol. A: Chem.* 1989, **48**, 161.

40 K. Venkatachalam, V. G. Gavalas, S. Xu, A. C. Leon, D. Bhattacharyya, L. G. Bachas, *J. Nanosci. Nanotechnol.* 2006, **6**, 2408.

41 W. Zhang, C. Wang, *Environ. Sci. Technol.* 1997, **31**, 2154.

42 F. Munteanu, A. Lindgren, J. Emneus, L. Gorton, T. Ruzgas, A. Ciucu, E. Csörregi, *Anal. Chem.* 1998, **70**, 2596.

43 J. P. Alarie, D. J. Bowyer, M. J. Sepaniak, A. M. Hoyt, T. Vo-Dinh, *Anal. Chim. Acta* 1990, **236**, 237.

44 P. Orozlan, G. L. Duveneck, M. Ehrat, H. M. Widmer, *Sens. Actuators B* 1993, **11**, 301.

45 Y. Ikariyama, S. Nishiguchi, E. Kobatake, M. Aizawa, M. Tsuda, T. Nakazawa, *Sens. Actuators B* 1993, **13**, 169.

46 S. Pirvutoiu, I. Surugiu, E. S. Dey, A. Ciucu, V. Magearu, B. Danielsson, *Analyst* 2001, **126**, 1612.

47 A. Lindgren, L. Stoica, T. Ruzgas, A. Ciucu, L. O. Gorton, *Analyst* 1999, **124**, 527; K. R. Rogers, J. N. Lin, *Biosens. Bioelectron.* 1992, **7**, 317; C. Nistor, J. Emneus, L. Gorton, A. Ciucu, *Anal. Chim. Acta* 1999, **387**, 309; K. Riedel, in G. Ramsay (ed.), *Commercial Biosensors: Applications to Clinical, Bioprocess and Environmental Samples* Wiley New York 1998; E. Dominguez, in O'Sullivan, G. G. Guilbault, S. Alcock, A. P. F. Turner (eds.), *Biosensors for Environmental Monitoring: Technology Evaluation*, University College Cork, Cork, 1998.

48 J. Liu, Yi Lu, *Adv. Mater.* 2006, **18**, 1667.

49 W. H. Brattain, J. Bardeen, *Bell Syst. Tech. J.* 1953, **32** (1), 1.

50 S. Jonda, M. Fleischer, H. Meixner, *Sens. Actuators B* 1996, **34**, 396.

51 G. Eranna, B. C. Joshi, D. P. Runthala, R. P. Gupta, *Crit. Rev. Solid State Mater. Sci.* 2004, **29**, 111.

52 H. Ogawa, M. Nishikawa, A. Abe, *J. Appl. Phys.* 1982, **53**, 4448.

53 H. Nanto, T. Minami, S. Takata, *J. Appl. Phys.* 1986, **60**, 482.

54 G. Faglia, P. Nelli, G. Sberveglieri, *Sens. Actuators B* 1994, **19**, 497.

55 A. Heilig, N. Barsan, U. Weimar, M.S-. Berberich, J. W. Gardner, W. Gopel, *Sens. Actuators B* 1997, **43**, 45.

56 Q. F. Pengfei, O. Vermesh, M. Grecu, *et al. Nano Lett.* 2003, **3**, 347.

57 O. Pummakarnchanaa, N. Tripathia, J. Dutta, *Sci. Technol. Adv. Mater.* 2005, **6**, 251.

58 M. K. Baller, H. P. Lang, J. Fritz, C. Gerber, J. K. Gimzewski, U. Drechsler, H. Rothuizen, M. Despont, P. Vettiger, F. M. Battiston, J. P. Ramseyer, P. Fornaro, E. Meyer, H. J. Guntherodt, *Ultramicroscopy* 2000, **82**, 1.

59 B. L. Weeks, J. Camarero, A. Noy, A. E. Miller, L. Stanker, J. J. De Yoreo, *Scanning* 2003, **25**, 297.

60 L. A. Pinnaduwage, A. Gehl, D. L. Hedden, G. Muralidharan, T. Thundat, R. T. Lareau, T. Sulchek, L. Manning, B. Rogers, M. Jones, J. D. Adams, *Nature* 2003, **425**, 474.

61 A. Subramanian, P. I. Oden, S. J. Kennel, K. B. Jacobson, R. J. Warmack, T. Thundat, M. J. Doktycz, *Appl. Phys. Lett.* 2002, **81**, 385.

62 H.-F. Ji, T. Thundat, *Biosens. Bioelectron.* 2002, **17**, 337.

63 N. V. Lavrik, M. J. Sepaniak, P. G. Datskos, *Rev. Sci. Instrum.* 2004, **75**, 2229.

64 J. Fritz, M. K. Baller, H. P. Lang, H. Rothuizen, P. Vettiger, E. Meyer, H.-J. Guntherodt, C. Gerber, J. K. Gimzewski, *Science* 2000, **288**, 316.

65 L. A. Pinnaduwage, A. Wig, D. L. Hedden, A. Gehl, D. Yi, T. Thundat, R. T. Lareau, *J. Appl. Phys.* 2004, **95**, 5871.

66 H.-F. Ji, R. Dabestani, G. M. Brown, P. F. Britt, *Chem. Commun.* 2000, 457.

67 H.-F. Ji, T. Thundat, R. Dabestani, G. M. Brown, P. F. Britt, P. V. Bonnesen, *Anal. Chem.* 2001, **73**, 1572.

68 K. Liu, H.-F. Ji, *Anal. Sci.* 2004, **20**, 9.

69 R. L. Gunter, W. G. Delinger, K. Manygoats, A. Kooser, T. L. Porter, *Sens. Actuators A* 2003, **107**, 219.

70 N. Nugaeva, K. Gfeller, N. Backmann, H. P. Lang, H.-J. Güntherodt, M. Hegner, *Microsc. Microanal.* 2007, **13**, 13.

71 M. Alvarez, A. Calle, J. Tamayo, L. M. Lechuga, A. Abad, A. Montoya, *Biosens. Bioelectron.* 2003, **18**, 649.

72 R. Raiteri, M. Grattarola, H.-J. Butt, P. Skla'dal, *Sens. Actuators B* 2001, **79**, 115.

73 P. V. Lambeck, *Sens. Actuators B* 1992, **8**, 103.

74 D. Andsager, J. Hilliard, M. H. Nayfeh, *Appl. Phys. Lett.* 1994, **64**, 1141.

75 M. T. Kelly, A. B. Bocarsly, *Coord. Chem. Rev.* 1998, **171**, 251.

76 J. Harper, M. J. Sailor, *Anal. Chem.* 1996, **68**, 3713.

77 G. Di Francia, V. La Ferrara, L. Quercia, G. Faglia, *J. Porous Mater.* 2000, **7**, 287.

78 B. Cullum, G. Griffin, G. Miller, T. Vo-Dinh, *Anal. Biochem.* 2000, **277**, 25.

79 H. Xu, J. W. Aylott, R. Kopelman, T. J. Miller, M. A. Philbert, *Anal. Chem.* 2001, **73**, 4124.

80 Y.-E. L. Koo, Y. Cao, R. Kopelman, S. M. Koo, M. Brasuel, M. A. Philbert, *Anal. Chem.* 2004, **76**, 2498.
81 M. Brasuel, R. Kopelman, T. J. Miller, R. Tjalkens, M. A. Philbert, *Anal. Chem.* 2001, **73**, 2221.
82 J. P. Sumner, J. W. Aylott, E. Monson, R. Kopelman, *Analyst* 2002, **127**, 11.
83 E. J. Park, M. Brasuel, C. Behrend, M. A. Philbert, R. Kopelman, *Anal. Chem.* 2003, **75**, 3784.
84 L. M. Liz-Marzan, *Mater. Today* 2004, **7**, 26–31.
85 G. L. Hornyak, C. J. Patrissi, C. R. Martin, J.-C. Valmalette, L. Lemaire, J. Dutta, H. Hofmann, *Nanostruct. Mater.* 1997, **9**, 571.
86 S. Link, M. A. El-Sayed, *Annu. Rev. Phys. Chem.* 2003, **54**, 331.
87 R. Jelinek, S. Kolusheva, *Biotechnol. Adv.* 2001, **19**, 109.
88 N. T. Kim Thanh, Z. Rosenzweig, *Anal. Chem.* 2002, **74**, 1624.
89 N. Nath, A. Chilkoti, *Anal. Chem.* 2002, **74**, 504.
90 I. Simon, N. Bârsan, M. Bauer, U. Weimar, *Sens. Actuators B* 2001, **73**, 1.
91 P. Hauptmann, *Sensors – Principles and Application*, Prentice-Hall, Englewood Cliffs, NJ, 1991.
92 Y. Shimizu, M. Egashira, *Bull. Mater. Res.* 1999, **24** (6), 18.
93 C. A. Pope, R. T. Burnett, M. J. Thun, E. E. Calle, D. Krewski, K. Ito, G. D. Thurston, *J. Am. Med. Assoc.* 2002; **287**, 1132.
94 M. V. Johnston, *J. Mass Spectrom.* 2000; **35**, 585.
95 Australian Agency for International Development , *Safe Water Guide for the Australian Aid Program*, Canberra, 2005, www.ausaid.gov.au/publications/pdf/safe_water_guide.pdf.
96 See http://www.idexx.com/water/colilert, IDEXX Laboratories, USA.
97 J. W. Yi, W. Y. Shih, R. Mutharasan, W.-H. Shih, *J. Appl. Phys.* 2003, **93**, 619.
98 World Health Organization , *Arsenic in Drinking Water*, Fact Sheet No. 210, 2001, http://www.who.int/inf-fs/en/fact210.html.
99 W. R. Cullen, K. J. Reimer, *Chem. Rev.* 1989, **89**, 713.
100 US Environmental Protection Agency , *Locating and Estimating Air Emissions from Sources of Arsenic and Arsenic Compounds*, EPA-454-R-98-013, Office of Air Quality Planning and Standards, Research Triangle Park, NC-27711, 1998.
101 J. A. Collins, *et al. Environ. Pollut.* 2001, **111**, 53.
102 K. Kristen, *Environ. Sci. Technol.* 2000, **34**, 376A.
103 J. L. G. Ariza, D. S. Rodas, I. Giraldez, E. Morales, *Talanta* 2000, **51**, 257.
104 US Environmental Protection Agency , Analytical Methods Support Document for Arsenic in Drinking Water, EPA-815-R-00-010, EPA, Office of Water, Targeting and Analysis Branch, Washington, DC, 1999.
105 M. H. A. Zavar, M. Hashemi, *Talanta* 2000, **52**, 1007.
106 S. Kundu, S. K. Ghosh, M. Mandal, T. Pal, A. Pal, *Talanta* 2002, **58**, 935; S. Kundu, S. K. Ghosh, M. Mandal, T. Pal, *New J. Chem.* 2002, **26**, 1081.
107 W. C. Trogler, S. Toal, in S. R. Pehrsson, P. Pehrsson (eds.), *Nanotechnology and the Environment*, Oxford University Press, Oxford, 2005, 169.

5
Benefits in Energy Budget

Ian Ivar Suni

5.1
Introduction

Nanomaterials in the 1–100 nm size range have unusual potential for applications within a wide variety of existing and emerging technologies. Nanomaterials have several intriguing properties that may be exploited for technological applications. Due to quantum confinement effects, when their dimensions are comparable to the electron mean free path or the optical wavelength, the electronic and optical properties of nanomaterials become size dependent. This is of course the origin of the unique properties of the widely popularized "quantum dots", which exhibit quantum confinement in all three dimensions. Another interesting property of nanomaterials, and in particular nanoparticles, is their unusually high chemical reactivity. This has led to the widespread use of metal and metal oxide nanoparticles as commercial catalysts in the chemical and petrochemical industries. Metal nanoparticles are also currently employed within catalytic converters in automobiles as three-way catalysts. Three-way catalysts catalyze the following three reactions: oxidation of unburned hydrocarbons, oxidation of CO, and reduction of nitrogen oxides.

Another interesting aspect of nanomaterials is their unusually high surface area per unit mass. Many potential applications that exploit the high surface area of nanomaterials involve their use within compacted solids as what are termed nanostructured materials, which in many cases are composite materials. Nanostructured materials thus have extremely high internal surface areas, although these may not be chemically accessible. Composite nanomaterials can be fabricated from nanowires or nanotubes of extremely high aspect ratio, allowing for low percolation thresholds. This means that high aspect ratio nanomaterials can more easily form interacting networks within a composite material to form a conductive electrical pathway or to increase mechanical strength.

Although nanomaterials have several existing applications, their potential for the development of new technologies is the main source of the excitement within academia, government and industry. Among the most widely anticipated applications of nanomaterials is the development of more environmentally friendly and more efficient energy sources. Interest in sustainable energy is driven in part by long-term concerns

Nanotechnology. Volume 2: Environmental Aspects. Edited by Harald Krug

ISBN: 978-3-527-31735-6

about the scarcity of hydrocarbon fuels, which are in increasingly great demand with the rapid industrialization of China, Russia, Brazil and other emerging economies. In addition, concerns about greenhouse gas emissions such as CO_2 that arise from combustion of hydrocarbons are generating interest in cleaner energy sources.

Applications of nanomaterials in the field of energy include fuel cell catalysts, fuel cell support materials, hydrogen storage, solar cells, lithium ion batteries and supercapacitors. The current discussion will focus on recent results, on clear demonstrations of the utility of nanomaterials and on the scientific basis for these applications. In low-temperature fuel cells, Pt and other noble metal nanoparticle catalysts have been widely studied for their ability to catalyze efficiently the electrochemical reduction of oxygen and the electrochemical oxidation of both hydrogen and methanol. Because these nanoparticle catalysts may be interspersed with less conductive materials, carbon nanotubes have been widely investigated as catalyst support materials to improve catalyst utilization in fuel cells. The development of fuel cells and other energy sources powered by molecular hydrogen, with water as the only chemical product, is an important goal for sustainable energy. One of the critical issues limiting hydrogen energy is the need for an infrastructure and new technology for hydrogen storage and distribution. Although early results have now been shown to be misleading, carbon nanotubes have been widely investigated for their hydrogen storage properties.

Further applications for nanomaterials can be envisioned in the area of solar energy cells. The classical example of nanotechnology is the variation in optical absorption/emission of semiconductor nanostructures with dimension. These size-dependent properties have been exploited to alter the wavelength of optical absorption to match the terrestrial window. In addition, nanostructured TiO_2 in dye-sensitized solar cells (DSSCs) has been widely investigated due to its high internal surface area, which increases the available dye for optical absorption and maximizes internal reflections within the DSSC.

Nanomaterials have also been employed within Li ion batteries, particularly as materials for anode construction. Many materials have been demonstrated to have higher Li capacities than the prototypical graphite anode material, but they have been prone to mechanical failure due to repeated expansion (contraction) during Li insertion (removal) as the battery is charged (discharged). Intensive research efforts have been expended to use these alternative Li anode materials in the form of nanoparticles, nanowires or nanotubes to minimize mechanical strain during Li insertion and removal.

5.2
Nanomaterials in Fuel Cells

5.2.1
Low-Temperature Fuel Cell Technology

The development of fuel cells as a clean, environmentally friendly energy source is widely anticipated. The use of a hydrogen as a fuel is particularly attractive, since the

main product produced would be H_2O, with effectively zero emissions. Even the economical use of other hydrocarbon fuels beyond gasoline and natural gas may have global benefits, as this may lessen the demands for hydrocarbon fuels. Fuel cells operate by converting chemical potential energy directly into a current or voltage by coupling an electrochemical oxidation reaction with an electrochemical reduction reaction. A wide variety of fuel cells have been investigated, including proton exchange membrane, direct methanol, molten carbonate, solid oxide and phosphoric acid fuel cells.

High-temperature fuel cells such as molten carbonate, solid oxide and phosphoric acid fuel cells have recently been employed for several applications, particularly those where waste heat can be employed to reach and maintain the operating temperature. For example, waste heat is widely generated throughout industrial chemical plants, sometimes making fuel cells an economical energy source. At the operating temperature of these fuel cells, the anode and cathode reactions are typically fairly facile, making the use of electrocatalysts, which are often in the form of nanoparticles, unnecessary. In addition, nanomaterials may be subject to grain growth, sintering, dissolution and other unwanted chemical reactions at high temperature.

On the other hand, nanomaterials are much more compatible with low-temperature fuel cells, which are needed for many transportation and consumer applications where intermittent operation is typical and power requirements are relatively modest. The most common low-temperature fuel cells are the polymer electrolyte membrane fuel cell (PEMFC) and the direct methanol fuel cell (DMFC), where the following reactions occur:

$$\text{Anode (PEMFC):} \quad H_2 \rightarrow 2H^+ + 2e^- \tag{5.1}$$

$$\text{Anode (DMFC):} \quad CH_3OH + H_2O \rightarrow CO_2 + 6H^+ + 6e^- \tag{5.2}$$

$$\text{Cathode (PEMFC and DMFC):} \quad O_2 + 4H^+ + 4e^- \rightarrow 2H_2O \tag{5.3}$$

The structure of a typical proton exchange membrane fuel cell (PEMFC) is illustrated in Figure 5.1. The cathode in a DMFC has a similar structure, whereas the anode structure depends on whether the methanol feed stream is in liquid or vapor form. In both electrodes of a PEMFC, the metal nanoparticle catalyst is dispersed atop larger carbon particles that are combined into a porous structure that allows mass transport of both reactants and products. Sandwiched between the two electrodes is a Nafion-type polymer material that serves as a proton conductor between the anode, where protons are produced and the cathode, where protons are consumed. Nafion, a perfluorosulfonated polymer, facilitates proton "hopping" along its sulfonate backbone through a series of electrostatic interactions.

Several technological challenges remain for commercialization of PEMFCs and DMFCs. For economic reasons, the use of precious metal catalysts should be reduced or eliminated. The cost of Nafion polymer membranes may also need to be reduced and their durability improved. Nafion membranes are employed for proton transport in both PEMFCs and DMFCs and are only conductive within a narrow temperature

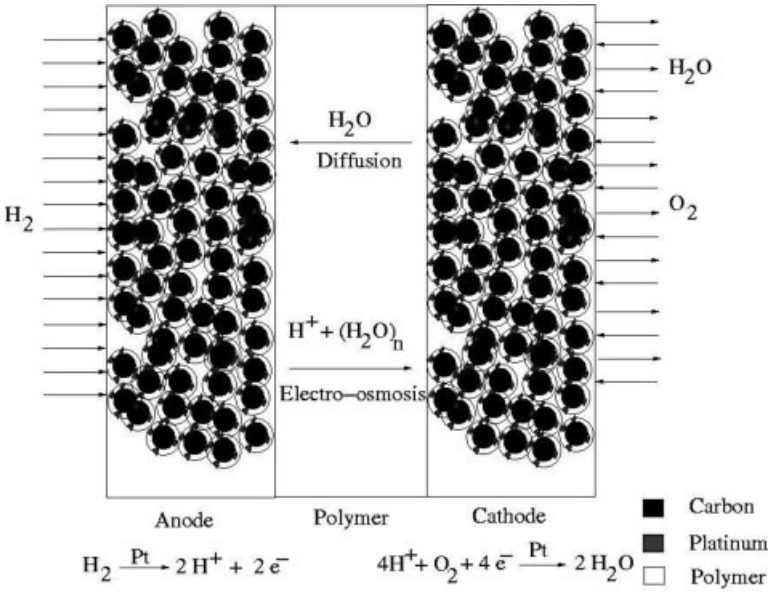


Figure 5.1 Schematic illustration of a typical proton exchange membrane fuel cell (PEMFC). (Courtesy of Renganathan Rengaswamy, Department of Chemical and Biomolecular Engineering, Clarkson University, Potsdam, NY, USA).

range, where the membrane is neither dried out nor flooded. In an H_2-fueled PEMFC, the need to maintain an appropriate humidification level throughout the fuel cell creates a complex water management problem. The local humidification depends on a complex balance between water production at the cathode, water consumption at the anode, water diffusion from the cathode to the anode, and water electro-osmosis from the anode to the cathode [1]. The durability of PEMFCs and DMFCs must also be improved, since Nafion degradation, catalyst agglomeration and dissolution and carbon corrosion can all occur upon prolonged operation at high current density.

5.2.2
Nanoparticle Catalysts in Low-Temperature Fuel Cells

When considering the use of nanomaterials in fuel cells, many observers would first consider the use of nanoparticle catalysts in both the anode and cathode. However, nanoparticle fuel cell catalysts will be discussed only briefly, since these reactions have been widely studied and the interested reader can consult recent reviews [2–9]. Hydrogen reduction by Reaction (5.1) at the anode of a PEMFC is the most facile due to its simple reaction mechanism, and Pt nanoparticles are widely used as electrocatalysts for this reaction. The main complication is that Pt catalysts can be easily poisoned by trace CO in the H_2 fuel, and so far the best performance has been attained by PtRu bimetallic nanoparticle catalysts, preferably with a 1 : 1 ratio of

Pt : Ru, that facilitate CO desorption. Ternary and quaternary catalysts have also been widely investigated.

The two cathode reactions above, methanol reduction by Reaction (5.2) and O_2 reduction by Reaction (5.3), involve more complex mechanisms and multi-step electron transfer, making electrocatalysis more difficult. O_2 reduction is most facile on Pt nanoparticle catalysts, and the use of Pt alloys with transition metals such as Co, Cr, Ti and Zr has been thoroughly investigated. However, for catalysts tested to date, the overpotential loss of 300–400 mV for O_2 reduction still accounts for about 80% of the voltage loss in a typical PEMFC [10]. Similarly, methanol oxidation has been widely studied on Pt nanoparticle catalysts alloyed with a wide variety of different transition metals, including Ru, Os and Sn. Given that the expensive Pt catalyst contributes significantly to the overall fuel cell cost, non-Pt catalyst materials are also under intensive investigation for both PEMFCs and DMFCs [11, 12]. However, Pt and its alloys in nanoparticle form remain the best catalysts for reactions (5.1)–(5.3) in low-temperature fuel cells.

5.2.3
Fuel Cell Catalyst Support Materials

Both PEMFCs and DMFCs employ porous catalyst support structures, typically some form of carbon, that perform multiple functions. The catalyst support material must be porous enough to provide a pathway for inlet and outlet of gaseous reactants and products, but it must also maintain electrical conductivity so that the voltage (current) created across the fuel cell can be captured for use or storage. The requirement to maintain electrical conductivity is complicated by the presence of Nafion polymer, which is typically far less conductive than the carbon support material.

Carbon nanotubes and carbon nanofibers have been widely investigated for possible application into the catalyst supports shown in Figure 5.1 for the PEMFC [10]. The main improvement that is envisioned is increased utilization for the Pt catalyst supported on carbon nanotubes. The high nanotube aspect ratio increases the likelihood that Pt catalyst will have direct electrical contact to the desired electrode, without electrical blockage by intervening Nafion particles. Another potential advantage of carbon nanotubes as catalyst supports is their improved resistance to oxidation. One of the primary barriers to commercialization of both PEMFCs and DMFCs is their poor durability. During long-term usage, catalyst agglomeration, catalyst dissolution and carbon corrosion all occur, resulting in a gradual loss of performance.

Wang *et al.* recently reported that the corrosion current for carbon nanotube catalyst support materials in a PEMFC cathode is 30% lower than that from Vulcan XC-72 carbon catalyst support materials [13]. In addition, these authors noted that the supported Pt catalyst better maintains its activity for the oxygen reduction reaction. Li and Xing recently used cyclic voltammetry to compare corrosion currents for carbon nanotube and Vulcan XC-72 carbon catalyst supports following prolonged oxidation [14]. They found that for the carbon nanotube-based support material, the

corrosion current eventually disappeared, whereas the corrosion current continued indefinitely for the standard carbon support material.

5.2.4 Carbon Nanotubes: Science and Technology

Carbon nanotubes, which are allotropes of carbon from the fullerene structural family, have been the most widely studied nanomaterial. They can be conceived as all-sp^2 carbons arranged in graphene sheets that have been rolled up into hollow tubes. The nanotubes can be capped at the ends by a fullerene-type hemisphere and can range in length from tens of nanometers to several micrometers. Carbon nanotubes can be subdivided into two categories, single-wall carbon nanotubes (SWNTs) and multi-wall carbon nanotubes (MWNTs). As the name suggests, SWNTs consist of a single hollow tube with a diameter of 0.4–3 nm, whereas MWNTs are composed of multiple concentric nanotubes spaced by 0.34 nm, with overall diameters of 2–200 nm.

Research interest in carbon nanotubes arises from several of their extraordinary properties. For example, their mechanical strength per unit weight is 100 times greater than that of steel, their electrical conductivity is similar to that of Cu and their thermal conductivity is comparable to that of diamond [15]. MWNTs have electrical conductivities greater than those of metals, but depending on the tube diameter and chirality, SWNTs can behave electronically as either metals or semiconductors. In addition, the aspect ratios of both SWNTs and MWNTs can be as high as 10^3–10^5, allowing for low percolation thresholds when they are employed in composite materials. Thus carbon nanotubes have been proposed for a diverse range of applications, including nanoscale transistors, chemical sensors, high-strength composites, hydrogen storage, and fuel cell electrode supports.

Carbon nanotubes can be made by laser ablation, electric arc discharge and chemical vapor deposition. The detailed synthesis conditions, such as temperature, pressure, or the presence of an inert gas, strongly influence the properties of the resulting carbon nanotubes, as does the presence and type of metal catalyst employed. One of the primary difficulties with these synthetic methods is that all create a complex mixture of different carbon forms, including amorphous carbon, graphite particles and carbon nanotubes. Thus synthesis must typically be followed by a difficult separation process.

For applications as fuel cell catalyst support materials, one should also consider the chemical reactivity of carbon nanotubes, since they must first be functionalized with metal nanoparticle catalysts and then formed into porous support materials. Both of these processes involve solution-phase chemistry, preferably aqueous chemistry. Dispersion of carbon nanotubes into aqueous solvents is difficult given their hydrophobicity, so the use of organic solvents is often required. In addition, carbon nanotubes are highly chemically inert, so chemical or electrochemical methods must be employed to attach the catalyst, typically Pt or its alloys.

Among the methods that have been employed for catalyst deposition are electroless deposition, otherwise known as chemical impregnation, electrodeposition, microwave

techniques, sputtering, and several different colloidal techniques [10]. Electroless impregnation methods are most commonly employed and involve surface oxidation/activation by a strong acid, Pt salt adsorption and Pt salt reduction by a reducing agent such as formaldehyde. Surface activation is necessary to increase the number of reactive hydroxyl, carbonyl, carboxylate and phenolic sites on the surface of the carbon nanotubes. The use of ultrasound during the surface oxidation/activation step has recently been shown to increase the Pt content by increasing the number of reactive sites generated.

The use of electrodeposition for Pt particle formation on carbon catalyst supports is attractive in that very small nanoparticles (<5 nm diameter) can be formed with very high activity [16, 17]. However, limited diffusion of Pt ions into the channels within carbon nanotube/Nafion composites severely restricts the Pt loading that may be attained. Several groups have electrodeposited Pt nanoparticles on carbon nanotube support materials through an alternative process, whereby application of a cathodic potential is preceded by metal salt precipitation on to the support surface [18–20]. This greatly enhances the catalyst loading that can be attained.

Although functionalization of carbon nanotubes with Pt catalyst nanoparticles is likely needed for creating an efficient PEMFC, carbon nanotubes also have intrinsic electrochemical behavior that may be beneficial for such applications. Electron transfer associated with different electrochemical reactions, including oxygen reduction, has been shown to be intrinsically more rapid at carbon nanotube electrodes than at electrodes made from other forms of carbon [21, 22]. For example, the oxygen reduction peak is shifted quite strongly in the anodic direction for an MWNT electrode prepared by an electric arc discharge process relative to a carbon paste electrode. Alternatively, one can compare the magnitude of the exchange current density for this reaction, which is about 5 times higher on an MWNT electrode than on a graphite electrode [21].

In addition, the electrochemical behavior of carbon nanotube electrodes is in general highly surface dependent and the introduction of certain functional groups, such as those introduced by oxidation treatments, can increase the rate of electron transfer dramatically [23]. For example, very different results can be obtained depending on whether the walls or the ends dominate the electrochemical nature of carbon nanotubes. The walls exhibit electrochemistry similar to that of basal planes of pyrolytic graphite, while the ends exhibit electrochemistry similar to the edges of pyrolytic graphite [23]. In addition, the electrochemical behavior of carbon nanotubes varies considerably with the methods used for purification and preparation [23].

5.2.5
Carbon Nanotubes within Operating PEMFCs

Many research groups have reported electrochemical studies of Pt and Pt alloy nanoparticles dispersed on to carbon nanotubes or graphite nanofibers. Given the size of this literature, discussion here will focus on results reported within an operating DMFC or an operating PEMFC or on those who have made direct comparisons to standard carbon support materials. This ensures that the electrochemical behavior

reported is retained following construction of a membrane electrode assembly (MEA), which involves the application of elevated temperature and pressure. In addition, carbon nanotubes will likely change their relative orientation during MEA fabrication, possibly affecting their behavior as catalyst support materials.

Results have been reported for oxygen reduction at carbon nanotube-supported cathode catalysts in both DMFCs and PEMFCs. Li and co-workers reported detailed studies of DMFCs in which carbon nanotubes are used to support Pt cathode catalysts [24, 25]. This group appears to have published the earliest report of the use of carbon nanotubes for fuel cell catalyst support materials. They compared two DMFCs with Pt nanoparticles as cathode catalysts, in one case supported on MWNTs and in the other case supported on Vulcan XC-72 carbon [24, 25]. Both DMFCs contained about 1.0 mg cm^{-2} Pt at the cathode and both employed commercial anode catalysts containing about 2.0 mg cm^{-2} PtRu. The carbon nanotubes were grown by electric arc discharge in a vacuum chamber, and Pt loading was accomplished by reduction of H_2PtCl_6 with ethylene glycol, yielding Pt particles of 2–5 nm diameter. Reduction with ethylene glycol was reported to yield much higher catalyst dispersion than reduction with formaldehyde.

Their best MWNT-supported cathode catalyst exhibited a limiting current density and maximum power density of 434 mA cm^{-2} and 103 mW cm^{-2}, respectively, while the corresponding Vulcan XC-72-supported catalyst yielded values of 309 mA cm^{-2} and 70 mW cm^{2-}, respectively [25]. Their results are shown in Figure 5.2 for two different preparations (A and B) of MWNT supports. In addition to the enhanced conductivity and connectivity of the carbon nanotubes, the authors also suggested

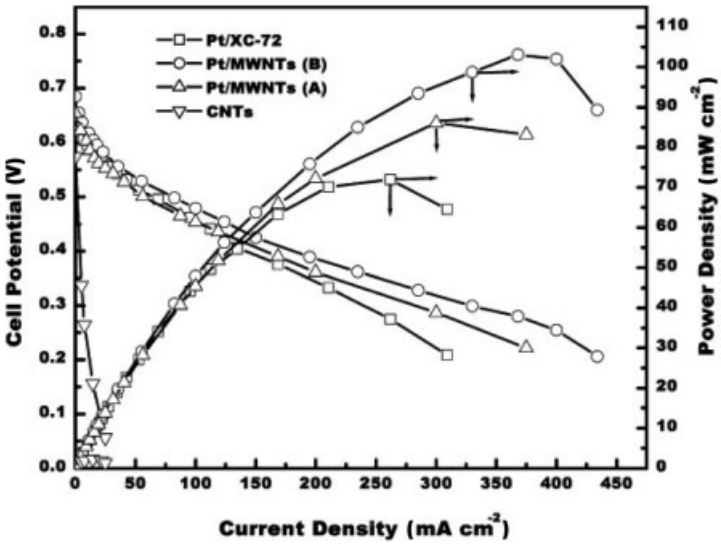


Figure 5.2 Current–voltage response and power density for a DMFC containing four different cathode catalysts supports. (From Ref. [25]).

that the catalyst activity on the standard support material may be compromised by trace amounts of organosulfur compounds, common catalyst poisons [24].

Not surprisingly, several reports have also appeared of carbon nanotube supports for Pt cathode catalysts in H_2-fueled PEMFCs, where the oxygen reduction reaction is the same [26–29]. Shaijumon *et al*. recently reported interesting results for an operating PEMFC with composite carbon catalyst supports containing a mixture of Pt-decorated MWNTs and commercial Pt/C samples from E-Tek [29]. MWNTs were fabricated by catalytic decomposition of acetylene in a CVD reactor and decorated with Pt nanoparticles by impregnation with H_2PtCl_6 followed by reduction with $NaBH_4$. This yields Pt particles of size 5–8 nm. Composite cathodes were fabricated by using 0, 25, 40, 50, 60, 75 and 100 wt.% Pt/MWNT, with the remainder of the catalyst as commercial E-Tek 20 wt.% Pt/C, with a total Pt loading of 0.5 mg cm^{-2}. The anode of the PEMFC was constructed from commercial E-Tek 20 wt.% Pt/C, with a loading of 0.25 mg cm^{-2}. Surprisingly, an optimum performance was observed with a composite catalyst support composed of a 50:50 wt.% mixture of the two carbon forms. This yielded a current density of 535 mA cm^{-2} at a voltage of 540 mV [29]. This is considerably higher than the corresponding current densities for pure E-Tek and pure MWNT catalyst supports of 258 and 362 mA cm^{-2}, respectively [29].

Waje *et al*. recently reported the PEMFC performance of CVD-grown carbon nanotubes that are pretreated by electrochemical reduction in a diazonium–acetonitrile electrolyte and then decorated with Pt nanoparticles by standard impregnation and reduction methods [27]. This treatment yields uniform Pt particles of about 2–2.5 nm diameter, with a mass loading of about 0.09 mg cm^{-2} [27]. This Pt/carbon nanotube catalyst layer was then employed as the cathode in a H_2-fueled PEMFC with a standard E-Tek/Vulcan XC-72 anode catalyst, and its performance was compared with that of a reference E-Tek/Vulcan XC-72 cathode catalyst with a slightly lower loading, about 0.075 mg cm^{-2}. The maximum power density obtained for the Pt/nanotube cathode catalyst was about 290 mW cm^{-2}, whereas that obtained from the reference cathode catalyst was about 160 mW cm^{-2} [27]. The authors contend that the superior performance at high current densities arises from the more open structure of the Pt/nanotube cathode catalyst, which enhances mass transport of reactants and products [30].

Several research groups have investigated the use of carbon nanotube catalyst supports on the anode side of an H_2-fueled PEMFC [31–34]. Li *et al*. recently described a filtration method for incorporating a Pt/carbon nanotube film into a PEMFC so that it is partially oriented [31]. These authors started with commercial MWNTs, oxidized them in a nitric acid–sulfuric acid mixture, and deposited Pt nanoparticles by ethylene glycol reduction of H_2PtCl_6, producing Pt nanoparticles of 2–5 nm diameter. The Pt/MWNT suspension was then filtered through a hydrophilic nylon filter-paper, which apparently forces the hydrophobic MWNTs to stand up and self assemble on the filter-paper [31]. These can then be pressed onto a Nafion membrane to create a partially oriented but somewhat loosely packed Pt/MWNT film. This Pt/MWNT film was then used as the cathode catalyst layer in an operating PEMFC and compared with two reference cathode catalysts, one made from a

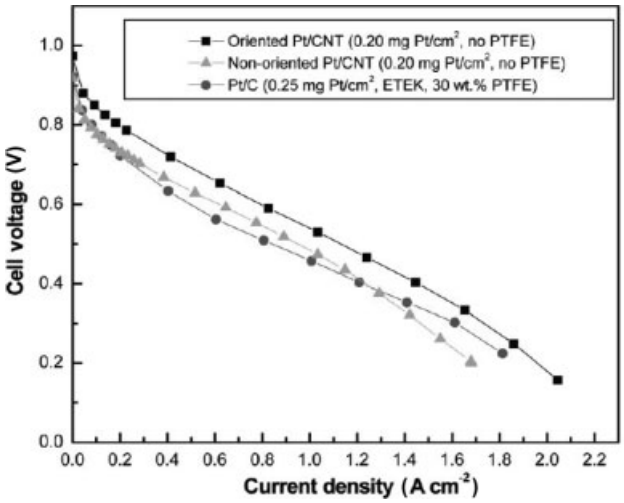


Figure 5.3 Current–voltage response for a PEMFC containing oriented nanotube, non-oriented nanotube and E-Tek cathode catalysts supports. (From Ref. [31]).

non-oriented Pt/MWNT film and one made from E-Tek Pt/C. All cathodes had approximately the same Pt loading, 0.20–0.25 mg cm^{-2}. The best performance was observed for the cathode catalyst containing oriented Pt/MWNT, which the authors argue arises partly from improved mass transport [31]. The results of Li *et al.* are shown in Figure 5.3, which compares oriented carbon nanotube cathode catalyst support materials with non-oriented carbon nanotube supports.

Carmo *et al.* also studied carbon nanotube catalyst supports for PEMFC anode fabrication using nanotubes grown by chemical vapor deposition [32]. Both Pt and PtRu nanoparticle catalysts were deposited by impregnation of H_2PtCl_6, with and without $RuCl_3$, followed by reduction at elevated temperature in an H_2 atmosphere. This produced an average Pt particle size of about 3.6 nm and an average PtRu particle size of about 4.6 nm, whereas the corresponding particle sizes on Vulcan XC-72 carbon were 6.8 and 6.4 nm, respectively. As is typically observed, the particle sizes measured by different techniques varied somewhat. The metal loading in all cases was approximately 0.4 mg cm^{-2}. At the anode side, the Pt/carbon nanotube catalyst showed significantly better performance than the Pt/Vulcan XC-72 catalyst, suggesting that carbon nanotubes may have some intrinsic role in suppressing CO poisoning [32]. However, the authors noted that this may be due to trace presence of the metal catalysts used for carbon nanotube growth. The same group also investigated the same set of catalysts and supports for the anode reaction in a direct methanol fuel cell, finding that the best performance was obtained for a PtRu catalyst supported on MWNTs [32].

Liang *et al.* reported a study of carbon nanotube anode catalyst supports that compared different techniques for PtRu nanoparticle formation [33]. They found that reduction of mixtures of H_2PtCl_6 and $RuCl_3$ in ethylene glycol yielded a much higher nucleation density of nanoparticles than reduction in aqueous solutions of

formaldehyde. This was attributed to the lower polarity of ethylene glycol, which therefore does not interfere with ion exchange reactions that deposit Pt and Ru [33]. The ethylene glycol reduction formed 2–8 nm diameter PtRu catalyst particles on MWNTs that were purchased commercially and purified before use. Anode catalysts with a loading of 0.22–0.32 mg cm^{-2} were then compared with commercial catalysts with a loading of 0.39 mg cm^{-2} in an operating hydrogen fuel cell. On a per weight basis, the MWNT-supported catalyst exhibited superior performance in the middle to high current density regime [33].

Several groups have studied the use of carbon nanotubes as anode catalyst supports in DMFCs [35–39]. Understanding of these studies is complicated by the need to separate effects associated specifically with the catalyst support from those associated with the exact catalyst composition.

5.3 Hydrogen Storage

The efforts to develop energy sources powered by molecular hydrogen, rather than by hydrocarbons, are motivated primarily by the desire to minimize the production of greenhouse gases. Carbon dioxide, which is one of the common greenhouse gases, is the inevitable product of energy sources fueled by hydrocarbons. By comparison, energy sources fueled by hydrogen would produce mainly water, dramatically reducing greenhouse gas emissions. The development of hydrogen-powered energy sources encompasses a number of technical challenges, including the development of low-cost, efficient fuel cells powered by hydrogen, in addition to low-cost, efficient methods for producing and storing hydrogen. One of the greatest challenges for applications in transportation is the lack of an infrastructure to store and distribute hydrogen. The infrastructure for storing and distributing hydrocarbons is well developed and the development of an alternative infrastructure for hydrogen is a daunting economic and technological obstacle. Hydrogen storage for vehicular applications has challenging constraints of weight and space.

Proposed hydrogen storage methods include compression, liquefaction, hydride formation and adsorption on carbon and other nanomaterials, although all currently have significant shortcomings [40]. Although hydrogen compression is the simplest method for hydrogen storage, the energy density is not high enough for most applications that are envisioned. In addition, this approach is thought to be more expensive than hydrogen liquefaction. On the other hand, hydrogen liquefaction is limited by the large energetic requirement for the liquefaction process and by the continuous loss of hydrogen due to boiling.

For systems based on hydrogen adsorption, the simplest way to compare their hydrogen storage capabilities is the weight percent of hydrogen that they are capable of adsorbing. In addition to the storage capacity, another important issue for hydrogen storage is reversibility. Practical hydrogen storage systems need to be reversible at moderate temperatures and pressures. Hence the mechanism of hydrogen storage is extremely important. Hydrogen chemisorption can likely only

be reversed at elevated temperature, whereas hydrogen physisorption is more likely to be readily reversible. The two main obstacles to commercialization are the low hydrogen storage capacity and the reversibility of the hydrogen storage process over a great many hydrogen adsorption/desorption cycles.

Two types of metal hydrides have been studied for hydrogen storage applications, metal hydrides and complex hydrides. The prototypical metal hydrides are formed from binary intermetallic compounds, A_xB_y [40]. Some examples include $LaNi_5$, TiFe, Ti_2Ni and $CeNi_3$. A is typically a rare earth or alkaline earth metal that tends to form a stable hydride, whereas B is typically a transition metal and does not form stable hydrides. The function of metal B, often Ni, is to catalyze hydrogen dissociation. The most studied metal hydrides that form from A_xB_y intermetallic compounds have relatively low hydrogen storage capacity, typically less than 3 wt.%. As a result, recent attention has turned towards lighter metals such as Mg, but to date suitable Mg hydrides have not been found.

More complex metal hydrides can be formed from hydrogen and combinations of metals from Groups I, II and II such as Li, Mg, B and Al [40]. Such materials include Mg_2FeH_6, $Al(BH_4)_3$, $NaAlH_4$ and $LiBH_4$, the last of which has the highest hydrogen storage capacity yet reported, 18 wt.%. Since it is able to store hydrogen reversibly at moderate temperature, $NaAlH_4$ has received greater attention than $LiBH_4$. One difficulty with these materials is that hydrogen desorption occurs via a multi-step mechanism where the optimum conditions for each step are different. In practice, this means either that hydrogen desorption is slow or that the full hydrogen storage capacity of these materials is not utilized. In addition, their stability over many cycle periods is inadequate, with gradual changes in composition and morphology.

5.3.1
Hydrogen Storage Using Carbon Nanomaterials

The primary motivation for the use of nanomaterials for hydrogen storage is their extremely high surface area per unit weight or unit volume. Sound fundamental reasons exist for investigating carbon nanomaterials relative to nanomaterials of other compositions. Carbon is well known for its ability to adsorb gases, so carbon materials are already widely employed as adsorbents. Carbon-based nanomaterials proposed for hydrogen storage include carbon nanotubes and graphite nanofibers. As discussed above, carbon nanotubes can be prepared either as SWNTs or MWNTs. SWNTs, which have the strong adsorption capability of carbon materials coupled with an enormous surface area per unit weight, are therefore promising as hydrogen storage materials.

Despite the widespread use as carbon as an adsorbent, the precise mechanism by which carbon nanomaterials adsorb hydrogen is not completely understood. For gases above the critical temperature, the expected adsorption mechanism is monolayer adsorption. Simple calculations based on the known surface area of different nanomaterials and the known dimensions of hydrogen molecules yield maximum hydrogen storage capacities in the range 2–4 wt.% [41]. More complex calculations are not substantially different. Such values are considerably less than the benchmark values provided by the US Department of Energy (DOE), which projects requirements

of 4.5 wt.% by 2007, 6 wt.% by 2010 and 9 wt.% by 2015 [42]. However, the existence of more complex hydrogen storage mechanisms, such as those involving defects, cannot be discounted.

The highest hydrogen storage capacities reported to date for carbon nanotubes are in the 5–10 wt.% range. However, the upper end of this range is now treated with considerable skepticism, since other investigators have had difficulty reproducing these results [43–45]. Instead, it has become generally accepted that room temperature hydrogen storage capacity is limited to less than 1 wt.%, although up to 6 wt.% hydrogen storage capacity can be attained at cryogenic temperatures [46]. Since hydrogen monolayers are generally bound by physisorption, hydrogen storage capacity typically decreases dramatically as the temperature is raised.

It should be noted that measurements of hydrogen adsorption are complicated by the very small extent of hydrogen adsorption relative to other gases, so experimental measurements are sensitive to the detailed procedures of how they are performed [45, 47, 48]. Agreement among reports of hydrogen storage capacity from different laboratories is often hampered by the lack of reliable methods for producing, purifying and quantifying carbon nanotubes. Indeed, if defects are critically involved in hydrogen storage, slight differences in preparation techniques may even result in intrinsic differences in hydrogen storage capacity [49].

One focus of current research efforts is on methods for improving the hydrogen storage capacity of carbon nanomaterials by treatments to increase its surface reactivity. Such treatments include reactive ball-milling [50–52], oxidation [53], acid treatment [47] and doping with transition metals such as Pd [54–58]. These transition metals serve a catalytic purpose of breaking the chemical bond in molecular hydrogen so that it can be stored in greater quantities on a carbon surface.

Non-carbon-based nanomaterials, such as boron nitride nanotubes, have also been studied for hydrogen storage applications [59–61]. Boron nitride nanotubes are fullerene materials with similar properties to carbon nanotubes. One advantage of boron nitride nanotubes is their greater oxidation resistance with respect to carbon nanotubes.

5.4
Solar Cells

5.4.1
Solar Energy Basics, Including Quantum Confinement

Solar power is highly desirable as a sustainable energy source due to the expected long lifespan of the Sun, on the order of 10 billion years. Solar energy has long been considered an attractive alternative to hydrocarbon fossil fuels, but its widespread adoption has been hindered by the coupled problems of low efficiency and high cost, in addition to the large area required for generation of significant power. Most commercial photovoltaic cells employ either crystalline, polycrystalline or amorphous Si [62]. Photovoltaic cells based on other materials, such as CdTe, $CuInSe_2$, $CuInGaSe_2$ and TiO_2, have comparable or higher efficiencies,

Table 5.1 Reported maximum efficiencies and other data for solar cell materials (From Ref. [62]).

Material	Efficiency (%)	Area (cm^2)	V_{oc} (V)	j_{sc} ($mA\,cm^{-2}$)
a-Si	12.7	1	0.887	19.4
$CuInSe_2$	15.4	0.408	0.515	41.2
$CuInGaSe_2$	19.2	0.470	0.689	35.7
$CuInAlSe_2$	16.9	1.032	0.621	36.0
CdTe	16.5		0.845	25.9

but none is yet cost-effective in comparison with Si photovoltaic cells [62]. The highest efficiency reported for several different solar cell materials is given in Table 5.1 [62].

Nanomaterials have been employed in several reported types of photovoltaic cells, for a variety of different purposes. Probably the classic demonstration of nanotechnology is the dependence of optical bandgap of a semiconductor nanostructure on its dimensions. When the size is comparable to the exciton Bohr radius, quantum confinement effects shift the bandgap of a semiconductor nanostructure to higher energy. This has been widely popularized by the term "quantum dot". The detailed dependence of the optical bandgap on nanostructure dimensions can be determined by solving the Schrödinger equation with the appropriate boundary conditions.

Hence the most straightforward application of nanomaterials to photovoltaic technology is to tune the optical bandgap by varying the size dimensions of a nanostructure. When optical absorption occurs within a nanoparticle, nanowire, nanotube or other nanostructure, the voltage created by exciton formation is captured in an external electrical circuit. The main focus of these efforts has been the metal chalcogenides, including CdS, ZnS, PbS and CdSe. CdTe is particularly appealing as its bandgap (1.45 eV, 855 nm) is nearly ideal for solar terrestrial photoconversion. Figure 5.4 illustrates the terrestrial solar irradiance determined by several different measurements, using detectors both normal to the Earth's surface and tilted [63]. The report of a CdS/CdTe solar cell with 15.8% efficiency in 1993 stimulated an enormous literature on this type of photovoltaic cell [64].

The blue shift of the optical bandgap with decreasing nanostructure dimension has motivated the use of nanomaterials in such solar cells. For example, losses due to optical absorption in the CdS n-type window can be minimized through the use of nanocrystalline CdS, where the optical absorption is blue shifted out of the terrestrial window for light collection [65]. This allows for use of a window material that does not absorb photons, allowing subsequent collection by the absorber material.

More generally, nanostructured absorber layers have several potential advantages over bulk absorber layers. The nanostructure dimension can be adjusted to tune the optical bandgap to match the terrestrial window [66]. The variation in optical properties of chalcogenide nanostructures has been extensively characterized [67, 68]. In addition, the presence of multiple interfaces within the absorber material may increase the effective pathlength by internal light scattering, allowing the use of much

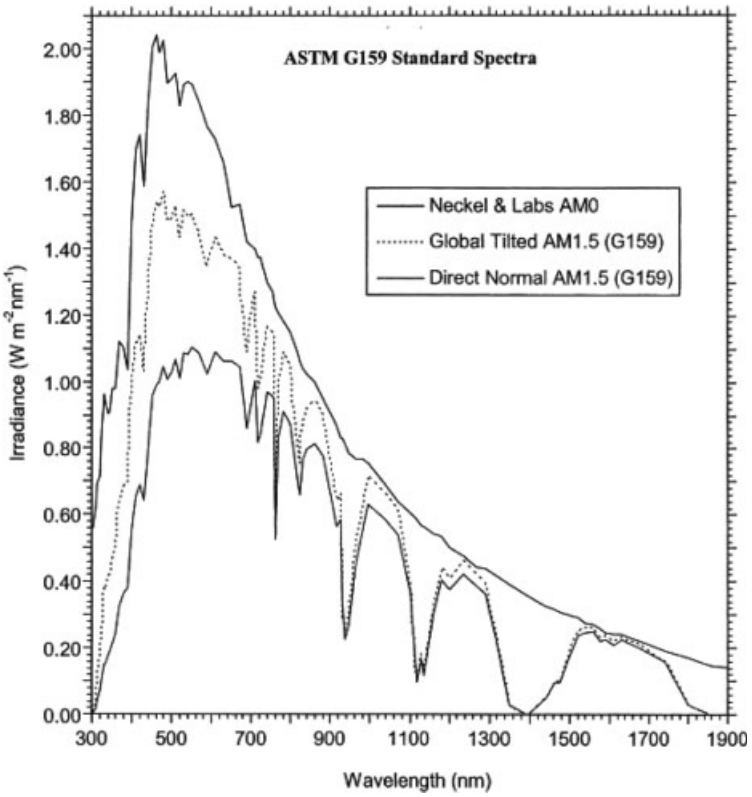


Figure 5.4 Terrestrial solar irradiance window (From Ref. [63]).

thinner absorber layers [69]. For example, the optical pathlength has been reported to increase by a factor of five in nanocrystalline TiO_2 films [70]. The competing effects of maximizing the internal surface area, which favors nanoscale particles and maximizing internal reflections, which favors microscale particles, make this a complex optimization problem [71].

5.4.2 Nanocrystalline Dye-Sensitized Solar Cells

Although TiO_2 is an effective photocatalyst for many applications, for solar cells it has the substantial drawback of an optical bandgap (3.2 eV, 387 nm) in the ultraviolet region of the spectrum, so pure TiO_2 will not absorb throughout most of the terrestrial solar irradiance window shown in Figure 5.4. The addition of a sensitizing dye, typically a transition metal complex that absorbs visible radiation, provides the necessary coverage throughout the terrestrial solar window.

The most widespread application of nanomaterials in solar cells is the use nanocrystalline TiO_2, and to a lesser extent nanocrystalline ZnO and SnO_2, as the absorber layer in DSSCs. First reported by O'Regan and Gratzel in 1991 [72],

the DSSC is an innovative, low-cost alternative to silicon-based photovoltaic cells with the absorber layer constructed from a mesoporous nanocrystalline TiO_2 film sensitized by a monolayer of dye molecules and containing an electrolyte, typically acetonitrile containing the iodide/triiodide redox couple. This can be conceptualized as two percolating continuous random networks: a network of connected TiO_2 nanoparticles and a complementary network of pores filled with organic electrolyte [73]. The dye molecules reside at the interface of these two networks. Because of the small TiO_2 nanoparticle size (~20 nm), the surface area (roughness factor) of the TiO_2 film can be more than 1000 times greater than the geometric surface area. The main obstacles to commercialization of DSSCs are their relatively low efficiency and the stability of the dye molecules during long-term usage.

The following reactions occur within a Gratzel-type DSSC:

$$TiO_2/S + h\nu \rightarrow TiO_2/S^* \tag{5.4}$$

$$TiO_2/S^* \rightarrow TiO_2/S^+ + e_{cb}^- \tag{5.5}$$

$$TiO_2/S^+ + \frac{3}{2}I^- \rightarrow TiO_2/S + \frac{1}{2}I_3^- \tag{5.6}$$

$$TiO_2/S^+ + e_{cb}^- \rightarrow TiO_2/S \tag{5.7}$$

$$\frac{1}{2}I_3^- + e_{pt}^- \rightarrow \frac{3}{2}I^- \tag{5.8}$$

$$\frac{1}{2}I_3^- + e_{cb}^- \rightarrow \frac{3}{2}I^- \tag{5.9}$$

Photon absorption by the dye monolayer (S) adsorbed at the TiO_2 surface in Reaction (5.4) creates a dye molecule in an electronically excited state (S^*). Electron transfer from this excited state to the conduction band (e_{cb}^-) of TiO_2 is shown in Reaction (5.5). The desired oxidation of I^- is shown in Reaction (5.6), while Reactions (5.7) and (5.8) represent two undesired reactions, back reaction of conduction band electrons with the dye and the reduction of I_3^- by injected electrons. An important criterion for high-efficiency operation of the DSSC is that Reaction (5.6) must occur much more quickly than Reactions (5.7) and (5.8). Reaction (5.9) describes I_3^- reduction at the counter electrode to regenerate the electron mediator, I^-. The rapidity of this reaction at the counter electrode, usually Pt, can often limit the overall cell performance, as described further below.

The main drawback of the Gratzel-type DSSC has been the low electron diffusion coefficient. Laser flash-induced transient photocurrent measurements and intensity-modulated photocurrent spectroscopy show that the electron diffusion coefficient in nanocrystalline TiO_2 films is about two orders of magnitude less than in bulk anatase TiO_2 [74–78]. More recent photocurrent transient measurements that systematically varied the average TiO_2 particle size suggest that this behavior arises from electron traps located predominantly at the TiO_2 nanoparticle surface, not within the bulk particle or at interparticle grain boundaries [79]. Ultrafast measurements using THz

spectroscopy suggest that low electron mobilities also arise from local electric field effects that are coupled to the TiO_2 morphology [80].

In analogy with the approaches described above for using carbon nanotubes as fuel cell catalyst support materials, several investigators have tried using one-dimensional TiO_2 nanostructures, such as nanotubes and nanowires, in order to obtain simultaneously both high surface area and the connectivity needed for rapid electron transfer [81]. TiO_2 nanostructures can be fabricated by a variety of different methods, including template synthesis within nanoporous membranes, hydrothermal methods and colloidal methods. Several authors have mixed TiO_2 nanowires or nanotubes with more standard TiO_2 materials, demonstrating improved energy conversion efficiency.

Yoon *et al*. reported the template synthesis of Ti nanowires and nanotubes from $TiCl_4$ within nanoporous alumina membranes [82]. Template synthesis of nanomaterials using commercially available nanoporous alumina and polycarbonate membranes has been widely employed to fabricate metal, semiconductor, ceramic and polymer nanowires and nanotubes [83]. Solar cells containing 10 wt.% of 180–250 nm TiO_2 nanowires and nanotubes with 90 wt.% Degussa P25 TiO_2 exhibited an energy conversion efficiency reported to be 42% higher than that obtained for Degussa P25 TiO_2 alone [82].

Jiu *et al*. reported the growth of TiO_2 nanowires of controlled length by a hydrothermal process in a solution containing a cetyltrimethylammonium bromide (CTAB) surfactant through addition of varying amounts of a triblock copolymer containing poly(ethylene oxide), poly(propylene oxide) and poly(ethylene oxide) [84]. The use of multiple surface-active species in such wet synthesis methods has been shown to allow for the anisotropic growth needed for nanowire formation. Surprisingly, the TiO_2 nanowire morphology survives intact following calcination at 450 °C and sintering at 550 °C, but only in the presence of the triblock copolymer. Nanowires of 20–30 nm diameter and 100–300 nm length have been produced by this method. Earlier studies from the same research group used a slightly different technique and produced TiO_2 nanotubes with diameters of 5–10 nm and lengths of 30–300 nm [85]. DSSCs with an absorber layer containing mixtures of these nanotubes and Degussa P25 TiO_2 showed superior efficiency to those containing only Degussa P25 TiO_2 [86].

Jiu *et al*. also incorporated their TiO_2 nanowires into a DSSC and compared their performance with that of Degussa P25 TiO_2 [84]. However, as these authors acknowledge, this comparison between these two materials is difficult due to their differing crystal structures, degree of crystallinity, packing orientation and porosity. For example, Degussa P25 TiO_2 contains about 80 wt.% anatase and 20 wt.% rutile phases, whereas the nanowire TiO_2 is purely in the anatase phase. Earlier studies have shown that DSSCs constructed using nanocrystalline TiO_2 in the anatase phase are more efficient that those constructed using Degussa P25 TiO_2 [87]. Despite these difficulties in comparison, the DSSC containing TiO_2 nanowires exhibited superior performance to those containing Degussa P25 TiO_2 for thick (<10 μm) absorber layers, where the advantages in terms of enhanced electron diffusion rates would be most evident.

This group has also grown TiO_2 nanowires using an "oriented attachment" method that ensures that the nanowires grow parallel to each other [88]. This process involves the use of surfactant-aided anatase crystal growth process near room temperature. This geometry is designed to maximize the rate of electron transfer in a DSSC and yielded an efficiency of 9.3%.

Several other groups have incorporated TiO_2 nanowires and/or nanotubes into DSSCs. The fabrication of TiO_2 nanowires has been demonstrated using the following sol–gel alkyl halide elimination reaction [89]:

$$TiCl_4 + Ti(OR)_4 \rightarrow 2TiO_2 + 4RCl \tag{5.10}$$

The size and the crystal phase (anatase vs. rutile) of the nanowires can be controlled to some extent by the injection rate of the titanium precursors. As the injection rate increases, the nanowire diameter decreases and the proportion of anatase phase increases. A DSSC fabricated from 81 wt.% anatase and 19 wt.% rutile TiO_2 nanowires, which is similar to the mass fraction of these phases in Degussa P25 TiO_2, exhibited a higher efficiency (3.83%) than DSSCs fabricated from purely anatase or purely rutile nanowires.

Several other examples of TiO_2 nanomaterials in DSSCs should also be noted. A DSSC has also been constructed using an absorber that contains titanate ($H_2Ti_3O_7$) nanotubes fabricated by a hydrothermal process [90]. TiO_2 nanowires have been directionally grown by electrospinning and incorporated into a DSSC with a highly viscous gel electrolyte, attaining an efficiency of 6.2% [91]. Rather than creating individual nanotubes, a Ti film can be anodized to create a continuous, oriented and highly ordered array of TiO_2 nanotubes that has been demonstrated as part of a DSSC [92].

Another intriguing example of nanomaterials in DSSCs is the fabrication of core–shell nanoparticles for use in the absorber layer. TiO_2 nanoparticles have been coated with insulating oxides or wide bandgap semiconductors. The coating is designed to prevent interfacial recombination, but must be thin enough to allow electron tunneling between the dye and TiO_2. Although this has been studied mainly with Al_2O_3 coatings [93, 94] this effect can be seen using a wide variety of oxide coatings [95, 96]. This idea has also been extended to oxide coating of ZnO nanowires [97].

5.4.3
Nanomaterials in Solar Cell Counter Electrodes

Nanomaterials are also important for other elements within a DSSC besides the absorber material. Equation (5.6) occurs at the counter electrode and is important for regenerating the electron acceptor, I^-. If the rate of I_3^- reduction is not sufficiently rapid, then this will limit the overall DSSC efficiency. Pt is typically the electrocatalytic cathode material of choice in a wide variety of electrochemical systems. However, owing to its high cost, minimizing the amount of Pt used is highly desirable. The rapidity of I_3^- reduction can be quantified by its charge transfer resistance (R_{ct}), which can be conveniently determined by electrochemical impedance spectroscopy (EIS). EIS studies of the reaction rate at the Pt counter electrode indicate that a 2–3 nm

Pt film is sufficient to obtain both rapid I_3^- reduction and adequate electrical conductivity [98, 99]. In order to reduce costs but maintain electrocatalytic activity, Pt nanoparticles dispersed within an indium tin oxide (ITO) film have also been proposed for use as the counter electrode in DSSCs [100].

5.5
Lithium Ion Battery Anode Materials

5.5.1
Lithium Ion Batteries

Lithium ion batteries have now become ubiquitous due to their high voltage (~3.6 V), high energy density (125 W h kg^{-1}) and long life cycle (>1000 cycles) relative to other battery types, such as Ni–Cd, Ag–Zn, Ni–hydride and lead acid batteries. Applications are widespread in portable consumer electronics, including notebook computers, cellular telephones, MP3 players and camcorders. Like other batteries, Li ion batteries are composed of a cathode and an anode, separated by an electrolyte. This type of battery has been called the "swing" battery because Li ions are exchanged alternately between the cathode and anode during battery charging and discharging.

The cathode material, which provides a source of Li ions during battery charging, is typically either layered $LiMO_2$, where M can be Co, Ni, Al or Mn; Li manganese oxide spinels ($LiMn_2O_4$); or other Li salts. Although $LiCoO_2$ is widely used as cathode material due to its long cycle life and reasonable energy density, the high cost and high toxicity of Co have limited its use to relatively small batteries. Efforts are under way to replace Co either partly or completely with Ni or Mn, which are less costly and less toxic. The prototypical anode material is carbon/graphite, which accepts Li ions during battery charging by intercalation between adjacent graphitic planes. Li metal was used as the anode for early Li ion batteries, but this caused safety problems due to the high activity of metallic Li. During battery discharge, the direction of Li ion migration is reversed. The electrochemical reactions at the anode and cathode are shown below for the prototypical cathode and anode materials during battery charging:

$$LiCoO_2 \rightarrow Li_{1-x}CoO_2 + xLi^+ + xe^- \tag{5.11}$$

$$6C + xLi^+ + xe^- \rightarrow Li_xC_6 \tag{5.12}$$

The intervening electrolyte material is normally a polymer material, most commonly poly(ethylene oxide) (PEO). Significant research efforts have been undertaken to find new materials for the cathode, anode and electrolyte to improve the life cycle and increase the energy density of the Li ion battery. Many of these efforts have involved the development of nanomaterials, in large part due to their higher internal surface area [101].

Nanomaterials that have been proposed as an anode material in Li ion batteries are mainly metal nanoparticles, carbon nanotubes and nanocomposites that combine

these two materials. The benchmark for comparison is intercalation of Li in graphite, where the theoretical limit is determined by the compound formed, LiC_6, which is equivalent to a storage capacity of 372 mA h g^{-1}. High Li capacity has been obtained for numerous alternative elements, including Ag, Sn, Al, Si, Sb, Bi and Pb [102]. One of the most widely studied anode materials is Sn, galvanized by researchers at Fuji Photo Film, who reported a high Li storage capacity for Sn composite electrodes [103]. The recently announced Sony Nexelion lithium ion battery is a composite containing several materials, including Sn and C. Sn has a maximum theoretical Li capacity arising from the stoichiometry $Li_{22}Sn_5$, corresponding to a storage capacity of 994 mA h g^{-1}.

The other widely studied lithium storage material is Si, which has a higher theoretical lithium storage capacity of 4200 mA h g^{-1}, arising from a stoichiometry of $Li_{22}Si_5$, than other widely studied materials. Applications of Si for lithium ion storage, which will be discussed later, are hindered by the even larger volume change upon lithiation (>300%) and by the poor electrical conductivity of Si. In addition to Sn and Si, the other alternative Li storage materials that have been studied include Sb and Cu_6Sn_5. The Li storage capacity of Sb has been reported to be 660 mA h g^{-1}, arising from the stoichiometry Li_3Sb. Another metal that has been widely investigated is Cu_6Sn_5, which first forms $Li_2Cu_6Sn_5$, which decomposes upon further lithium addition and forms $Li_{22}Sn_5$ surrounded by a Cu matrix. The Li storage capacity of Cu_6Sn_5 is about 350 mA h g^{-1}.

5.5.2
Nanomaterials for Lithium Ion Storage: Sn Nanoparticles

Unfortunately, the large volume change associated with Li compound formation during repeated cycling of these alternative lithium storage materials gradually causes a loss of electrical contact and mechanical degradation, with the anode material cracking repeatedly and eventually becoming pulverized. Graphite materials, on the other hand, show little volume change during Li cycling. In order to accommodate the volume change associated with Li storage, Li battery anodes containing micro- and nanocrystalline metals have been tested and show a greater ability to tolerate mechanical strain [102]. However, although the performance of anodes constructed from nanoparticles is greatly improved, they still suffer gradual mechanical degradation and a new problem is encountered, nanoparticle agglomeration [104].

In an elegant quantitative study, Noh *et al*. compared the capacity retention for Sn nanoparticles of various diameters, as shown in Table 5.2 [105]. They considered these results to be somewhat surprising, given that the surface area involved in the formation of the semiconductor electrolyte interface (SEI) increases with decreasing particle size, as does the extent of irreversible Li consumption during SEI formation. They reported that the initial capacity of Li anode constructed from 10 nm diameter Sn nanoparticles was as high as 1000 mA h g^{-1}, with little or no capacity fade.

Several ingenious strategies have been attempted to circumvent the problem of volume cycling during repeated Li ion battery charging/discharging. Anodes have

Table 5.2 Capacity retention for Sn nanoparticles of different diameters after 50 cycles of 0.5 C (From Ref. [105]).

Diameter (nm)	Capacity retention (%)
300	58
200	68
20	87
10	94

been constructed that are nanocomposites of active and inactive materials, with the inactive material serving as a mechanical buffer to accommodate the volume change [106–108]. For example, composites between Sn_2Fe (active) and $SnFe_3C$ (inactive) prepared by high energy ball-milling have been reported to have reversible Li capacities twice that of graphite materials [106, 107]. A serious disadvantage of this technique is that the energy density per unit mass of the anode material is reduced, sometimes dramatically, by the presence of the inactive component.

To circumvent this problem, several groups have instead created composites of Li-active nanoparticles (usually Sn) with a buffer material that is also active, usually carbon in graphitic form. Since graphite is fairly soft, it is expected to form a mechanical buffer during repeated Li insertion/removal [109–111]. One complication of this approach is that Sn nanoparticles are relatively difficult to fabricate. As is well known in colloid chemistry, formation of metal and metal oxide nanostructures requires careful control of reaction conditions, typically requiring the use of surfactants, stabilizers and/or capping agents in order both to control nanostructure growth and simultaneously to prevent aggregation [112].

Wang *et al*. have reported the fabrication of Sn nanoparticles of 2–5 and 7–13 nm diameter by reduction of $SnCl_4$ with $NaBH_4$ [110]. The Sn nanoparticles were then dispersed in graphite to form a composite Li anode containing 10.3 wt.% Sn. One of the challenges of such techniques is obtaining a high fraction of Sn in the composite anode. During both the preparation and dispersion steps, the Sn nanoparticles were protected from agglomeration by the presence of 1,10-phenanthroline, which forms a coordination compound with Sn [110]. This anode exhibited a 415 mA h g^{-1} Li storage capacity and was 91.3% reversible after 60 charge/discharge cycles.

Caballero *et al*. reported the formation of Sn nanoparticles by reduction of $SnCl_4$ by KBH_4 in the presence of cellulose fibers [111]. The Sn nanoparticles apparently nucleate and grow only on the surface of the cellulose fibers, preventing agglomeration. This cellulose–Sn nanocomposite exhibits a storage capacity of 600 mA h g^{-1} as an Li anode and was reported to exhibit reversible charge/discharge cycling, although specific numbers for capacity retention were not provided [111].

A particularly ingenious method for forming a nanocomposite anode from an Li active metal (usually Sn) and an active support (usually carbon) is to encapsulate the metal nanoparticles within carbon spheres [113–117]. This protects the metal nanoparticles both from agglomeration and from mechanical degradation arising

from volume expansion. So far, this technique has been demonstrated only for particles larger than the nanoscale (1–100 nm), but could in principle be extended to smaller particles. Unless the carbon coating can be made graphitic, the Li storage capacity will not be very high. In some cases, the addition of a conductive material such as Cu silicide within carbon coating Si particles may improve electrical contact during repeated lithium charge/discharge cycles [113].

Some of the studies that involve active metals within carbon shells provide some interesting, direct comparisons between cycling stability of these materials, with and without their carbon shells [114, 117]. Jung *et al.* created carbon encapsulation about commercial Sn particles by hydrophobization in 1-octanethiol, dispersion in a resorcinol–formaldehyde microemulsion, polymerization and carbonization of the coating by high-temperature annealing [114]. Without first making the Sn particle surface hydrophobic, the Sn particle surface will not be wetted during this synthesis. The performance of this material (CSP) as an Li anode was then compared with two control materials, Sn nanoparticles that are not encapsulated (SN) and a random mixture of spherical carbon powder with Sn particles (MIX) with the same nominal composition, 20 wt.% Sn. The results for 40 charge/discharge cycles are shown in Figure 5.5 [114]. Clearly, the cyclability of the carbon-encapsulated Sn particles is far greater than either control anode, demonstrating the stabilizing effect of the carbon encapsulation.

Another study used a similar fabrication technique to encapsulate Sn_2Sb particles with carbon through polymerization of resorcinol–formaldehyde microemulsion followed by high-temperature annealing [117]. The Li anode performance of the carbon microsphere (CM)-encapsulated Sn_2Sb particles was compared with that of Sn_2Sb powder, as shown in Figure 5.6. While the initial Li storage capacity of the Sn_2Sb powder is 689 mA h g^{-1}, only 20.3% of this capacity is retained after 60 charge/discharge cycles [117]. On the other hand, the CM-encapsulated Sn_2Sb particles retained 87.7% of their original storage capacity of 649 mA h g^{-1} after 60 charge/discharge cycles [117].

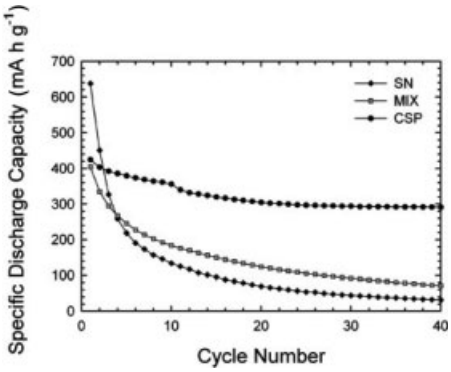


Figure 5.5 Li discharge capacity with cycle time for Sn particles in Li battery anode. (From Ref. [114]).

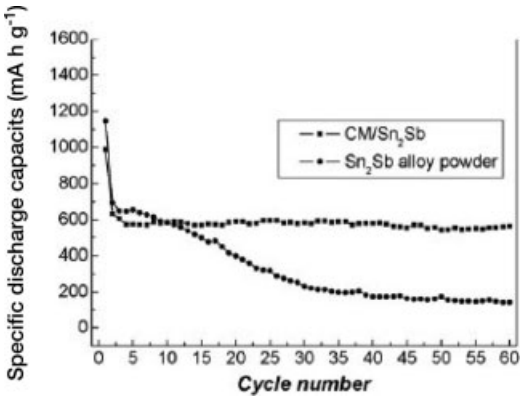


Figure 5.6 Li discharge capacity with cycle time for Sn_2Sb powder, with and without a carbon microsphere (CM) coating. (From Ref. [117]).

5.5.3
Nanomaterials for Lithium Ion Storage: Si Nanocomposites

Most of the examples given above describe ingenious techniques to improve the cycle stability of Sn-based lithium storage materials through the use of nanoparticles and nanocomposites. After Sn, the most widely studied alternative lithium storage material is Si. Si has an even higher theoretical lithium storage capacity (4200 mA h g^{-1}) than does Sn (994 mA h g^{-}), although the low conductivity of Si requires the use of some type of conductive filler. Not surprisingly, the most widely studied composite Si anode materials are Si–carbon composites [118–122]. One difficulty with Si–C composite anodes is that high-temperature processing may form the compound SiC, which is inactive for lithium storage [118].

Holzapfel *et al.* deposited nanocomposite Si–graphite films by chemical vapor deposition (CVD) of Si from SiH_4 on a graphitized fine particle carbon film, Timrex KS6 [121]. This active material was then mixed with one of two different binders, dissolved in a petroleum ether solution and dried under vacuum. The resulting electrode film contains about 7 wt.% Si in the form of Si nanoparticles ranging from 10 to 20 nm in diameter [121]. The Li storage capacity of this Si–graphite electrode declined gradually from 2500 to 1900 mA h g^{-1} during 100 charge/discharge cycles.

Wang *et al.* formed an Si–C nanocomposite by high-energy ball-milling of commercially available powders, 80 nm Si nanoparticles and 10 μm spherical mesocarbon microbeads (MCMB) [120]. MCMB represent an industry standard for lithium storage and exhibit capacities ranging from 300 to 340 mA h g^{-1} with excellent stability during cycling. After 20 h of ball-milling, the spherical MCMB lose their original structural integrity and the Si nanoparticles become dispersed within the MCMB. These composite anodes were tested and compared with Li anodes constructed from 80 nm Si powder, 10 μm MCMB and composite anodes containing much larger Si particles (20 μm). The best performance was obtained for

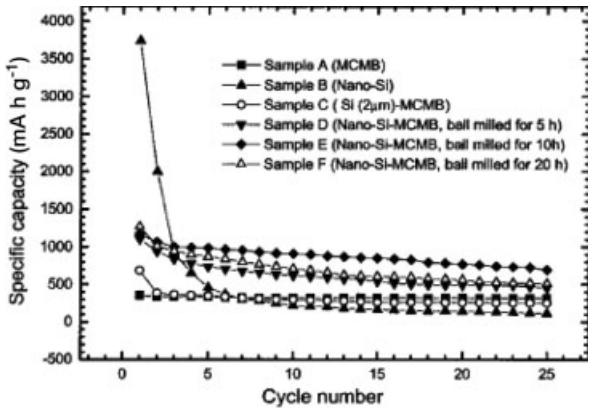


Figure 5.7 Li discharge capacity with cycle time for nanocrystalline Si, MCMB and ball-milled mixtures of the two. (From Ref. [120]).

composite anodes constructed from 20 wt.% Si and 80 wt.% MCMB, which showed good reversibility over 25 charge/discharge cycles with a capacity of 1066 mA h g^{-1}, as illustrated in Figure 5.7 [120].

5.5.4
Nanomaterials for Lithium Ion Storage: Carbon Nanotubes and Carbon Nanotube-Based Composites

MWNTs have been reported to exhibit a reversible Li capacity of 400–1000 mA h g^{-1} [123, 124], whereas SWNTs have been reported to exhibit a reversible Li capacity in the range of 450–600 mA h g^{-1}, which can be increased to 1000 mA h g^{-1} by ball-milling to increase the surface area [125]. This capacity is believed to be associated with defect sites within the nanotubes, sites between adjacent grapheme sheets within MWNTs and sites between adjacent SWNTs. Even these early reports noted significant shortcomings that would likely prevent commercialization, including an extremely high irreversible capacity, a gradual decline in reversible capacity and a strong dependence of Li capacity on carbon nanotube structure and preparation technique. Subsequent investigations concluded that carbon nanotubes alone were unlikely to provide an adequate Li anode material, but were fairly promising for the development of supercapacitors [126, 127].

Despite diminishing interest in Li anode materials constructed solely from carbon nanotubes, interest remains in the use of carbon nanotubes within nanocomposite Li anode materials. Several research groups have studied Li ion storage in nanocomposites composed of mixtures of carbon nanotubes and Sn, Sb or Ni nanoparticles [128–134]. These nanocomposite materials generally exhibit extremely high initial Li insertion capacity, which then gradually declines.

Kumar *et al*. reported a method for creating Sn–carbon nanotube nanocomposites that involves *in situ* formation of Sn nanoparticles within the carbon nanotubes [130]. Carbon nanotubes were grown by chemical vapor deposition from acetylene using a

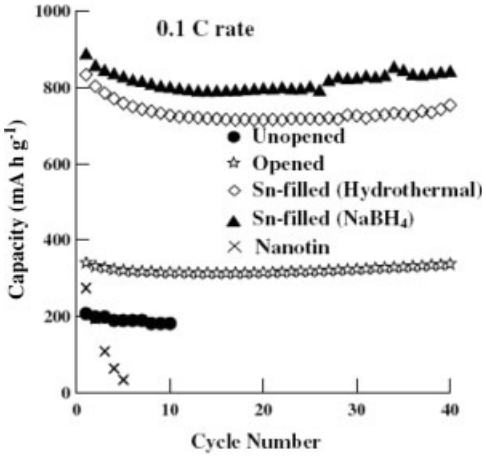


Figure 5.8 Li discharge capacity with cycle time for Sn–carbon nanotube composite with Sn reduced by two different methods. (From Ref. [130]).

ferrocene catalyst. Following removal of catalyst and amorphous carbon, the carbon nanotube ends were oxidatively opened using concentrated HNO_3. Aqueous $SnCl_2$ was introduced into the carbon nanotubes by capillary action, then reduced by either hydrothermal treatment or $NaBH_4$. The Li insertion capacity of the Sn nanoparticle–carbon nanotube composite material is shown in Figure 5.8, which illustrates that an approximately stable Li capacity on the order of 800 mA h g^{-1} is attained after 10 cycles [130].

Nanocomposites fabricated by ball-milling of carbon nanotubes with Si have also been prepared and tested for their Li insertion capacity [135, 136]. High and stable reversible Li capacity has not yet been obtained by such techniques.

5.5.5
Lithium Ion Storage: Further Considerations

An often overlooked aspect of lithium ion battery anodes is the inherent difference in electrocatalytic behavior between graphite and active metals [137]. Graphite surfaces exhibit well-known differences between the electrochemistry of basal planes and edge surfaces, whereas the surfaces of active metals appear to be more electrochemically homogeneous. The solid electrolyte interface (SEI) on basal planes of graphite is known to be relatively thin and contain many organic decomposition products, while the edge sites exhibit a thicker SEI that contains mainly inorganic species [137]. The mechanism of SEI formation is also found to differ in ethylene carbonate and propylene carbonate electrolytes, which are the two most common electrolytes in lithium ion batteries.

The studies discussed here report many impressive results for lithium battery anodes constructed from composite materials containing a variety of identifiable nanoparticle and nanotube ingredients. However, Dahn's group suggested that an

alternative approach, deposition of nearly homogeneous amorphous samples of the same materials, may provide a superior approach to obtaining the maximum lithium storage capacity [138–140]. Observations of large 10–20 μm $Si_{0.64}Sn_{0.36}$ particles suggest that they move intact over a much longer length scale during volume changes associated with lithium cycling then previously believed [140]. This supports the authors' assertion that the problem with degradation observed in metal nanoparticle lithium anodes arises not from the inability of the particles to absorb the volume change, but from the inability of their binder to absorb the volume change. In support of this argument, the cycling stability of lithium anodes containing $Si_{0.64}Sn_{0.36}$ particles was shown to be greatly improved by the use of an elastomeric binder, which provides a mechanical buffer for volume changes during charge/discharge cycling [138].

Dahn's group has also produced amorphous $Si_{1-x}Sn_x$ films with extremely high lithium storage capacity, more than 3500 mA h g^{-1}, with only 1–2% capacity fade per cycle [139]. These films have been produced by magnetron sputtering and investigated over a wide range of $Si_{1-x}Sn_x$ compositions. Amorphous films are produced for compositions with $x < 0.45$. The failure mode of such anode materials is argued to be qualitatively different than that observed for composite anodes containing metal nanoparticles [139]. Since the volume change occurs uniformly across the anode material, internal stresses at interfacial boundaries do not occur. Instead, the entire anode homogeneously expands and contracts during lithium charging and discharging. The most common failure mode is then reported to arise from an eventual loss of electrical contact [139]. One possible shortcoming of the use of amorphous thin films is that in some cases, a thermodynamic driving force may exist to form crystalline structures.

References

1 P. Costamagna, S. Srinivasan, *J. Power Sources* 2001, **102**, 253–69.
2 T.R. Ralph, M.P. Hogarth, *Platinum Met. Rev.* 2002, **46**, 3–14.
3 T.R. Ralph, M.P. Hogarth, *Platinum Met. Rev.* 2002, **46**, 117–35.
4 M.P. Hogarth, T.R. Ralph, *Platinum Met. Rev.* 2002, **46**, 146–64.
5 H.A. Gasteiger, S.S. Kocha, B. Sompalli, F.T. Wagner, *Appl. Catal. B* 2005, **56**, 9–35.
6 E. Antolini, J.R.C. Salgado, E.R. Gonzalez, *J. Power Sources* 2006, **160**, 957–68.
7 E. Antolini, J.R.C. Salgado, E.R. Gonzalez, *Appl. Catal. B* 2006, **63**, 137–49.
8 H. Liu, C. Song, L. Zhang, J. Zhang, H. Wang, D.P. Wilkinson, *J. Power Sources* 2006, **155**, 95–110.
9 J.H. Wee, K.Y. Lee, *J. Power Sources* 2006, **157**, 128–35.
10 K. Lee, J. Zhang, H. Wang, D.P. Wilkinson, *J. Appl. Electrochem.* 2006, **36**, 507–22.
11 B. Wang, *J. Power Sources* 2005, **152**, 1–15.
12 L. Zhang, J. Zhang, D.P. Wilkinson, H. Wang, *J. Power Sources* 2006, **156**, 171–82.
13 X. Wang, W. Li, Z. Chen, M. Waje, Y. Yan, *J. Power Sources* 2006, **158**, 154–59.
14 L. Li, Y. Xing, *J. Electrochem. Soc.* 2006, **153**, A1823–28.
15 B.S. Files, B.M. Mayeaux, *Adv. Mater. Proc.* 1999, **10**, 47–49.
16 N.R.K.V. Reddy, E.B. Anderson, E.J. Taylor, *US Patent* 5 084 144, 1992.
17 E.J. Taylor, E.B. Anderson, N.R.K. Vilambi, *J. Electrochem. Soc.* 1992, **139**, L45–46.

18 D.J. Guo, H.L. Li, *J. Electroanal. Chem.* 2004, **573**, 197–202.

19 Z. He, J. Chen, D. Liu, H. Zhou, Y. Kuang, *Diamond Relat. Mater.* 2004, **13**, 1764–70.

20 C. Wang, M. Waje, X. Wang, J.M. Tang, R.C. Haddon, Y. Yan, *Nano Lett.* 2004, **4**, 345–48.

21 P.J. Britto, K.S.V. Santhanam, A. Rubio, J.A. Alonso, P.M. Ajayan, *Adv. Mater.* 1999, **11**, 154–57.

22 J.M. Nugent, K.S.V. Santhanam, A. Rubio, P.M. Ajayan, *Nano Lett.* 2001, **1**, 87–91.

23 J.J. Gooding, *Electrochim. Acta* 2005, **50**, 3049–60.

24 W. Li, C. Liang, W. Zhou, H. Han, Z. Wei, G. Sun, Q. Xin, *Carbon* 2002, **40**, 791–94.

25 W. Li, C. Liang, W. Zhou, J. Qiu, Z. Zhou, G. Sun, Q. Xin, *J. Phys. Chem. B* 2003, **107**, 6292–99.

26 Z. Liu, L.M. Gan, L. Hong, W. Chen, J.Y. Lee, *J. Power Sources* 2005, **139**, 73–78.

27 M.M. Waje, X. Wang, W. Li, Y. Yan, *Nanotechnology* 2005, **16**, S395–400.

28 X. Li, M. Hsing, *Electrochim. Acta* 2006, **51**, 5250–58.

29 M.M. Shaijumon, S. Ramaprabhu, N. Rajalakshmi, *Appl. Phys. Lett.* 2006, **88**, 253105.

30 X. Wang, M. Waje, Y. Yan, *Electrochem. Solid-State Lett.* 2005, **8**, A42–44.

31 W. Li, X. Wang, Z. Chen, M. Waje, Y. Yan, *Langmuir* 2005, **21**, 9386–89.

32 M. Carmo, V.A. Paganin, J.M. Rosolen, E.R. Gonzalez, *J. Power Sources* 2005, **142**, 169–76.

33 Y. Liang, H. Zhang, B. Yi, Z. Zhang, Z. Tan, *Carbon* 2005, **43**, 3144–52.

34 G. Grishkumar, M. Rettker, R. Underhile, D. Binz, K. Vinodgopal, P. McGinn, P. Kamat, *Langmuir* 2005, **21**, 8487–94.

35 E.S. Steigerwalt, G.A. Deluga, C.M. Lukehart, *J. Phys. Chem. B* 2002, **106**, 760–66.

36 G. Grishkumar, T.D. Hall, K. Vinodgopal, P. Kamat, *J. Phys. Chem. B* 2006, **110**, 107–14.

37 W. Li, X. Wang, Z. Chen, M. Waje, Y. Yan, *J. Phys. Chem. B* 2006, **110**, 15353–58.

38 K.T. Jeng, C.C. Chien, N.Y. Hsu, S.C. Yen, S.D. Chiou, S.H. Lin, W.M. Huang, *J. Power Sources* 2006, **160**, 97–104.

39 J. Pabhuram, T.S. Zhao, Z.K. Tang, R. Chen, Z.X. Liang, *J. Phys. Chem. B* 2006, **110**, 5245–52.

40 L. Zhou, *Renew. Sustain. Energy Rev.* 2005, **9**, 395–408.

41 M.S. Dresselhaus, K.A. Williams, P.C. Eklund, *MRS Bull.* 1999, **24** (11), 45–50.

42 US Department of Energy, Energy Efficiency and Renewable Energy (EERE), *Hydrogen, Fuel Cells and Infrastructure Technologies Program, Multi-year R&D Plan*, 2005, http://www1.eere.energy.gov/hydrogenandfuelcells/mypp/pdfs/storage.pdf.

43 M. Hirscher, H. Becher, M. Haluska, U. Detlaff-Weglikowska, A. Quintel, G.S. Duesberg, Y.M. Choi, P. Downes, M. Hulman, S. Roth, I. Stepanek, P. Bernier, *Appl. Phys. B* 2001, **72**, 129–132.

44 G.G. Tibbetts, G.P. Meisner, C.H. Olk, *Carbon* 2001, **39**, 2291–2301.

45 A. Lan, A. Mukasyan, *J. Phys. Chem. B* 2005, **109**, 16011–16.

46 R. Zacharia, K.Y. Kim, A.K.M. Fazle-Kibria, K.S. Nahm, *Chem. Phys. Lett.* 2005, **412**, 369–75.

47 H. Takagi, H. Hatori, Y. Soneda, N. Yoshizawa, Y. Yamada, *Mater. Sci. Eng. B* 2004, **108**, 143–47.

48 J.M. Blackman, J.W. Patrick, C.E. Snape, *Carbon* 2006, **44**, 918–27.

49 A.D. Lucking, L. Pan, D.L. Narayanan, C.E.B. Clifford, *J. Phys. Chem. B* 2005, **109**, 12710–17 16.

50 S. Orimo, T. Matsushima, H. Fujii, T. Fukunaga, G. Majer, *J. Appl. Phys.* 2001, **90**, 1545–49.

51 F. Liu, X. Zhang, J. Cheng, J. Tu, F. Kong, W. Huang, C. Chen, *Carbon* 2003, **41**, 2527–32.

52 X.B. Yu, G.S. Walker, N. Bowering, D.M. Grant, J. Shen, Z. Wu, B. J. Xia, *Electrochem. Solid-State Lett.* 2005, **6**, A596–98.

53 M.R. Smith, E.W. Bittner, W. Shi, J.K. Johnson, B.C. Bockrath, *J. Phys. Chem. B* 2003, **107**, 3752–60.

54 D. Lupu, A.R. Biris, I. Misan, A. Jianu, G. Holzhuter, E. Burkel, *Int. J. Hydrogen Energy* 2004, **29**, 97–102.

55 E. Yoo, L. Gao, T. Komatsu, N. Yagai, K. Arai, T. Yamazaki, K. Matsuishi, T. Matsumoto, J. Nakamura, *J. Phys. Chem. B* 2004, **108**, 18903–07.

56 H. Takagi, H. Hatori, Y. Yamada, *Carbon* 2005, **43**, 3037–39.

57 A. Anson, E. LaFuente, E. Urriolabeitia, R. Navarro, A.M. Benito, W.K. Maser, M. T. Martinez, *J. Phys. Chem. B* 2006, **110**, 6643–48.

58 C.K. Back, G. Sandi, J. Prakash, J. Hranisavljevic, *J. Phys. Chem. B* 2006, **110**, 16225–31.

59 T. Oku, M. Kuno, *Diamond Relat. Mater.* 2003, **12**, 840–45.

60 T. Oku, M. Kuno, I. Narita, *J. Phys. Chem. Solids* 2004, **65**, 549–52.

61 X. Chen, X.P. Gao, H. Zhang, Z. Zhou, W.K. Hu, G.L. Pan, H.Y. Zhu, T.Y. Yan, D.Y. Song, *J. Phys. Chem. B* 2005, **109**, 11525–29.

62 A. Morales-Acevedo, *Solar Energy* 2006, **80**, 675–81.

63 C.A. Gueymard, D. Myers, K. Emery, *Solar Energy* 2002, **73**, 443–67.

64 J. Britt, C. Ferekides, *Appl. Phys. Lett.* 1993, **62**, 2851–52.

65 V.P. Singh, R.S. Singh, G.W. Thompson, V. Jayaraman, S. Sanagapalli, V.K. Rangari, *Solar Energy Mater. Solar Cells* 2004, **81**, 293–303.

66 R.S. Singh, V.K. Rangari, S. Sanagapalli, V. Jayaraman, S. Mahendra, V.P. Singh, *Solar Energy Mater. Solar Cells* 2004, **82**, 315–330.

67 A.P. Alivisatos, A.L. Harris, N.J. Levinos, M.L. Steigerwald, L.E. Brus, *J. Chem. Phys.* 1988, **89**, 4001–11.

68 A.P. Alivisatos, *J. Phys. Chem.* 1996, **100**, 13226–39.

69 A. Usami, *Chem. Phys. Lett.* 1997, **277**, 105–08.

70 K. Ernst, A. Belaidi, R. Konenkamp, *Semicond. Sci. Technol.* 2003, **18**, 475–79.

71 J. Ferber, J. Luther, *Solar Energy Mater. Solar Cells* 1998, **54**, 265–75.

72 B. O'Regan, M. Gratzel, *Nature* 1991, **353**, 737–40.

73 A.J. Frank, N. Kopidakis, J. van de Lagemaat, *Coord. Chem. Rev.* 2004, **248**, 1165–79.

74 A. Solbrand, H. Lindstrom, H. Rensmo, A. Hegfeldt, S.E. Lindquist, S. Sodergren, *J. Phys. Chem. B* 1997, **101**, 2514–18.

75 L. Dloczik, O. Ileperuma, I. Lauermann, L.M. Peter, E.A. Ponomarev, G. Redmond, N.J. Shaw, I. Uhlendorf, *J. Phys. Chem. B* 1997, **101**, 10281–89.

76 S. Nakade, S. Kambe, T. Kitamura, Y. Wada, S. Yanagida, *J. Phys. Chem. B* 2001, **105**, 9150–52.

77 S. Kambe, S. Nakade, T. Kitamura, Y. Wada, S. Yanagida, *J. Phys. Chem. B* 2002, **106**, 2967–72.

78 A.C. Fisher, L.M. Peter, E.A. Pomonarev, A.B. Walker, K.G.U. Wijayantha, *J. Phys. Chem. B* 2004, **104**, 949–58.

79 N. Kopidakis, N.R. Neale, K. Zhu, J. van de Lagemaat, A.J. Frank, *Appl. Phys. Lett.* 2005, **87**, 202106.

80 E. Hendry, M. Koeberg, B. O'Regan, M. Bonn, *Nano Lett.* 2006, **6**, 755–59.

81 K.P. Jayadevan, T.Y. Tseng, *J. Nanosci. Nanotechnol.* 2005, **5**, 1768–84.

82 J.H. Yoon, S.R. Jang, R. Vittal, J. Lee, K.J. Kim, *J. Photochem. Photobiol. A* 2006, **180**, 184–88.

83 A. Huczko, *Appl. Phys. A* 2000, **70**, 365–76.

84 J. Jiu, S. Isoda, F. Wang, M. Adachi, *J. Phys. Chem. B* 2006, **110**, 2087–92.

85 M. Adachi, Y. Murata, I. Okada, S. Yohikawa, *J. Electrochem. Soc.* 2003, **150**, G488–93.

86 S. Ngamsinlapasathian, S. Sakulkhaemaruethai, S. Pavasupree, A. Kitiyanan, T. Sreethawong, Y. Suzuki, S. Yoshikawa, *J. Photochem. Photobiol. A* 2004, **164**, 145–51.

87 S. Kambe, K. Murakoshi, T. Kitamura, Y. Wada, S. Yanagida, H. Kominami, Y. Kera, *Solar Energy Mater. Solar Cells* 2000, **61**, 427–41.

88 M. Adachi, Y. Murata, J. Takao, J. Jiu, M. Sakamoto, F. Wang, *J. Am. Chem. Soc.* 2004, **126**, 14943–49.

89 B. Koo, J. Park, Y. Kim, S.H. Choi, Y.E. Sung, T. Hyeon, *J. Phys. Chem. B* 2006, **110**, 24318–23.

90 M. Wei, Y. Konishi, H. Zhou, H. Sugihara, H. Arakawa, *J. Electrochem. Soc.* 2006, **153**, A1232–36.

91 M.Y. Song, Y.R. Ahn, S.M. Jo, D.Y. Kim, J. Y. Ahn, *Appl. Phys. Lett.* 2005, **87**, 113113.

92 G.K. Mor, K. Shankar, M. Paulose, O.K. Varghese, C.A. Grimes, *Nano Lett.* 2006, **6**, 215–18.

93 A. Zaban, S.G. Chen, S. Chappel, B.A. Gregg, *Chem. Commun.* 2000, 2231–32.

94 G.R.R.A. Kumara, K. Tennakone, V.P.S. Perera, A. Konno, S. Kaneko, M. Okuya, *J. Phys. D* 2001, **34**, 868–873.

95 A. Kay, M. Grätzel, *Chem. Mater.* 2002, **14**, 2930–35.

96 D. Menzies, Q. Dai, Y.B. Cheng, G.P. Simon, L. Spiccia, *Mater. Lett.* 2005, **59**, 1893–96.

97 M. Law, L.E. Greene, A. Radenovic, T. Kuykendall, J. Liphardt, P. Yang, *J. Phys. Chem. B* 2006, **110**, 22652–63–23.

98 A. Hauch, A. Georg, *Electrochim. Acta* 2001, **46**, 3457–66.

99 X. Fang, T. Ma, G. Guan, M. Akiyama, T. Kida, E. Abe, *J. Electroanal. Chem.* 2004, **570**, 257–63.

100 S. Katusic, P. Albers, R. Kern, F.M. Petrat, R. Sastrawan, S. Hore, A. Hinsch, A. Gutsch, *Solar Energy Mater. Solar. Cells* 2006, **90**, 1983–99.

101 H.K. Liu, G.X. Wang, Z. Guo, J. Wang, K. Konstantinov, *J. Nanosci. Nanotechnol.* 2006, **6**, 1–15.

102 J.O. Besenhard, J. Yang, M. Winter, *J. Power Sources* 1997, **68**, 87–90.

103 Y. Idota, T. Kubota, A. Matsufuji, Y. Maekawa, T. Miyasaka, *Science* 1997, **276**, 1395–97.

104 J. Yang, M. Wachtler, M. Winter, J.O. Besenhard, *Electrochem. Solid-State Lett.* 1999, **2**, 161–63.

105 M. Noh, Y. Kim, G. Kim, H. Lee, H. Kim, Y. Kwon, Y. Lee, J. Cho, *Chem. Mater.* 2005, **17**, 3320–24.

106 O. Mao, R.L. Turner, I.A. Courtney, B.D. Fredericksen, M.I. Buckett, L.J. Krause, J.R. Dahn, *Electrochem. Solid-State Lett.* 1999, **2**, 3–5.

107 O. Mao, J.R. Dahn, *J. Electrochem. Soc.* 1999, **146**, 423–27.

108 J. Yang, Y. Takeda, N. Imanishi, O. Yamamoto, *J. Electrochem. Soc.* 1999, **146**, 4009–13.

109 L. Shi, H. Li, Z. Wang, X. Huang, L. Chen, *J. Mater. Chem.* 2001, **11**, 1502–05.

110 Y. Wang, J.Y. Lee, B.H. Chen, *J. Electrochem. Soc.* 2004, **151**, A563–70.

111 A. Caballero, J. Morales, L. Sanchez, *Electrochem. Solid-State Lett.* 2005, **8**, A464–466.

112 L. Balan, R. Schneider, D. Billaud, J. Lambert, J. Ghanbaja, *Mater. Lett.* 2005, **59**, 2898–2902.

113 J.H. Kim, H. Kim, H.J. Sohn, *Electrochem. Commun.* 2005, **7**, 557–61.

114 Y.S. Jung, K.T. Lee, J.H. Ryu, D. Im, S.M. Oh, *J. Electrochem. Soc.* 2005, **152**, A1452–57.

115 M. Noh, Y. Kwon, H. Lee, J. Cho, Y. Kim, M.G. Kim, *Chem. Mater.* 2005, **17**, 1926–29.

116 K. Wang, X. He, J. Ren, C. Jiang, C. Wan, *Electrochem. Solid-State Lett.* 2006, **9**, A320–23.

117 K. Wang, X. He, J. Ren, L. Wang, C. Jiang, C. Wan, *Electrochim. Acta* 2006, **52**, 1221–25.

118 H. Li, X. Huang, L. Chen, Z. Wu, Y. Liang, *Electrochem. Solid-State Lett.* 1999, **2**, 547–49.

119 I.S. Kim, P.K. Kumta, *J. Power Sources* 2004, **136**, 145–49.

120 G.X. Wang, J. Yao, H.K. Liu, *Electrochem. Solid-State Lett.* 2004, **7**, A250–53.

121 M. Holzapfel, H. Buqa, F. Krumeich, P. Novak, F.M. Petrat, C. Veit, *Electrochem. Solid-State Lett.* 2005, **8**, A516–20.

122 X. Yang, Z. Wen, X. Zhu, S. Huang, *Electrochem. Solid-State Lett.* 2005, **8**, A1481–83.
123 E. Frackowiak, S. Gautier, H. Gaucher, S. Bonnamy, F. Beguin, *Carbon* 1999, **37**, 61–69.
124 G.T. Wu, C.S. Wang, X.B. Zhang, H.S. Yang, F. Qi, P.M. He, W.Z. Li, *J. Electrochem. Soc.* 1999, **146**, 1696–1701.
125 B. Gao, A. Kleinhammes, X.P. Tang, C. Bower, L. Fleming, Y. Wu, O. Zhou, *Chem. Phys. Lett.* 1999, **307**, 153–57.
126 E. Frackowiak, F. Beguin, *Carbon* 2001, **39**, 937–50.
127 E. Frackowiak, F. Beguin, *Carbon* 2002, **40**, 1775–87.
128 W.X. Chen, J.Y. Lee, Z. Liu, *Electrochem. Commun.* 2002, **4**, 260–65.
129 W.X. Chen, J.Y. Lee, Z. Liu, *Carbon* 2003, **41**, 959–66.
130 T.P. Kumar, R. Ramesh, Y.Y. Lin, G.T K. Fey, *Electrochem. Commun.* 2004, **6**, 520–25.
131 J. Xie, X.B. Zhao, G.S. Cao, M.J. Zhao, *Electrochim. Acta* 2005, **50**, 2725.
132 Z.P. Guo, Z.W. Zhao, H.K. Liu, S.X. Dou, *Carbon* 2005, **43**, 1392–99.
133 J. Yin, M. Wada, Y. Kitano, S. Tanase, O. Kajita, T. Sakai, *J. Electrochem. Soc.* 2005, **152**, A1341–46.
134 H. Huang, W.K. Zhang, X.P. Gan, C. Wang, L. Zhang, *Mater. Lett.* 2007, **61**, 296–99.
135 J.Y. Eom, J.W. Park, H.S. Kwon, S. Rajendran, *J. Electrochem. Soc.* 2006, **153**, A1678–84.
136 Y. Zhang, X.G. Zhang, H.L. Zhang, Z.G. Zhao, F. Li, C. Liu, H.M. Cheng, *Electrochim. Acta* 2006, **51**, 4994–5000.
137 M.R. Wagner, P.R. Raimann, A. Trifonova, K.C. Moeller, J.O. Besenhard, M. Winter, *Electrochem. Solid-State Lett.* 2004, **7**, A201–05.
138 Z. Chen, L. Christensen, J.R. Dahn, *Electrochem. Commun.* 2003, **5**, 919–23.
139 T.D. Hatchard, J.R. Dahn, *J. Electrochem. Soc.* 2004, **151**, A1628–35.
140 A. Timmons, J.R. Dahn, *J. Electrochem. Soc.* 2006, **153**, A1206–10.

6
An Industrial Ecology Perspective

Shannon M. Lloyd, Deanna N. Lekas, and Ketra A. Schmitt

6.1
Introduction

6.1.1
Industrial Ecology

Industrial ecology (IE) is a framework for analyzing the impacts and interactions of industrial, social and ecological systems. White [1] defined IE as 'the study of the flows of materials and energy in industrial and consumer activities, of the effects of these flows on the environment and of the influences of economic, political, regulatory and social factors on the flow, use and transformation of resources'.

IE as a field is fairly young. The intellectual underpinnings of IE can be found in systems analysis research conducted by Forrester [2], research on the flows of materials in economies by Ayres and Kneese [3] and the use of systems analysis to evaluate environmental degradation trends [4, 5]. Two major developments in 1989 are generally cited as founding IE as a discipline. First, Ayres [6] developed the concept of industrial metabolism, which compares processes for converting materials, energy and labor into finished products and waste to living organisms. Second, Frosch and Gallopoulos [7] published 'Strategies for Manufacturing', in which they developed the biological analogy for industrial systems [8].

The primary objectives of IE are to understand how industrial and economic systems behave and interact with ecological and social systems, to transition from open systems to closed-loop systems where waste from one industry can be used an input for another industry and to develop industrial and regulatory strategies that function effectively with natural systems – allowing resources to be replenished and avoiding damage to biological and natural systems. The field of IE encompasses several related areas of research, practice and tools. The following list was identified by the International Society for Industrial Ecology [9].

Nanotechnology. Volume 2: Environmental Aspects. Edited by Harald Krug

ISBN: 978-3-527-31735-6

- material and energy flow studies ('industrial metabolism')
- dematerialization and decarbonization
- technological change and the environment
- life-cycle planning, design and assessment
- design for the environment ('eco-design')
- extended producer responsibility ('product stewardship')
- eco-industrial parks ('industrial symbiosis')
- product-oriented environmental policy
- eco-efficiency.

The conceptual framework provided by IE and the tools listed above can be used retrospectively to evaluate the impacts of current industrial processes, consumer activities and government regulations. However, when retrospective studies identify specific changes for reducing negative ecological impact, it is often difficult to make changes to existing products or practices. For example, decisions made during product development determine what a product will be made of, how it will be produced, where it will be produced, how it will be used and how it will be disposed. Consequently, most of the costs and material, energy and environmental loadings that will be experienced during a product's life cycle are likely to be committed during product development [10–12]. Actual costs and environmental impact are not realized until later in the product life cycle. Changing a product to reduce its environmental impact after the product has been developed can cost orders of magnitude more than making the change during product development [11]. If infrastructures have been built around a commercialized product, it may be difficult to make any changes at all.

Rather than wait until a product is developed and commercialized, IE concepts and tools can be applied prospectively to provide a forward-looking analysis that allows the design and manufacture of products in a manner that prevents or reduces negative environmental impacts and interactions in nature. For example, life cycle engineering (LCE), design for the environment (Df E) and green design approaches are used during product development to estimate the environmental impacts of different product designs and support decision-making aimed at reducing the environmental impact of products [12–14].

6.1.2 Applying Industrial Ecology to Nanotechnology

Control of properties at the nanoscale offers opportunities to use materials and energy more efficiently and reduce waste and pollution. As a result, nanotechnology has the potential to provide more appealing products while also improving environmental performance and sustainability. However, a technological push towards greater investment in nanotechnology without a commensurate consideration of its potential impacts on human health and the environment could lead to new (and sometimes unforeseen) impacts or to cases where the nanotechnology substitute is inferior to the product or process replaced when evaluated over its full life cycle.

Nanotechnology offers a prime opportunity for industrial ecologists to assess an emerging technology in the nascent stages of product development. A holistic industrial ecology perspective can be used to provide a complete picture of the resource requirements and potential human health and environmental impacts associated with the full life cycle of nanotechnology-based goods and services and should reveal products and applications where a precautionary approach may be needed. In this chapter, we describe several IE concepts: life cycle assessment (LCA), material flow analysis (MFA), substance flow analysis (SFA) and corporate social responsibility – and explore their implications and application to nanotechnology.

6.2
Life Cycle Assessment

6.2.1
Background on Life Cycle Assessment

The life cycle stages associated with a product are shown in Figure 6.1. Materials, energy and labor are required to extract, process and transport raw materials and to manufacture, transport, use, dispose of, reuse and recycle products. In addition to consuming resources, the transformation of materials and energy into products results in environmental discharges and generates waste. Evaluating the total

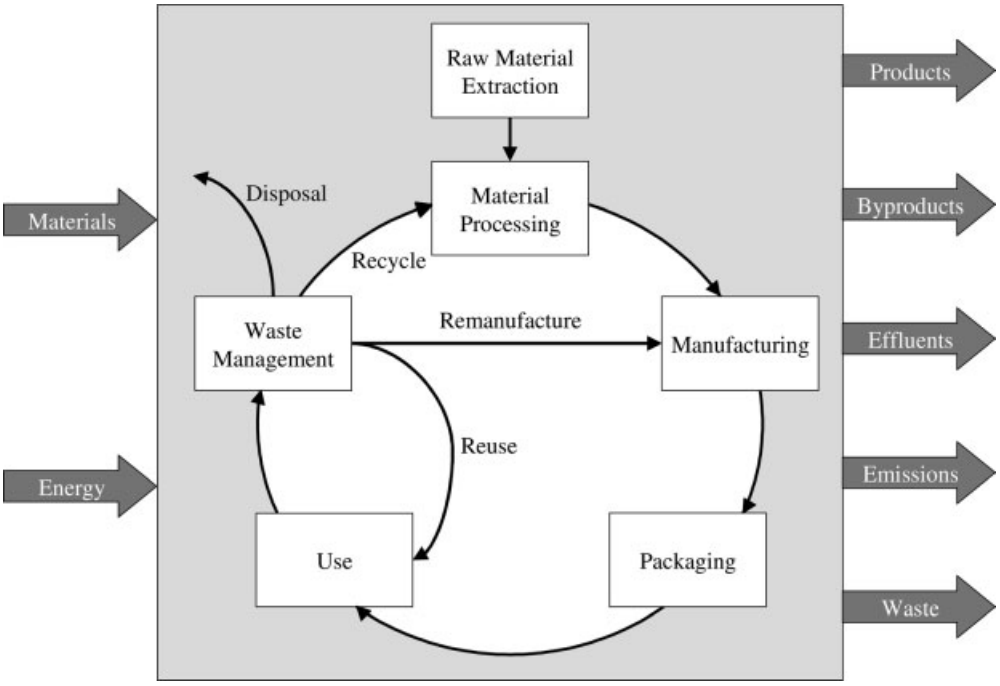


Figure 6.1 Product life cycle stages.

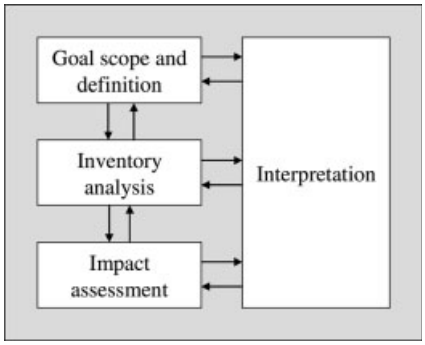


Figure 6.2 ISO 14040 life cycle assessment framework.

environmental impacts of a product requires analysis of material usage, energy usage and environmental discharges associated with each life cycle stage. This makes it possible to identify the most important environmental implications and recommend changes to design specifications, industrial processes or use activities to reduce the overall environmental impact of a product. For example, Lave *et al.* [15] found that mining and smelting lead and making and recycling a battery for early battery-powered vehicles would result in 4–60 times as much lead being discharged into the environment per vehicle mile as a comparable car using leaded gasoline. This insight arose from considering the product's entire life cycle instead of narrowly focusing on the vehicle's zero emissions during vehicle use.

LCA is a systematic, analytical process for assessing the inputs and outputs associated with each life cycle stage for a given product. The International Organization for Standardization (ISO) LCA standards describe the principles and framework for conducting an LCA [16, 17]. These stages are illustrated in Figure 6.2. The goal and scope definition phase of an LCA specifies the reason for conducting the study and identifies system boundaries, data requirements and study limitations. The life cycle inventory (LCI) analysis phase consists of collecting, validating and aggregating input and output data to quantify material use, energy use, environmental discharges and waste associated with each life cycle stage. The life cycle impact assessment (LCIA) phase consists of translating LCI results in potential environmental effects using impact categories, category indicators, characterization models, equivalency factors and weighting values. The interpretation phase occurs iteratively with other LCA phases to assess whether results are in line with defined goals and scope, provide an unbiased summary of the results, define significant impacts and recommend methods for reducing material use, energy use and environmental burdens.

6.2.2
Life Cycle Implications for Nanotechnology

In an early report on the societal implications of nanoscience and nanotechnology, Lave [18] warned that every technology has undesirable consequences and called

for a life cycle perspective in assessing the effects of nanotechnology on environmental quality and sustainability. More recently, reports published by the US National Science and Technology Council [19], the European Commission [20] and the UK Royal Society and Royal Academy of Engineering [21] recommended the use of LCA to evaluate nanotechnology products. In particular, the Royal Society and Royal Academy of Engineering report recommended that a series of LCAs be undertaken for applications arising from existing and expected developments in nanotechnologies, to ensure that the savings in resource consumption during the use of the product are not offset by increased consumption during manufacture and disposal.

Numerous projections of nanotechnology-based products and processes assert that they will use materials and energy more efficiently and reduce waste and pollution. For example, one projection estimates that nanoscale lighting technologies could reduce total global energy consumption by 10%, save $100 billion annually and reduce carbon emissions by 200 million tons per year [22]. Another projection estimates that using lightweight nanocomposites in automotive parts could save 11.5 billion liters of gasoline and reduce carbon dioxide emissions by 5 billion kilograms over the life of one year's fleet of vehicles [23].

These examples focus on the use stage of the product life cycle. In particular, environmental and resource savings are expected from improving use-phase lighting and vehicle efficiency through nanotechnology. However, a complete assessment must include an analysis of material usage, energy usage and environmental discharges associated with the full life cycle of commercialized nanotechnology products and processes. This includes extraction of raw materials, production, use and end-of-life. Table 6.1 lists several possible life cycle benefits and disadvantages of nanotechnology products. The assertions provided in Table 6.1 are generalizations, which will not likely apply to every nanotechnology-based product. Instead, LCA can be used to evaluate the net environmental impact of a specific product.

6.2.3
Life Cycle Studies Conducted to Date

To date, LCA has been used only on a limited basis to assess the potential life cycle impacts of nanotechnology products. Table 6.2 provides a summary of LCAs of nanotechnology-based products conducted to date. Additional studies have been identified by Lekas [24]. For the most part, these studies focused on the life cycle implications associated with reducing the mass of materials incorporated into products and use-phase energy consumption. Evaluation of the impacts associated with processing nanomaterials, manufacturing and retiring nanotechnology products and releasing engineered nanoparticles into the environment are limited by lack of data on producing nanomaterials, recovery of materials from nanotechnology products, fate and transport of engineered nanoparticles and risks from ecological and human exposure to the numerous types of engineered nanoparticles.

There are currently insufficient life cycle inventory data for nanoscale manufacturing processes. LCA studies of nanotechnology-based products conducted to date have relied on surrogate data for these processes. However, recent work has focused on quantifying

Table 6.1 Potential life cycle benefits and issues of nanotechnology.

	Implication	Environmental impact	Life cycle stage(s)
Potential benefits	Improved ability to detect and eliminate pollution	Improved air, water and soil quality	All
	Improved pollution control technology	Improved air, water and soil quality	All
	High-precision manufacturing	Reduced waste	Manufacturing
	Design and control chemistry	Reduced reliance on toxic and scarce materials	Materials processing, manufacturing
	More efficient processes for materials synthesis	Lower energy usage	Materials processing
	More efficient energy production and storage	Lower energy usage	All
Potential risks	Development and release of toxic engineered nanoparticles	Negative impact on human and ecosystem health	All
	Top-down methods with high waste-to-product ratios	Increase in materials required and waste during manufacturing	Materials processing, manufacturing
	High energy requirements for synthesizing nanomaterials	Increased energy usage	Materials processing
	Self-assembly reactions using toxic substances	Increase in toxic releases	Materials processing
	Complex issues with material recovery	Lower recycling rates, increased energy usage for recycling	End-of-life

the inputs and outputs associated with producing nanomaterials. For example, Zhang *et al.* [25] qualitatively examined the energy requirements for several important nanoscale manufacturing technologies. Preliminary findings indicated that bottom-up manufacturing processes are more energy intensive than top-down processes. With bottom-up approaches, nanoscale structures with fundamentally new molecular organization are built by precisely locating individual atoms and molecules where they are needed. With top-down approaches, nanoscale structures are made by reducing the dimensions of a larger structure using machining and etching techniques.

Khanna and Bakshi [26] conducted a cradle-to-gate LCA to evaluate two nanoparticles: nanoclay synthesis from montmorillonite clay and carbon nanofibers via catalytic pyrolysis of hydrocarbons on a metallic catalyst. Rather than re-lying on surrogate life cycle inventory data for nanoscale manufacturing technologies, the inputs and outputs

Table 6.2 Summary of comparative life cycle assessments.

Reference	Application	Conventional	Nanotechnology	Potential impact with nanotechnology
Lloyd and Lave [54]	Automotive body panels	Steel and aluminum	Nanoclay-based nanocomposite	Lighter automobiles, use-phase fuel savings and CO_2 emissions reduction. Perhaps not better than aluminum
Europa [55]	Coatings for magnesium alloys	Chromium-based coatings	Nano-coatings	Unknown
Beaver [56]	Inorganic sunscreens	Bulk titania	Nano-sized titania	Unknown
	Membrane	Sol–gel	Alumoxane nanoparticles	Unknown
Steinfeldt *et al.* [57]	Aluminum coating	Water, solvent and powder varnish	Sol–gel nano-varnish	Less thick coating layer, lower VOC and greenhouse gas emissions during varnish production and use
	Styrene synthesis	Iron oxide catalytic converter	Nanotube-based catalytic converter	Energy savings during synthesis and reduced heavy metals emissions. Risks of carbon nanotubes unknown
	Displays	CRT, LCD and plasma	OLED and CNT	Lower production input requirements, increased use-phase energy efficiency
	Lighting	Conventional and energy-saving bulbs	LEDs	Current LED technology less efficient than energy-saving bulbs
Lloyd *et al.* [58]	Automotive catalysts	Conventional fabrication	Nanofabrication	Reduced PGM loading levels, decreased energy consumption and lower emissions
Fluharty [59]	Copying/printing	Conventional pulverization technique	Emulsion aggregation	Reduced toner usage, emissions and waste and energy use. Increased liquid waste
Jolliet *et al.* [29]	Electronics, computing	Not specified	Not specified	Improved use-phase performance, dematerialization, higher electricity consumption during manufacturing
	Medical applications	Not specified	Not specified	Dominated by use-phase risks and benefits

associated with producing nanomaterials should be quantified and incorporated into future LCAs of nanotechnology-based products.

Efforts have also been initiated to develop a fundamental understanding of the behavior of engineered nanoparticles in natural systems and their influence on biological systems [27, 28]. This understanding should eventually improve the ability to evaluate the fate and transport of nanoparticles and their effects on ecological and human health. While current LCAs of nanotechnology-based products are not able to quantify the potential impacts of nanoparticle releases, future LCAs will incorporate findings from studies such as these to assess the potential impact of ecological and human exposure to nanoparticles during a product's life cycle.

From a life cycle perspective, the use of nanotechnology-based products will not necessarily result in environmental benefits. LCAs conducted to date indicate high potential savings from reducing use-phase energy consumption or in-product material usage and a potential increase in environmental impact from energy-intensive bottom-up nanomanufacturing processes. However, these generalizations are not the case for all nanotechnology-based products. In addition, very little is known about the environmental implications of other life cycle stages, such as the disposal of nanotechnology-based products.

Other efforts focus on developing tools and approaches for supporting LCA and engineering of nanotechnology products. Jolliet *et al.* [29] developed a matrix for identifying the additional risks and benefits of a nano-product's life cycle. DuPont and Environmental Defense have developed a framework to identify and evaluate key factors for the responsible development, production, use and disposal of nanomaterials. One of the steps associated with this framework is to profile the life cycle of the nanotechnology [30]. Using life cycle approaches to evaluate the life cycle environmental implications of alternative courses of action during nanotechnology research and development (R&D) would improve the ability to identify the potential hazards of nanotechnology, evaluate tradeoffs, establish regulatory mechanisms, optimize products for all aspects of life cycle performance and make more strategic R&D choices.

6.3 Substance Flow Analysis

6.3.1 Background on Substance Flow Analysis

While LCA helps us understand the benefits and costs of nanotechnology-based products at different life cycle stages as compared with conventional products, other tools can help us improve understanding better nanomaterials themselves and their implications at different life stages. Material flow analysis (MFA) and substance flow analysis (SFA) are used to evaluate the flows and accumulation of resources within an economy, sector or geographic region. MFA evaluates bulk flows of materials and SFA tracks flows of specific substances. SFA offers a framework for better understanding the amount, spatial location and flow of specific materials throughout the economy.

SFA evaluates the flow of specific materials from cradle to grave, quantifying the input of a substance during production, end-use applications and end-of-life. The context used for such an analysis may vary from the specific (e.g., facility level) to the broad (e.g., world scope). Udo de Haes *et al.* [31] refer to SFA as a tool for analyzing the societal metabolism of substances, whereby materials are exchanged between the anthroposphere with the environment. SFA can be used to identify potential environmental problems by quantifying the flows and accumulation of a specific substance in different environmental systems.

6.3.2
Substance Flow Analysis Implications for Nanotechnology

Faced with many unknowns about nanomaterials and their penetration into our everyday lives, SFA provides an approach for investigating the quantity and location of specific nanomaterials in the economy (e.g., the quantity of carbon nanotubes that are produced and exist in Japan versus the USA). It is less useful to compare quantities of nanomaterials in general because the properties, effectiveness and hazard potential differ by nanomaterial.

Consideration of quantity may also be useful for characterizing the dematerialization or waste reduction that may result when nanomaterials replace conventional or larger materials. For instance, a back-of-the-envelope MFA calculation on lead in cathode-ray tube (CRT) monitors revealed that disposing of one CRT monitor sends as much as 0.45 kg of lead to a landfill; switching to nanotechnology-based substitutes such as flat panel displays with organic light-emitting diodes will reduce this type of lead waste in the future [32, 33]. By finding out where nanomaterials are produced and what they are used for, we can better understand where they will ultimately end up and who will be exposed to nanomaterials during the production, distribution or use of products containing these substances.

SFA can be an appropriate tool when the material of interest is linked to a particular impact and thus warrants a more focused analysis on the stocks and flows and concentrations in the environment [34]. Additionally, by identifying large accumulations of nanomaterials, SFA may also highlight unexpected impact areas.

6.3.3
Summary of Substance Flow Analysis Work Conducted to Date

Few SFAs have been conducted on nanomaterials to date. One of the authors performed a preliminary SFA on carbon nanotubes [35]. Carbon nanotubes were chosen because of the growing interest in, manufacture, and use of these materials combined with increased concerns about their potential risks. For this analysis, carbon nanotube production and use information was gathered from the literature (both journals and news sources) and nanotube company Web sites. Information on nanotube production, raw material inputs and nanotube destination was requested from nanotube producers identified in a *Small Times Magazine* survey [36] and other producers identified during initial research.

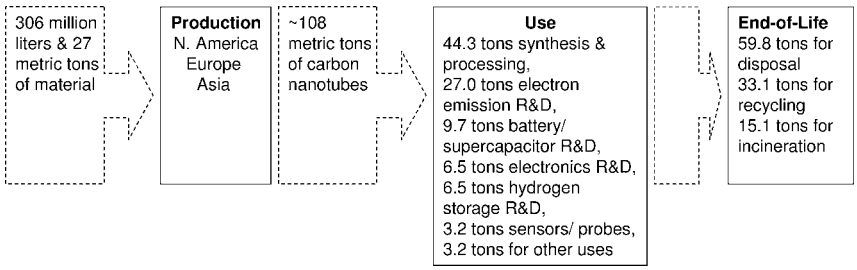


Figure 6.3 Approximate substance flow diagram for carbon nanotubes (2004).

Since nanomaterials are relatively new and information on their production and use is often proprietary, this SFA required estimation and assumptions to characterize production, use and end-of-life flows of carbon nanotubes based on available information. Since the most comprehensive data were available on production, the initial analysis focused on this life cycle stage. Information provided by one firm was used to approximate the inputs and outputs from carbon nanotube growth and then as a basis for characterizing carbon nanotubes at other life stages. The dissipation of carbon nanotubes into various end uses was estimated using carbon nanotube patent filings for eight broad application categories. Sales data or product-specific projections may also provide ways to approximate nanomaterial penetration into end uses. End-of-life outcomes for nanotubes was modeled using surrogate data from existing waste management scenarios published by the US Environmental Protection Agency (EPA) for US municipal solid waste. In reality, different products and applications will likely result in different waste management outcomes. For instance, electronics containing carbon nanotubes may be refurbished and reused, whereas carbon nanotubes in synthesis and processing may be incinerated.

The flow diagram in Figure 6.3 presents a general overview of the SFA findings on carbon nanotubes. Estimates presented here are generalized based on limited information, a small amount of nanotubes in commercial applications and a rapidly changing market. However, the findings improve our understanding of carbon nanotube flows throughout the economy. SFA provides a snapshot of a substance's flow in the economy at a given time, providing a better understanding of where these materials exist and are expected to go. Widespread projected application of these nanomaterials in many everyday products (from vehicle composites to tennis racquets and batteries) amidst uncertain risks makes the characterization of the flow of nanotubes increasingly important.

6.4
Corporate Social Responsibility

6.4.1
Background on Corporate Social Responsibility

Corporate Social Responsibility (CSR) is broadly defined as behavior by corporate entities that consider social and environmental impacts concurrently with the

economic well-being of the corporation. CSR is believed by many to pay dividends in terms of improved recruiting, branding and consumer loyalty and also through decreased risk of litigation and regulation. The following practices are often adopted by CSR-minded companies:

- Include stakeholders in decision-making.
- Maintain transparency with corporate activities affecting workers of the community.
- Seek to protect workers and their community through limiting environmental and human health impacts of products during manufacture, use and disposal.
- Use materials, labor and locations sustainably.
- Participate in CSR certification and reporting programs such as ISO 14000.

6.4.2
Corporate Social Responsibility Implications for Nanotechnology

Small Times Magazine estimated that over 3500 firms were involved in some aspect related to micro- or nanotechnology [37]. This includes companies that research and manufacture nanomaterials, nanotechnology-based products and equipment for nano-production, trade magazines and firms that offer legal expertise, venture capital investment and intellectual property strategizing on nanotechnology. Nanomaterials are used in numerous sectors, including energy, cosmetics, medicine and electronics.

Nanomaterials offer significant opportunities for decreased environmental impacts in terms of materials and energy consumption, and also the possibility of targeted human health benefits in the future. Along with the potential benefits also comes the potential for uncertain risks. Uncertainty in the risks of nanotechnology stem from scientific uncertainty in several areas, such as transport and dispersion of nanomaterials; ability of nanomaterials to enter biological systems; potential mode of action and toxicological effects of nanomaterials; suitability of traditional worker protection techniques in nano-environments; and the ability of the life cycle benefits of specific nanotechnology applications to outweigh the life cycle costs.

The level of uncertainty surrounding nanotechnology makes the work of protecting the public, workers and the environment and communicating about potential risks and benefits challenging for CSR-committed corporations. How can a corporate entity prevent or reduce exposures when the mode of exposure is not understood? How can a corporation communicate a complex, uncertain message to the public? Given the potential advantages and risks offered by nanomaterials, what is the appropriate path forward? These challenges also represent a significant opportunity for CSR-oriented firms to collaborate with regulators, non-governmental organisations (NGOs), and other firms to achieve common corporate social responsibility. By being proactive in seeking partners, sharing information and conducting research, nanofirms can hope to reduce risk and obtain competitive advantage.

Increased public awareness of the potential risks of nanotechnology, the lack of sound strategies for identifying and assessing these risks and increased production and use of nanomaterials prompted NGOs, activist groups and members of the scientific community to call for more research investigating nanotechnology's

risks [38–41]. For example, a report prepared for Greenpeace called for in-depth assessments of the environmental risks of near-term nanotechnology – specifically products or processes that might result in the release of nanoparticles into the environment [40]. The Canadian-based NGO ETC Group [38, 39] called for a global moratorium on the commercial production of nanomaterials. In particular, ETC Group [39] cautioned against scaling up of nanomaterial production without understanding the potential adverse side-effects from using them in many diverse commercial applications. Similar campaigns launched by activist groups have resulted in backlash against genetically modified (GM) crops. In the light of GM setbacks and a responsibility to balance pursuit of science with sensible precaution, editorials in *Nature* warned against leaving legitimate questions unanswered [41, 42]. Instead, they called for more research into nanotechnology risks accompanied by an open public discussion.

Companies face potential financial risks from the investment and insurance communities who are paying greater attention to the environmental, health and safety (EHS) discussions surrounding nanotechnology. Insurers are encouraging more rigorous review on both the opportunities and hazards of nanotechnology. A report by the Swiss Reinsurance Company asserts that the insurance industry should analyze nanotechnology to identify potential risks [43]. MunichRe, GenRE and Allianz have also published reports on the need for nanotechnology risk assessment [44]. Cientifica urges companies working with nanomaterials to prepare for the 'potential impact of future liabilities or changes in legislation' [53].

Companies that take the lead in dealing with EHS management and public perception concerns with nanotechnology (i.e. are 'first movers' in their industry group) may attain competitive advantage. They will not only be ready for possible regulation and may avoid retrofitting operations later (in the event of future nano-specific regulations), but they may also appeal to consumers who value CSR efforts and be in the position to influence regulatory or reporting schemes. For example, the CEO of NanoDynamics testified before Congress [45] and DuPont and Environmental Defense worked together to develop a framework for the responsible development, production, use and disposal of nanomaterials and products [30, 46]. By taking measures to consider and evaluate environmental, health and safety risks that nanomaterials may pose, corporations may avoid future risks, backlash and liability.

6.4.3
Summary of Work Conducted to Understand Nanofirm EHS Concerns and Actions

Several corporations have adopted environmental management strategies to identify, prevent or reduce the risks of nanotechnology. For example, ApNano Materials Inc. followed the European Commission's *Good Laboratory Practices Directives* to assess toxicity of its nano-based lubricant [44]. The Carbon Nanotech Research Institute adopted a policy against making products available in a powdered form in order to avoid any dispersion risks [47].

Various studies have sought to understand awareness of EHS issues at nanotechnology firms and to identify what firms are doing to deal with potential and uncertain

risks and what they could be doing to protect their workers, the environment and consumers. A few examples include the following:

- Lux Research [48] interviewed nanotechnology firms and found that companies want more certainty in the type of regulations to expect. In order to handle real, perceived and regulatory risks, Lux Research suggested that firms inventory nanomaterials, map them to exposures throughout the life cycle, characterize the risks with available knowledge and mitigate them with appropriate controls, toxicity testing and product redesigns.
- Innovest [44] reviewed hundreds of public and private nanotech companies to develop a list of best practices for offsetting potential perception risk based on product strategy, risk and stewardship.
- Lekas [49] surveyed nanotechnology startup firms in Connecticut and New York and found that they have varied concerns and degrees of progress in addressing EHS issues; they indicated that they need information and guidance and prefer communication about nanotechnology risks through an electronic or online venue from a government source.
- Lindberg and Quinn [50] surveyed nanotechnology firms in the northeastern USA and learned that small firms, in particular, need a roadmap from suppliers, industry, and government bodies to manage risks.
- A research team at the University of California, Santa Barbara [60] conducted an international survey of nanotechnology workplace safety practices for the International Council on Nanotechnology (ICON). The results of their interviews and analysis revealed that companies and laboratories recognize nanomaterial risks, but are following conventional practices faced with a lack of information.
- A European Commission survey of 380 European nanotechnology startups indicated that they do not consider social acceptance and environmental and health regulations as important barriers for the applications of nanomaterials [51].
- The European NanoBusiness Association [52] surveyed 142 European businesses (of which 18% were startups) and found that most respondents believed that health and environmental impacts of nanotechnology needed to be studied.

6.5 Conclusions

Scientific understanding of the human health and environmental impacts of nanotechnology and actions to regulate nanotechnology lag behind the research, development and commercialization of nanotechnology-based products. IE provides a framework for analyzing the impacts of nanotechnology on social and ecological systems. A number of IE tools, including LCA, MFA, SFA and CSR, can aid in assessing the potential risks and benefits of nanotechnology, in addition to developing

methods to reduce or mitigate potential risks. When applied prospectively, IE tools can be used to provide a forward-looking analysis that permits the design and manufacture of nanotechnology-based products in a manner that prevents or reduces their negative impacts. The current level of uncertainty surrounding the potential impacts and risks of nanotechnology present a significant challenge in applying IE tools. As scientific understanding of nanotechnology's environmental and health impacts evolves, the efficacy of IE in mitigating potential harm and designing more effective products using nanotechnology will continue to increase.

References

1 White, R. (1994) in *The Greening of Industrial Ecosystems. Preface*, (eds B. Allenby and D. Richards), The National Academy of Engineering, National Academy Press, Washington, DC, v–vi.

2 Forrester, J.W. (1968) *Principles of Systems*, Wright-Allen Press, Cambridge, MA.

3 Ayres, R.U. and Kneese, A.V. (1969) Production, consumption and externalities. *American Economic Review*, **59** (3), 282–297.

4 Meadows, D., Randers, J. and Meadows, D. (1972) *Limits to Growth*, Universe Books, New York.

5 Garner, A., and Keoleian, G.A. (1995) *Industrial Ecology: an Introduction*, National Pollution Prevention Center for Higher Education, University of Michigan, Ann Arbor, MI.

6 Ayres, R. (1989) Industrial metabolism. in *Technology and Environment*, (eds J.H. Ausubel and Sladovich), National Academy Press, Washington, DC, 23–49.

7 Frosch, R.A. and Gallopoulos, N.E. (1989) Strategies for manufacturing. *Scientific American*, **189** (3), 144–152.

8 Ehrenfeld, J. (2002) Industrial ecology: coming of age. *Environmental Science and Technology*, **36** (13), 280A–285A.

9 ISIE (2006) A History of Industrial Ecology. International Society for Industrial Ecology. http://www.is4ie.org/history.htm [Accessed 14 October 2006].

10 Ullmann, D.G. (1992) *The Mechanical Design Process*, McGraw-Hill, New York.

11 Mueller, D.G., Court, A.W. and Besant, C.B. (1999) Energy life cycle design: a method. *Proceedings of the Institution of Mechanical Engineers*, **213B**, 415–419.

12 Gediga, J., Florin, H. and Eyerer, P. (2002) Life cycle engineering: a tool for optimizing technologies, parts and systems. in *Mechanical Life Cycle Handbook*, (ed. M.S. Hundal), Marcel Dekker, New York.

13 OTA (1992) Green Products by Design: Choices for a Cleaner Environment. OTA-E-541. US Congress, Office of Technology Assessment, Washington, DC

14 Graedel, T.E. and Allenby, B.R. (1996) *Design for Environment*, Prentice Hall, Upper Saddle River, NJ.

15 Lave, L.B., Hendrickson, C.T. and McMichael, F.C. (1995) Environmental implications of electric cars. *Science*, **268** (5213), 992–995.

16 ISO (2006) Environmental Management – Life Cycle Assessment – Principles and Framework. International Organization for Standardization, ISO 14040:2006.

17 ISO (2006) Environmental Management – Life Cycle Assessment – Requirements and Guidelines. International Organization for Standardization, ISO 14044:2006.

18 Lave, L.B. (2001) Lifecycle/sustainability implications of nanotechnology. in *Societal Implications of Nanoscience and Nanotechnology*, (eds M.C. Roco and W.S.

Bainbridge), National Science Foundation, Arlington, VA, 162–168.

19 NSTC. (2004) Nanotechnology Grand Challenge in the Environment: Research Planning Workshop Report. National Science and Technology Council, Committee on Technology, Subcommittee on Nanoscale Science, Engineering and Technology, Arlington, VA.

20 Commission of the European Communities (2004) Communication from the Commission: Towards a European Strategy for Nanotechnology. Report No. COM (2004) 338 Final. Commission of the European Communities, Brussels.

21 The Royal Society and The Royal Academy of Engineering (2004) *Nanoscience and Nanotechnologies: Opportunities and Uncertainties*. The Royal Society and The Royal Academy of Engineering, London.

22 NSTC (2000) National Nanotechnology Initiative: the Initiative and its Implementation Plan. National Science and Technology Council, Committee on Technology, Subcommittee on Nanoscale Science, Engineering and Technology, Arlington, VA.

23 NIST (1997) Nanocomposites New Low-cost, High-strength Materials for Automotive Parts. ATP Project Brief 97-02-0047. National Institute of Standards and Technology, Arlington, VA.

24 Lekas, D. (2005) Analysis of Nanotechnology from an Industrial Ecology Perspective. Part I: Inventory and Evaluation of Life Cycle Assessments of Nanotechnologies. Independent project conducted at Yale's School of Forestry and Environmental Studies. November 2005.

25 Zhang, T.W., Boyd, S., Vijayaraghavan, A. and Dornfield, D. (2006) Energy use in nanoscale manufacturing. Proceedings of the 2006 IEEE International Symposium on Electronics and Environment, May, 266–271.

26 Khanna, V. and Bakshi, B.R. (2006) Towards a systems view in nanotechnology – life cycle assessment of nanoparticle synthesis, presented at the 2006 American Institute of Chemical Engineers Annual Meeting, 15 November.

27 Holsapple, M.P. (2005) Forum series: research strategies for safety evaluation of nanomaterials. *Toxicological Sciences*, **87** (2), 315.

28 Borm, P.J.A., Robbins, D., Haubold, S., Kuhlbusch, T., Fissan, H., Donaldson, K., Schins, R., Stone, V., Kreyling, W., Lademann, J., Krutmann, J., Warheit, D. and Oberdorster, E. (2006) The potential risks of nanomaterials: a review carried out for ECETOC. *Particle and Fibre Toxicology*, **3**, 11. http://www.particleandfibreto- xicology.com/content/3/1/11.

29 Jolliet, O., Wenger, Y. and Philbert, M. (2006) Environmental life cycle risks and benefits of nantoechnologies, presented at the Nano Science and Technology Institute Nanotechnology Conference and Trade Show, 7–11 May, Boston, MA.

30 Environmental Defense and DuPont . (2007) NANO Risk Framwork. Environmental Defense–DuPont Partnership, June. http://nanoriskframework.com.

31 Udo de Haes, H., Heijungs, R., Huppes, G., van der Voet, E. and Hettelingh, J. (2000) Full mode and attribution mode in environmental analysis. *Journal of Industrial Ecology*, **4** (1), 45–56.

32 Karn, B. (2006) A Proactive Environmental Perspective on Nanotechnology, presentation given at the US EPA Office of Research and Development and Woodrow Wilson International Center for Scholars/Project on Emerging Technologies at ONAMI, Portland, OR, 6 March.

33 EPA (2007) Nanotechnology White Paper. Prepared for the US Environmental Protection Agency by members of the Nanotechnology Workgroup, a group of EPA's Science Policy Council, US Environmental Protection Agency, Washington, DC. EPA 100 (B-07) 001, February 2007.

34 Bringezu, S., Schutz, H. and Moll, S. (2003) Rationale for and interpretation of economy-wide materials flow analysis and derived indicators. *Journal of Industrial Ecology*, **7** (2), 43–64.

35 Lekas, D. (2005) Analysis of Nanotechnology from an Industrial Ecology Perspective. Part II: Substance Flow Analysis Study of Carbon Nanotubes. Independent project conducted at Yale's School of Forestry and Environmental Studies. November 2005.

36 Small Times Survey (2004) 2004 National Small Tech Commercialization Survey, presented at NanoCommerce 2004, September.

37 Small Times (December 2005) 2006 Small Tech Business Directory. *Small Times Magazine*, **5** (9).

38 ETC Group (2002) No Small Matter! Nanotech Particles Penetrate Living Cells and Accumulate in Animal Organs. ETC Group Communique, May/June.

39 ETC Group (2003) No Small Matter II: the Case for a Global Moratorium: Size Matters! ETC Group Occasional Paper Series, April.

40 Arnall, A.H. (2003) Future Technologies, Today's Choices, Greenpeace Environmental Trust, July.

41 Editorial (2003) Don't believe the hype. *Nature*, **424** (6946), 237.

42 Brumfiel, G. (2003) A little knowledge . . .. *Nature*, **424** (6946), 246–248.

43 Swiss Re (2004) *Nanotechnology: Small Matter, Many Unknowns*, Swiss Reinsurance Company, Zurich.

44 Innovest (2005) Nanotechnology: Non-traditional Methods for Valuation of Nanotechnology Producers. *Introducing the Innovest Nanotechnology Index*, Innovest Strategic Value Advisors, New York, 29 August.

45 NanotechWire (2005) NanoDynamics CEO Address the US House Committee Today. 16 November. http://nanotechwire.com/news.asp?nid=2588.

46 ED (2005) Environmental Defense and DuPont: Global Nanotechnology Standards of Care Partnership. Environmental Defense. 11 October. http://www.environmentaldefense.org/article.cfm?contentID=4821.

47 Rawstern, R. (2004) Nanotubes and Buckyballs. NanoNews – Now. Issue 11. May. http://www.nanotech-now.com/products/nanonewsnow/issues/011/011.htm.

48 Lux Research (2006) Taking Action on Nanotech Environmental, Health and Safety Risks. NTS-R-P 06-003. Lux Research. May.

49 Lekas, D., Lifset, R. and Rejeski, D. (2006) Nanotech Startup Concerns, Information Needs and Opportunities to Proactively Address Environmental, Health and Social Issues: Focus on Firms in Connecticut and New York. Master's Project. Yale School of Forestry and Environmental Studies. July.

50 Lindberg, J. and Quinn, M. (2007) A Survey of Environmental, Health and Safety Risk Management Information Needs and Practices among Nanotechnology Firms in the Massachusetts Region, Project on Emerging Nanotechnologies, Woodrow Wilson International Center for Scholars, December.

51 European Commission (2005) European Survey on Success Factors, Barriers and Needs for the Industrial Uptake of Nanomaterials in SMEs. Report funded by European Commission, NanoroadSME, Sixth Framework Programme. July.

52 European NanoBusiness Association (2005) The 2005 European NanoBusiness Survey, The European NanoBusiness Association.

53 Cientifica (2005) Nanotechnologies: Risks and Rewards. Cientifica White Paper. June.

54 Lloyd, S.M. and Lave, L.B. (2003) Life cycle economic and environmental implications of using nanocomposites in automobiles. *Environmental Science and Technology*, **37** (15), 3458–3466.

55 Europa (2003) Nanocoatings lighten environmental burden of mobility.

http://ec.europa.eu/research/industrial_technologies/articles/article_346_en.html [Accessed 9 October 2006].

56 Beaver, E. (2004) Implications of Nanomaterials Manufacture and Use, presentation given at US EPA Nanotechnology STAR Progress Review Workshop.

57 Steinfeldt, M., Petschow, U., Haum, R. and von Gleich, A. (2004) Nanotechnology and Sustainability. Institute for Ecological Economy Research, Discussion Paper IOEW 65/04, October.

58 Lloyd, S.M., Lave L.B. and Matthews, H.S. (2005) Life cycle benefits of using nanotechnology to stabilize PGM particles in automotive catalysts. *Environmental Science and Technology*, **39** (5), 1384–1392.

59 Fluharty, A. (2005) A Comparison of Conventional Toner to Chemically Produced Toner using Life Cycle Assessment. National Science Foundation, Research Experiences for Teachers, Final Report.

60 Gerritzen, G., Huang, L., Killpack, K., Mircheva, M. and Couti, J. (2006) A Survey of Current Practices in the Nanotechnology Workplace. Produced for the International Council on Nanotechnology by the University of California, Santa Barbara, November. http://icon.rice.edu/projects.cfm?doc_id=4388.

7
Composition, Transformation and Effects of Nanoparticles in the Atmosphere

Ulrich Pöschl

7.1
Introduction

The effects of airborne particles in the nanometer to micrometer size range (aerosols) on the atmosphere, climate and public health are among the central topics in current environmental research [1, 2]. The particles scatter and absorb solar and terrestrial radiation, they are involved in the formation of clouds and precipitation as cloud condensation and ice nuclei and they affect the abundance and distribution of atmospheric trace gases by heterogeneous chemical reactions and other multiphase processes [3–6]. Moreover, they play an important role in the spread of biological organisms, reproductive materials and pathogens (pollen, bacteria, spores, viruses, etc.) and they can cause or enhance respiratory, cardiovascular, infectious and allergic diseases [3, 7–9].

An aerosol is generally defined as a suspension of liquid or solid particles in a gas, with particle diameters in the range 10^{-9}–10^{-4} m (lower limit, molecules and molecular clusters; upper limit, rapid sedimentation) [6, 9]. The most evident examples of aerosols in the atmosphere are clouds, which consist primarily of condensed water with particle diameters on the order of $\sim$10 µm. In atmospheric science, however, the term aerosol traditionally refers to suspended particles which contain a large proportion of condensed matter other than water, whereas clouds are considered as separate phenomena [10].

Atmospheric aerosol particles originate from a wide variety of natural and anthropogenic sources. Primary particles are directly emitted as liquids or solids from sources such as biomass burning, incomplete combustion of fossil fuels, volcanic eruptions and wind-driven or traffic-related suspension of road, soil and mineral dust, sea salt and biological materials (plant fragments, microorganisms, pollen, etc.). Secondary particles, on the other hand, are formed by gas-to-particle conversion in the atmosphere (new particle formation by nucleation and condensation of gaseous precursors). As illustrated in Figure 7.1, airborne particles undergo various physical and chemical interactions and transformations (atmospheric aging),

Nanotechnology. Volume 2: Environmental Aspects. Edited by Harald Krug

ISBN: 978-3-527-31735-6

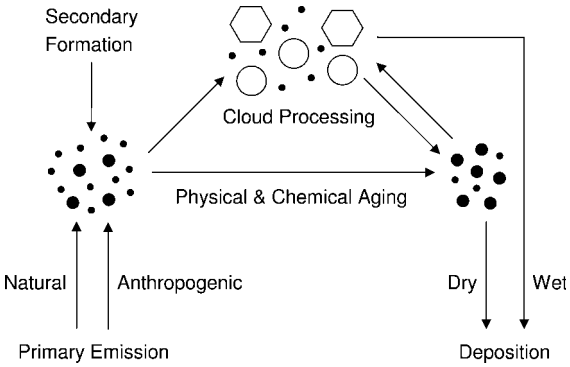


Figure 7.1 Atmospheric cycling of airborne nano- and microparticles (aerosols) [1, 2].

i.e. changes of particle size, structure and composition (coagulation, restructuring, gas uptake, chemical reaction). Particularly efficient particle aging occurs in clouds, which are formed by condensation of water vapor on pre-existing aerosol particles (cloud condensation and ice nuclei, CCN/IN). Most clouds re-evaporate and modified aerosol particles are again released from the evaporating cloud droplets or ice crystals (cloud processing). If, however, the cloud particles form precipitation which reaches the Earth's surface, not only the condensation nuclei but also other aerosol particles are scavenged on the way to the surface and removed from the atmosphere, which is actually the main sink of atmospheric aerosol particles ("wet deposition"). Particle deposition without precipitation of hydrometeors (airborne water particles), i.e. "dry deposition" by convective transport, diffusion and adhesion to the Earth's surface, is less important on a global scale, but it is highly relevant for local air quality, health effects (inhalation and deposition in the human respiratory tract) and the soiling of buildings and cultural monuments. Depending on aerosol properties and meteorological conditions, the characteristic residence times (lifetimes) of aerosol particles in the atmosphere range from hours to weeks [11, 12].

The lifetime of small nanoparticles with diameters on the order of ~10 nm or less is limited by rapid diffusion and coagulation with each other or with larger particles. Under highly polluted conditions (e.g. in exhaust plumes), it can be as short as a few seconds to minutes and in general it rarely exceeds a few hours. In contrast, large nanoparticles with diameters on the order of ~100 nm typically have the longest lifetime of all atmospheric particles. They are too small for efficient sedimentation and too large for efficient diffusion and coagulation. The only efficient removal process for these particles is wet deposition on average time scales of days to weeks.

The concentration, composition and size distribution of atmospheric aerosol particles are temporally and spatially highly variable. In the lower atmosphere (troposphere), the total particle number and mass concentrations typically vary in the approximate ranges 10^2–$10^5\,cm^{-3}$ and 1–100 $\mu g\,m^{-3}$, respectively [11–14]. In general, the predominant chemical components of air particulate matter (PM) are sulfate, nitrate, ammonium, sea salt, mineral dust, organics and black or elemental carbon, each of them typically contributing about 10–30% of the overall mass loading.

For different locations, times, meteorological conditions and particle size fractions, however, the relative abundance of different chemical components can vary by an order of magnitude or more [3, 6, 11, 15]. In atmospheric research, the term "fine air particulate matter" is usually restricted to particles with aerodynamic diameters $\leq 1\,\mu m$ or $2.5\,\mu m$ (PM_1 or $PM_{2.5}$, respectively). In air pollution control, it sometimes also includes larger particles up to $10\,\mu m$ (PM_{10}).

The total number concentration of aerosol particles in the atmosphere is usually dominated by nanoparticles with diameters up to ~100 nm, whereas the total mass concentration is generally dominated by particles with diameters >100 nm (microparticles). Characteristic examples of particle number concentration, size distribution and chemical composition of fine particulate matter in urban and high alpine air are illustrated in Figure 7.2. The displayed particle number size distributions (particle number concentration per logarithmic decade of particle diameter, $dN/d\log d_p$, plotted against particle diameter) have been observed in the city of Munich [500 m above sea level (asl); 8–14 December 2002] and at the Schneefernerhaus research station on Mount Zugspitze (2600 m asl; 6 November 2002) in Southern Germany. They correspond to total particle number concentrations of about $10^2\,cm^{-3}$ in alpine air and $10^4\,cm^{-3}$ in urban air and to particle mass concentrations of about 1

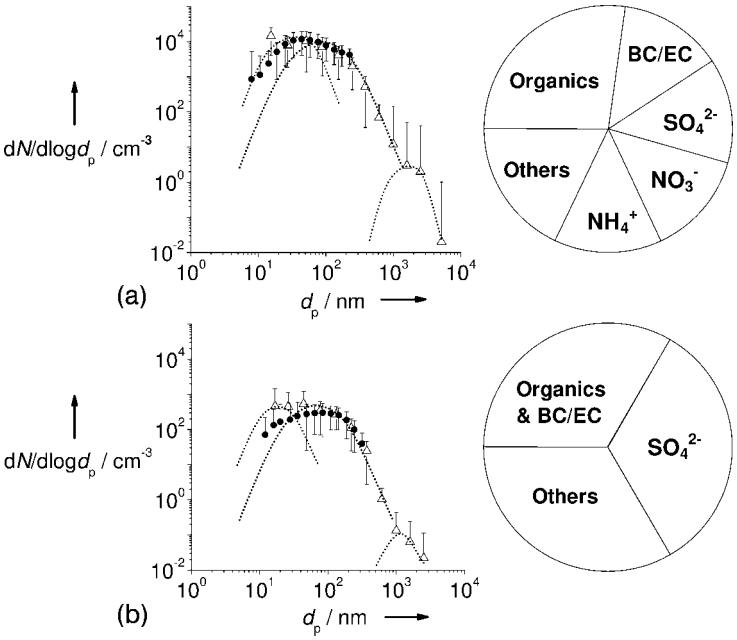


Figure 7.2 Characteristic examples of aerosol particle size distribution and chemical composition in urban (a) and high alpine air (b). Diagrams: number size distribution function $dN/d\log d_p$ (symbols and error bar, arithmetic mean values and standard deviations; open triangles, ELPI; full circles, SMPS; dotted lines, characteristic particle size modes); pie charts, typical mass proportions of main components [1].

and $10\,\mu g\,m^{-3}$, respectively. The measurements were performed with two complementary techniques, an electrical low pressure impactor (ELPI, 10 nm–10 µm, flow rate $30\,L\,min^{-1}$, measurement interval 1 min) and a scanning mobility particle sizer (SMPS, 10–300 nm, flow rate $1\,L\,min^{-1}$, measurement interval 30 min) [16, 17]. The deviations at very low particle size can be attributed to wall losses by diffusion in the SMPS system.

The dotted lines indicate characteristic particle size modes, which can be attributed to different sources, sinks and aging processes of atmospheric particles: nucleation (Aitken), accumulation and coarse mode [6, 9]. In corresponding mass size distributions, which are obtained by multiplication with particle volume ($d_p^3\pi/6$) and density (typically around $2\,g\,cm^{-3}$), the nucleation mode is usually negligible, whereas accommodation and coarse particle modes are of comparable magnitude. The composition pie charts are based on chemical analyses of $PM_{2.5}$ filter samples from the same locations and literature data for urban and remote continental background air [3, 6, 11, 15, 18–21].

Figure 7.3 illustrates the interdependencies between composition, composition-dependent properties, atmospheric interactions and transformation, climate and health effects and sources of aerosols. The resulting feedback loops are of central importance in the science and policy of environmental pollution and global change. Thus a comprehensive characterization (climatology) and mechanistic understanding of particle sources, properties and transformation is required for the quantitative assessment, reliable prediction and efficient control of natural and anthropogenic aerosol effects on climate and public health.

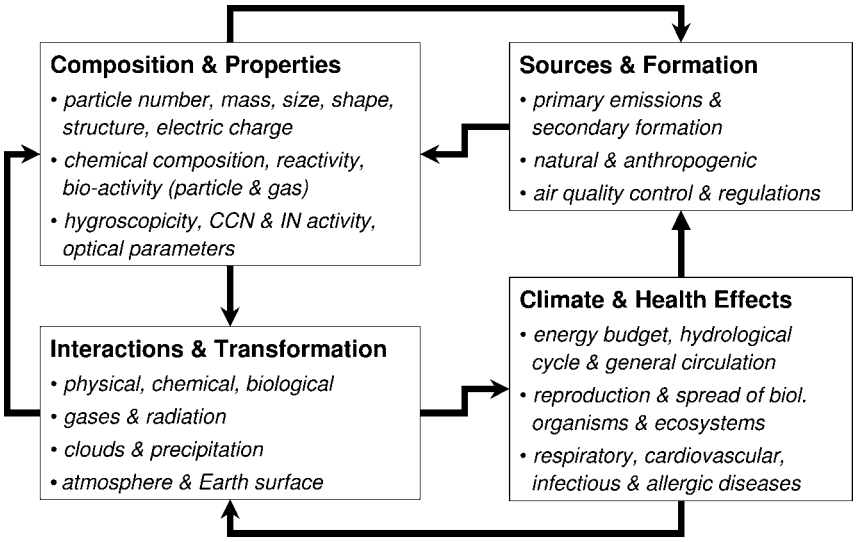


Figure 7.3 Interdependencies and feedbacks between atmospheric aerosol composition, properties, interactions and transformation, climate and health effects and sources [2].

7.2
Composition

Accurate determination of the chemical composition of air particulate matter is a formidable analytical task [1]. Minute sample amounts are usually composed of several main constituents and hundreds of minor and trace constituents. Moreover, the composition of the individual particles can be fairly uniform (internally mixed aerosols) or very different from the ensemble composition (externally mixed aerosols), depending on the involved particle sources and atmospheric aging processes (coagulation, gas–particle partitioning, chemical reactions). Especially in populated environments, air particulate matter can be pictured as the result of an "exploded pharmacy", comprising just about any non- or semi-volatile chemical compound occurring in the biosphere, hydrosphere and lithosphere or released by human activity. In addition to primary chemical components, which are directly emitted by natural and anthropogenic sources, air particulate matter mostly also contains secondary chemical components, which are formed by gas-phase reactions and subsequent gas-to-particle conversion or by chemical transformation of primary particle components in the atmosphere. By definition, an aerosol is composed of particulate and gas-phase components, i.e. the term aerosol component can refer to chemical compounds in both the condensed and gaseous state. In practice and in the remainder of this chapter, however, the term aerosol component usually refers to semi- and non-volatile particle components but not to volatile compounds residing almost exclusively in the gas phase.

7.2.1
Carbonaceous Components

Carbonaceous aerosol components (organic compounds and black or elemental carbon) account for a large fraction of air particulate matter, exhibit a wide range of molecular structures and have a strong influence on the physicochemical, biological and climate- and health-related properties and effects of atmospheric aerosols [1–3, 6, 21–25].

Traditionally, the total carbon (TC) content of air particulate matter is defined as the sum of all carbon contained in the particles, except in the form of inorganic carbonates. TC is usually determined by thermochemical oxidation and evolved gas analysis (CO_2 detection) and divided into an organic carbon (OC) fraction and a black carbon (BC) or elemental carbon (EC) fraction. Measurements of BC and EC are generally based on optical and thermochemical techniques and OC is operationally defined as the difference between TC and BC or EC (TC = BC + OC or TC = EC + OC) [21]. As illustrated in Figure 7.4, however, there is no real sharp cut but a continuous decrease in thermochemical refractiveness and specific optical absorption going from graphite-like structures to non-refractive and colorless organic compounds [26]. Both BC and EC, comprise the carbon content of the graphite-like material usually contained in soot (technically defined as the black product of incomplete hydrocarbon combustion or pyrolysis) and other combustion aerosol

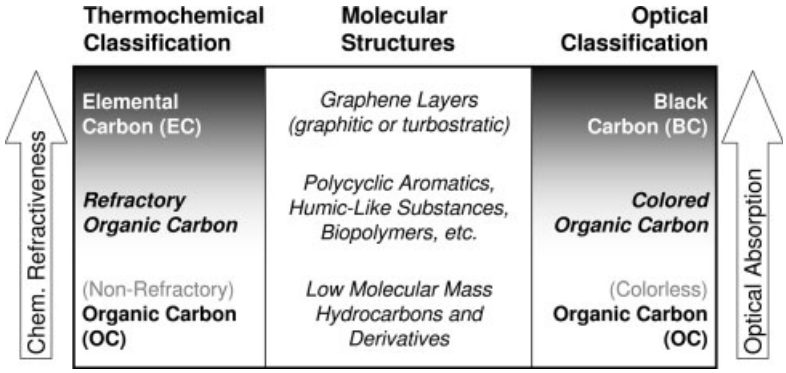


Figure 7.4 Optical and thermochemical classification and molecular structures of black carbon (BC), elemental carbon (EC) and organic carbon (OC = TC − BC or TC − EC) [26]. Depending on the method of analysis, different amounts of carbon from refractory and colored organic compounds are included in BC, EC or OC [1].

particles, which can be pictured as more or less disordered stacks of graphene layers or large polycyclic aromatics [27–29]. Depending on the applied optical or thermochemical methods (absorption wavelength, temperature gradient, etc.), however, BC and EC measurements also include the carbon content of colored and refractory organic compounds, which can lead to substantially different results and strongly limits the comparability and suitability of BC, EC and OC data for the determination of mass balances and physicochemical properties of air particulate matter.

Nevertheless, most information available on the abundance, properties and effects of carbonaceous aerosol components so far is based on measurement data of TC, BC/EC and OC [21, 25]. These data are now increasingly complemented by measurements of water-soluble organic carbon (WSOC), its macromolecular fraction (MWSOC) and individual organic compounds as detailed below. Moreover, the combination of thermochemical oxidation with ^{14}C isotope analysis (radiocarbon determination in evolved CO_2 by accelerator mass spectrometry) allows one to distinguish fossil fuel combustion and other sources of carbonaceous aerosol components. Recent results confirm that the EC is dominated by fossil fuel combustion and indicate highly variable anthropogenic and biogenic sources and proportions of OC [30].

Characteristic mass concentrations and concentration ratios of fine air particulate matter ($PM_{2.5}$) and carbonaceous fractions in urban, rural and alpine air in central Europe are summarized in Table 7.1. The reported data were obtained on an altitude transect through Southern Germany, from the city of Munich (500 m asl), via the meteorological observatory Hohenpeissenberg (1000 m asl), to the environmental research station Schneefernerhaus on Mount Zugspitze (2600 m asl) throughout the period 2001–2003. The sampling locations and measurement procedures have been described in detail elsewhere [31, 32] and the results are consistent with those of other studies performed at comparable locations [11, 13, 15, 18–21].

Table 7.1 Characteristic mass concentrations of fine particulate matter ($PM_{2.5}$) and proportions of total carbon (TC), elemental carbon (EC), organic carbon (OC), water-soluble OC (WSOC) and macromolecular WSOC (MWSOC, molecular mass >5 kDa) mass in urban, rural and high alpine air in central Europe (rounded arithmetic mean values ± standard deviation of about 30 filter samples collected at each location over the period 2001–2003).

	Urban (Munich)	Rural (Hohenpeissenberg)	Alpine (Zugspitze)
$PM_{2.5}$ ($(g\,m^{-3})$)	20 ± 10	10 ± 5	4 ± 2
TC in $PM_{2.5}$ (%)	40 ± 20	30 ± 10	20 ± 10
EC in TC (%)	50 ± 20	30 ± 10	30 ± 10
OC in TC (%)	40 ± 20	70 ± 10	70 ± 10
WSOC in TC (%)	20 ± 10	40 ± 20	60 ± 20
MWSOC in WSOC (%)	30 ± 10	50 ± 20	40 ± 20

On average, the total $PM_{2.5}$ mass concentration decreases by about a factor of two from urban to rural and from rural to alpine air, while the TC mass fraction decreases from ~40 to ~20%. The EC/TC ratios in $PM_{2.5}$ are as high as ~50% in the urban air samples taken close to a major traffic junction and on the order of ~30% in rural and high alpine air, demonstrating the strong impact of diesel soot and other fossil fuel combustion or biomass burning emissions on the atmospheric aerosol burden and composition. The water-soluble fraction of organic carbon (WSOC in OC), on the other hand, exhibits a pronounced increase from urban (~20%) to rural (~40%) and high alpine (~60%) samples of air particulate matter. This observation can be attributed to different aerosol sources (e.g. water-insoluble combustion particle components versus water-soluble biogenic and secondary organic particle components; see below) but also to chemical aging and oxidative transformation of organic aerosol components, which generally increases the number of functional groups and thus the water solubility of organic molecules.

Black or elemental carbon accounts for most of the light absorption by atmospheric aerosols and is therefore of crucial importance for the direct radiative effect of aerosols on climate [33–35]. Despite a long tradition of soot and aerosol research, however, there is still no universally accepted and applied operational definition of BC and EC. Several studies have compared the different optical and thermal methods applied by atmospheric research groups to measure BC and EC. Depending on techniques and measurement locations, fair agreement has been found in some cases, but mostly the results deviated considerably (up to 100% and more) [36–38].

Optical methods for the detection of BC are usually non-destructive and allow (near) real-time operation, but on the other hand they are particularly prone to misinterpretation. They generally rely on the assumptions that BC is the dominant absorber and has a uniform mass-specific absorption coefficient or cross-section. While these assumptions may be justified under certain conditions, they are highly questionable in the context of detailed chemical characterization of aerosol particles ("How black is black carbon?") [26]. In addition to different types of graphite-like

material, there are at least two classes of organic compounds which can contribute to the absorption of visible light by air particulate matter ("light-absorbing, yellow or brown carbon"): [21] polycyclic aromatics and humic-like substances. Therefore, optically determined BC values have to be considered as mass equivalent values but not as absolute mass or concentration values. Moreover, most conventional optical methods such as the aethalometer and integrating-sphere- and integrating-plate techniques are based on the measurement of light extinction rather than absorption. As a consequence, these methods require aerosol composition-dependent calibrations or additional sample work-up processes to compensate or minimize the influence of scattering aerosol components such as inorganic salts and acids on the measurement signal [37, 39, 40]. Alternatively, photoacoustic spectroscopy allows direct measurements of light absorption by airborne aerosol particles and in recent years several photoacoustic spectrometers have been developed and applied for the measurement of aerosol absorption coefficients and BC equivalent concentrations [41–43].

Among the few methods available for the characterization of the molecular and crystalline structures of BC and EC (graphite-like carbon proportion and degree of order) are high-resolution electron microscopy, X-ray diffraction and Raman spectroscopy [28, 29, 44]. These measurement techniques have revealed dependences of the microstructure and spectroscopic properties of flame soot, diesel soot and related carbonaceous materials on the processes and conditions of particle formation and aging. So far, however, these methods have been too labor intensive for routine investigations of atmospheric aerosol samples and their applicability to quantitative analyses remains to be proven [28, 29]. Nevertheless, recently developed measurement systems promise to allow the quantification of graphite-like carbon and soot in aerosol filter samples by Raman spectroscopy [45, 46].

7.2.2 Primary and Secondary Organic Components

The total mass of organic air particulate matter (OPM), i.e. the sum of organic aerosol (OA) components, is usually estimated by multiplication of OC by a factor of about 1.5–2, depending on the assumed average molecular composition and accounting for the contribution of elements other than carbon contained in organic substances (H, O, N, S, etc.) [21, 47]. The only way, however, to determine accurately the overall mass, molecular composition, physicochemical properties and potential toxicity of OPM is the identification and quantification of all relevant chemical components. Also, trace substances can be hazardous to human health and potential interferences of refractive and colored organics in the determination of BC or EC can be assessed only to the extent to which the actual chemical composition of OPM is known [26, 48].

Depending on their origin, OA components can be classified as primary or secondary. Primary organic aerosol (POA) components are directly emitted in the condensed phase (liquid or solid particles) or as semi-volatile vapors, which are condensable under atmospheric conditions. The main sources of POA particles and

components are natural and anthropogenic biomass burning (wildfires, slashing and burning, domestic heating); fossil fuel combustion (domestic, industrial, traffic related); and wind-driven or traffic-related suspension of soil and road dust, biological materials (plant and animal debris, microorganisms, pollen, spores, etc.), sea spray and spray from other surface waters with dissolved organic compounds.

Secondary organic aerosol (SOA) components are formed by chemical reaction and gas-to-particle conversion of volatile organic compounds (VOCs) in the atmosphere, which may proceed via different pathways:

1. new particle formation: formation of semi-volatile organic compounds (SVOCs) by gas-phase reactions and participation of the SVOCs in the nucleation and growth of new aerosol particles;
2. gas–particle partitioning: formation of SVOCs by gas-phase reactions and uptake (adsorption or absorption) by pre-existing aerosol or cloud particles;
3. heterogeneous or multiphase reactions: formation of low-volatility or non-volatile organic compounds (LVOCs, NVOCs) by chemical reaction of VOCs or SVOCs at the surface or in the bulk of aerosol or cloud particles.

The formation of new aerosol particles from the gas phase generally proceeds via the nucleation of nanometer-sized molecular clusters and subsequent growth by condensation of condensable vapor molecules. Experimental evidence from field measurements and model simulations suggest that new particle formation in the atmosphere is most likely dominated by ternary nucleation of H_2SO_4-H_2O–NH_3 and subsequent condensation of SVOCs [49–51]. Laboratory experiments and quantum chemical calculations indicate, however, that SVOCs might also play a role in the nucleation process (H_2SO_4–SVOC complex formation) [52]. The actual importance of different mechanisms of particle nucleation and growth in the atmosphere has not yet been fully unraveled and quantified. In any case, however, the formation of new particles exhibits a strong and non-linear dependence on atmospheric composition and meteorological conditions, may be influenced by ions and electric charge effects and competes with gas–particle partitioning and heterogeneous or multiphase reactions [53]. Among the principal parameters governing secondary particle formation are temperature, relative humidity and the concentrations of organic and inorganic nucleating and condensing vapors, which depend on atmospheric transport in addition to local sources and sinks such as photochemistry and pre-existing aerosol or cloud particles [23, 25, 49, 50]. The rate and equilibrium of SVOC uptake by aerosol particles depend on the SVOC accommodation coefficients and on the particle surface area, bulk volume and chemical composition (kinetics and thermodynamics of gas–particle partitioning) [54].

Most earlier studies of SOA formation had focused on pathways (1) and (2). Several recent studies indicate, however, that heterogeneous and multiphase reactions may also play an important role and substantially contribute to the overall atmospheric burden of OPM [21, 55–58]. The term "heterogeneous reaction" generally refers to reactions of gases at the particle surface, whereas the term "multiphase reaction" refers to reactions in the particle bulk involving species from the gas phase. A variety of different reversible and irreversible mechanisms of acid-catalyzed condensation and

Table 7.2 Prominent organic substance classes, characteristic magnitudes of their proportion in fine organic particulate matter (OPM; approximate upper limit of mass fraction) and their main sources.

Substance classes	Proportion	Sources
Aliphatic hydrocarbons	10^{-2}	Biomass, fossil fuel combustion
Aliphatic alcohols and carbonyls	10^{-2}	Biomass, SOA/aging
Levoglucosan	10^{-1}	Biomass burning
Fatty acids and other alkanoic acids	10^{-1}	Biomass, SOA/aging
Aliphatic dicarboxylic acids	10^{-1}	SOA/aging
Aromatic (poly-)carboxylic acids	10^{-1}	SOA/aging, soil/dust
Multifunctional aliphatics and aromatics (OH, CO, COOH)	10^{-1}	SOA/aging, soil/dust
Polycyclic aromatic hydrocarbons (PAHs)	10^{-3}	Fossil fuel combustion, biomass burning
Nitro- and oxy-PAHs	10^{-3}	Fossil fuel combustion, biomass burning, SOA/aging
Proteins and other amino compounds	10^{-1}	Biomass
Cellulose and other carbohydrates	10^{-2}	Biomass
Secondary organic oligomers/polymers and humic-like substances	10^{-1}	SOA/aging, soil/dust

radical-initiated oligo- or polymerization reactions involving organic and inorganic acids and photo-oxidants can lead to secondary formation of LVOCs and NVOCs of high molecular mass (SOA oligomers/polymers; Table 7.2). The actual atmospheric relevance and contributions of the different SOA formation pathways and involved chemical reaction mechanisms, however, still remain to be sorted out [21, 25].

Depending on local sources, meteorological conditions and atmospheric transport and thus on location, season and daytime, the composition of OPM can be dominated by POA or by SOA components. Recent studies indicate high abundance of POAs in tropical air masses due to intense biomass burning, whereas SOAs from biogenic and anthropogenic emissions of precursor VOCs seems to dominate in mid-latitude air masses. On a global scale, the formation of SOAs appears to be dominated by oxidation of biogenic VOCs (mostly by ozonolysis of terpenes) [59] and to amount to at least 50% of POA emissions [25, 60]. In the atmosphere, POA and SOA components are mixed with each other, BC/EC and inorganic aerosol components (externally and internally mixed aerosols) [61]. Moreover, both POA and SOA components can be efficiently transformed upon interaction with reactive trace gases and solar radiation (chemical aging, Section 7.3).

Hundreds of organic compounds have been detected in air particulate matter. Even in the most comprehensive investigations, however, only 10–40% of the OPM content estimated from OC measurements have been unambiguously identified on a molecular level. Prominent organic substance classes, characteristic magnitudes of their proportion in fine OPM (approximate upper limit of mass fraction) and their main sources are summarized in Table 7.2 [3, 6, 21, 25, 62–67].

Several studies have shown that macromolecules such as cellulose and proteins (molecular mass $\gg$1 kDa) and other compounds with relatively high molecular mass ($\gg$100 Da) such as humic-like substances (HULIS) account for large proportions of OPM and WSOC [21, 32, 68–72]. Obviously, biopolymers and humic substances are emitted as POA components (soil and road dust, sea-spray, biological particles), which may be modified by chemical aging and transformation in the atmosphere (e.g. formation of HULIS by oxidative degradation of biopolymers). On the other hand, organic compounds with high molecular mass can also originate from SOAs formation by heterogeneous and multiphase reactions at the surface and in the bulk of atmospheric particles as outlined above (SOA oligomers/polymers).

For the identification and quantification of individual organic compounds, filter and impactor samples are usually extracted with appropriate solvents and the extracts are analyzed by advanced instrumental or bioanalytical methods of separation and detection: gas and liquid chromatography; capillary electrophoresis; absorption, fluorescence and mass spectrometry; immunosorbent, enzyme and dye assays; etc. [26, 32, 62, 63, 66, 73–76]. Alternatively, deposited or suspended particles can be partially or fully vaporized by thermal or laser desorption and directly introduced into a gas chromatograph or spectrometer [3, 77–80].

In recent studies, nuclear magnetic resonance [18], Fourier transform infrared spectroscopy [47], scanning transmission X-ray microscopy [81] and aerosol mass spectrometry [82] have been applied to the efficient characterization and quantification of functional groups in OPM (alkyl, carbonyl, carboxyl and hydroxy groups; carbon double bonds and aromatic rings). These methods give valuable insight into the overall chemical composition, oxidation state and reactivity of OPM, but they provide only limited information about the actual identity of the individual compounds that present in the complex mixture. The molecular mass and structure of organic compounds, however, are crucial parameters for their physicochemical and biological properties and thus for their climate and health effects (volatility, solubility, hygroscopicity, CCN and IN activity, bioavailability, toxicity, allergenicity; Sections 7.3 and 7.4).

7.3 Transformation

Chemical reactions proceed at the surface and in the bulk of solid and liquid aerosol particles and can influence atmospheric gas-phase chemistry and properties of atmospheric particles and their effects on climate and human health[3, 6, 54, 83–92].

For example, aerosol chemistry leads to the formation of reactive halogen species, changes of reactive nitrogen and depletion of ozone – especially in the stratosphere, upper troposphere and marine boundary layer [93–103] On the other hand, chemical aging of aerosol particles generally changes their composition, decreases their reactivity, increases their hygroscopicity and cloud condensation activity and can change their optical properties [21, 26, 89, 104–109].

Because of their high surface-to-volume ratio, fine aerosol particles can be very efficiently transformed upon interaction with solar radiation (photolysis) and reactive trace gases (oxidation, nitration, acid–base reactions, hydrolysis, condensation or radical-initiated oligomerization, etc.). For example, oxidation and nitration reactions lead to the formation or degradation of hazardous aerosol components [8, 31, 48, 72], they cause artifacts upon collection and analysis of air particulate matter [3, 21, 110] and they play a major role in technical processes and devices for the control of combustion aerosol emissions [44, 111–113]. Moreover, the interaction with water can lead to structural rearrangements of solid aerosol particles, to the formation of highly concentrated aqueous solution droplets (hygroscopic growth) and to the formation of cloud droplets and ice crystals (Section 7.3.2).

Atmospheric aerosol transformations and gas–particle interactions generally involve multiple physicochemical processes such as mass transport, phase transition and chemical reaction at the interface or in the bulk of the gas phase, molecular clusters and liquid or solid particles, as illustrated in Figure 7.5. These multiphase processes are pivotal for the aerosol and cloud interactions and feedback loops outlined in Figures 7.1 and 7.3 and thus for the climate and health effects of atmospheric aerosols detailed below (Section 7.4).

Efficient investigation, elucidation and description of the interactions between multiple phases and chemical components of aerosols and clouds by laboratory experiments, field measurements, remote sensing and model studies require consistent terminologies and universally applicable mathematical formalisms and physical parameters. However, the current understanding of the mechanisms and kinetics of mass transport, phase transitions and chemical reactions in atmospheric aerosols and clouds is very limited. In addition to a lack of experimental data, one of the limitations is that the formalisms applied in different studies have mostly been restricted to specific systems and boundary conditions: liquid water, ice, acid

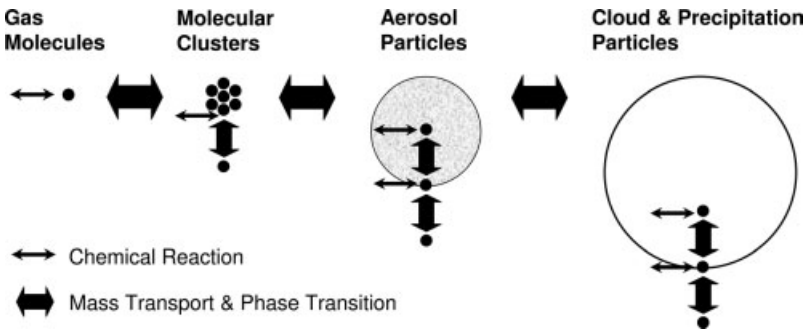


Figure 7.5 Schematic illustration of multiphase aerosol and cloud processes: mass transport and phase transitions of (semi-) volatile molecules between gas-phase, molecular clusters, aerosol particles and cloud or precipitation particles (thick arrows); chemical reactions in the gas phase, at the interface and in the particle bulk (thin arrows) [2].

hydrates, soot or mineral dust; fresh or aged surfaces; low or high reactant concentration levels, transient or (quasi-)steady-state conditions; limited selection of chemical species and reactions (see Ref. [54] and references therein). The different and sometimes inconsistent rate equations, parameters and terminologies make it difficult to compare, extrapolate and integrate the results of different studies over the wide range of reaction conditions relevant for the atmosphere, laboratory experiments, technical processes and emission control.

A comprehensive kinetic model framework for aerosol and cloud surface chemistry and gas–particle interactions has recently been proposed by Pöschl, Rudich and Ammann, abbreviated to PRA [54]. It allows to describe mass transport and chemical reactions at the gas–particle interface and to link surface processes with gas-phase and particle bulk processes in aerosol and cloud systems with unlimited numbers of chemical components and physicochemical processes. The key elements and essential aspects of the PRA framework are as follows:

1. a simple and descriptive double-layer surface model (sorption layer and quasi-static layer);
2. straightforward and additive flux-based mass balance and rate equations;
3. clear separation of mass transport and chemical reactions;
4. well-defined rate parameters (uptake and accommodation coefficients, reaction and transport rate coefficients);
5. clear distinction between different elementary and multistep transport processes (surface and bulk accommodation, etc.);
6. clear distinction between different elementary and multistep heterogeneous and multiphase reactions (Langmuir–Hinshelwood and Eley–Rideal mechanisms, etc.);
7. mechanistic description of complex concentration and time dependences;
8. flexible inclusion or omission of chemical species and physicochemical processes;
9. flexible convolution or deconvolution of species and processes;
10. full compatibility with traditional resistor model formulations.

Figure 7.6 illustrates the PRA model compartments and elementary processes at the gas–particle interface. The individual steps of mass transport are indicated by bold arrows besides the model compartments: gas-phase diffusion; reversible adsorption; mass transfer between sorption layer, quasi-static surface layer and near-surface particle bulk; diffusion in the particle bulk. The slim arrows inside the model compartments represent different types of chemical reactions: gas-phase reactions; gas–surface reactions; surface layer reactions; surface–bulk reactions; particle–bulk reactions [54]. Exemplary practical applications and model calculations demonstrating the relevance of these aspects have been presented in a companion paper [114].

The PRA framework is meant to serve as a common basis for experimental and theoretical studies investigating and describing atmospheric aerosol and cloud surface chemistry and gas–particle interactions. In particular, it will support the following research activities: planning and design of laboratory experiments for the

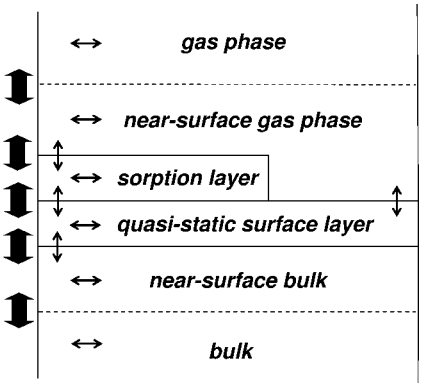


Figure 7.6 PRA framework model compartments, transport processes and chemical reactions at the gas–particle interface (double-layer surface model): fluxes of diffusion in the gas phase and particle bulk, adsorption and desorption, transfer between sorption layer and quasi-static surface layer and between quasi-static surface layer and near-surface particle bulk indicated by vertical thick arrows on the left side; elementary chemical reactions between species in the same or in different model compartments indicated by horizontal and vertical thin arrows [1, 54].

elucidation and determination of elementary processes and rate parameters; the establishment, evaluation and quality assurance of comprehensive and self-consistent collections of kinetic parameters; and the development of detailed master mechanisms for process models and the derivation of simplified but yet realistic parameterizations for atmospheric and climate models in analogy with atmospheric gas-phase chemistry [59, 115–119].

7.3.1
Chemical Transformation of Carbonaceous Aerosol Components

Organic aerosol components and the surface layers of BC or EC can react with atmospheric photo-oxidants (OH, O_3, NO_3, NO_2, etc.), acids (HNO_3, H_2SO_4, etc.), water and UV radiation. The chemical aging of OA components basically follows the generic reaction pathways outlined in Figure 7.7 and it tends to increase the oxidation state and water solubility of OC. In analogy with atmospheric gas-phase photochemistry of VOCs (methane, isoprene, terpenes, etc.) [59, 116, 117], oxidation, nitration, hydrolysis and photolysis transform hydrocarbons and derivatives with one or few functional groups into multifunctional hydrocarbon derivatives. The cleavage of organic molecules and release of SVOCs, VOCs, CO or CO_2 can also lead to volatilization of OPM. On the other hand, oxidative modification and degradation of biopolymers may convert these into HULIS (analogy with the formation of humic substances in soil, surface water and groundwater processes). Moreover, condensation reactions and radical-initiated oligo- or polymerization can decrease the volatility of OA components and promote the formation of SOAs particulate matter (SOA oligomers or HULIS, respectively; Table 7.2; Section 7.2.2).

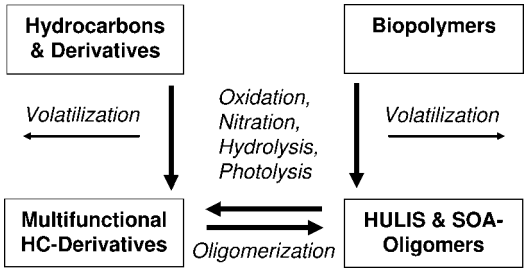


Figure 7.7 Generic reaction pathways for the atmospheric transformation (chemical aging) of organic aerosol components (left side, low molecular mass; right side, high molecular mass) [1].

The actual reaction mechanisms and kinetics, however, have been elucidated and fully characterized only for a small number of model reaction systems and components. So far, most progress has been made in the kinetic investigation and modeling of chemical reactions in cloud droplets [120, 121]. For the reasons outlined above, very few reliable and widely applicable kinetic parameters are available for organic reactions at the surface and in the bulk of liquid and solid aerosol particles [21, 54, 89, 122–124].

Several studies have shown that surface reactions of organic molecules and black or elemental carbon with gaseous photo-oxidants such as ozone and nitrogen dioxide tend to exhibit non-linear concentration dependences and competitive co-adsorption of different gas-phase components, which can be described by Langmuir–Hinshelwood reaction mechanisms and rate equations [54, 84, 89, 114, 125].

An example of such reactions is the degradation of benzo[*a*]pyrene (BaP) on soot by ozone. BaP is a polycyclic aromatic hydrocarbon (PAH) and prominent air pollutant with the chemical formula $C_{20}H_{12}$, consisting of five six-membered aromatic rings. It is one of the most hazardous carcinogens and mutagens among the 16 "priority polycyclic aromatic hydrocarbon pollutants" defined by the US Environmental Protection Agency (EPA). The main source of BaP in the atmosphere is combustion aerosols and it resides to a large extent on the surface of soot particles [48, 90, 110].

Figure 7.8 shows pseudo-first-order rate coefficients for the degradation of BaP on soot by ozone at gas-phase mole fractions or volume mixing ratios (VMR) up to 1 ppm under dry conditions and in the presence of water vapor [relative humidity (RH) 25%, 296 K, 1 atm]. These and complementary results of aerosol flow tube experiments and model calculations indicate reversible and competitive adsorption of O_3 and H_2O, followed by a slower, rate-limiting surface reaction between adsorbed O_3 and BaP on the soot surface. The kinetic parameters determined from the displayed non-linear least-squares fits (maximum pseudo-first-order rate coefficients and effective Langmuir adsorption equilibrium constants) allow the prediction of the half-life (50% decay time) of BaP on the surface of soot particles in the atmosphere. At typical ambient ozone VMR of ~30 ppb it would be only ~5 min under dry conditions and ~15 min at 25% RH.

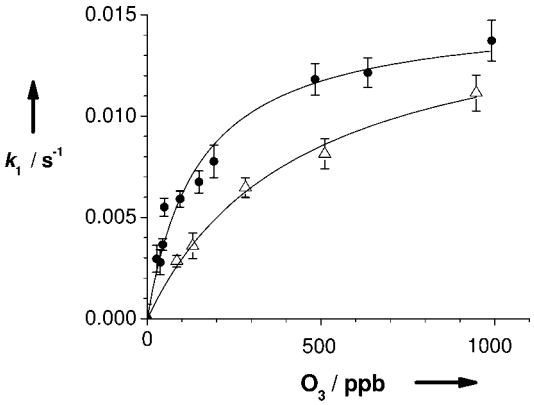


Figure 7.8 Pseudo-first-order rate coefficients (k_1) for the degradation of benzo[*a*]pyrene (BaP) on soot by ozone: measurement data from aerosol flow tube experiments under dry and humid conditions (symbols and error bars, arithmetic mean ± standard deviation; full circles, RH <1%; open triangles, RH ≈ 25%) and non-linear least-squares fit lines based on Langmuir–Hinshelwood rate equation [1, 90].

Figure 7.9 illustrates the recovery ratio (RR) of BaP from fine air particulate matter ($PM_{2.5}$) collected with a regular filter sampling system from urban air at ambient ozone VMRs up to 80 ppb (Munich, 2001–2002). The plotted RRs refer to filter samples collected in parallel with a system that removes ozone and other photo-oxidants from the sample air flow by means of an activated carbon diffusion denuder [110]. Thus deviations from unity represent the fraction of BaP degraded by reaction with ozone and other photo-oxidants from the sampled air during the

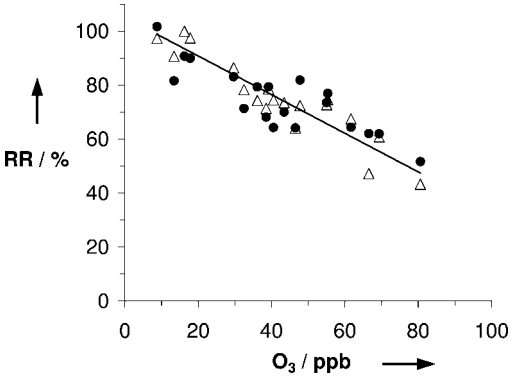


Figure 7.9 Recovery ratio for benzo[*a*]pyrene (BaP) (full circles) and the sum of all particle-bound five- and six-ring US EPA priority PAH pollutants, PAH(5,6) (open triangles), plotted against ambient ozone volume mixing ratio upon filter sampling of urban air particulate matter: measurement data points and linear least-squares fit [1, 110].

sampling process, i.e. the BaP loss by filter reaction sampling artifacts. The BaP recovery ratio is nearly identical with the recovery ratio to the sum of all particle-bound five- and six-ring US EPA priority PAH pollutants, PAH(5,6), and exhibits a negative linear correlation with ambient ozone. It decreases from unity at low ozone to ~0.5 at ~80 ppb O_3, which is a characteristic concentration level for polluted urban air in summer. Similar correlations have been observed in experiments performed at different locations and with different filter sampling and denuder systems [110].

With regard to chemical kinetics, the linear correlation between PAH recovery ratio and O_3 VMR can be attributed to the near-linear dependence of the PAH degradation rate coefficient on O_3 at low VRMs (VMR $\ll$ inverse of effective adsorption equilibrium constant; Figure 7.8) [54, 84, 90, 114]. Moreover, it indicates efficient protection and shielding of the PAH on deposited particles from further decay by coverage with subsequently collected particulate matter (build-up of "filter cake") on time scales similar to the half-life of PAH at the surface. Otherwise, the PAH recovery should be even lower and the ozone concentration dependence should be less pronounced.

In any case, the sampling artifacts observed by Schauer and co-workers and illustrated in Figure 7.9 imply that the real concentrations of particle-bound PAHs in urban air are up to ~100% higher than the measurement values obtained with simple filter sampling systems (without activated carbon diffusion denuder or equivalent equipment) as applied for most atmospheric research and air pollution monitoring purposes [48, 90, 110]. Clearly, other OA components with similar or higher reactivity towards atmospheric oxidants (e.g. alkenes) are also prone to similar or even stronger sampling artifacts, which have to be avoided or at least minimized and quantified for accurate and reliable determination of atmospheric aerosol composition and properties. These and other potential sampling and analytical artifacts caused by reactive transformation of fine air particulate matter have to be taken into account not only in atmospheric and climate research activities, but also in air pollution control. In particular, the control and enforcement of emission limits and ambient threshold level values for OA components which pose a threat to human health (Section 7.4.2) require the development, careful characterization and validation and correct application of robust analytical techniques and procedures [48].

As far as the atmospheric aerosol cycling and feedback loops are concerned (Figures 7.1 and 7.3), chemical aging and oxidative degradation of organics present on the surface and in the bulk generally make aerosol particles more hydrophilic or hygroscopic and enhance their ability to act as a CCN. Besides their contribution to the water-soluble fraction of particulate matter, partially oxidized organics can act as surfactants and influence the hygroscopic growth, CCN and IN activation of aerosol particles (Section 7.3.2).

The chemical reactivity of carbonaceous aerosol components also plays an important role in technical applications for the control of combustion aerosol emissions. For example, the lowering of emission limits for soot and related diesel exhaust particulate matter (DPM) necessitates the development and implementation of efficient exhaust aftertreatment technologies such as diesel particulate filters or particle traps with open deposition structures. These systems generally require

regeneration by oxidation and gasification of the soot deposits in the filter or catalyst structures. Usually the regeneration is based on discontinuous oxidation by O_2 at high temperatures (>500 °C) or continuous oxidation by NO_2 at moderate exhaust temperatures (200–500 °C) [29, 44, 111–113, 125]. Efficient optimization of the design and operating conditions of such exhaust aftertreatment systems requires comprehensive kinetic characterization and mechanistic understanding of the involved chemical reactions and transport processes. Recent investigations have shown that the oxidation and gasification of diesel soot by NO_2 at elevated concentration and temperature levels (up to 800 ppm NO_2 and 500 °C) follows a similar Langmuir–Hinshelwood reaction mechanism as the oxidation of BaP on soot by O_3 at ambient concentration and temperature level (up to 1 ppm O_3 and 30 °C) [29, 90, 125].

7.3.2 Restructuring, Phase Transitions, Hygroscopic Growth and CCN/IN Activation of Aerosol Particles upon Interaction with Water Vapor

Water vapor molecules interacting with aerosol particles can be adsorbed on the particles' surface or absorbed into the particles' bulk. For particles consisting of water-soluble material, the uptake of water vapor can lead to aqueous solution droplet formation and substantial increase of the particle diameter (hygroscopic growth) even at low relative humidities (RH <100%; atmospheric gas-phase water partial pressure < equilibrium vapor pressure of pure liquid water) [10].

At water vapor supersaturation (RH >100%), aerosol particles can act as nuclei for the formation of liquid cloud droplets [cloud condensation nuclei (CCN)]. For the formation of water droplets from a homogeneous gas phase devoid of aerosol particles supersaturations up to several hundred percent would be required (thermodynamic barrier for the homogenous nucleation of a new phase). In the atmosphere, however, water vapor supersaturations with respect to liquid water generally remain below 10% and mostly even below 1%, because aerosol particles induce heterogeneous nucleation, condensation and cloud formation. At low temperatures and high altitudes, clouds consist of mixtures of liquid water droplets and ice crystals or entirely of ice crystals. The formation of ice crystals is also induced by pre-existing aerosol particles, so-called ice nuclei (IN), as detailed below. Ice nucleation in clouds usually requires temperatures well below 0 °C, which can lead to high water vapor supersaturations with respect to ice [10, 126–131].

The minimum supersaturation at which aerosol particles can be effectively activated as CCN or IN is called critical supersaturation. It is determined by the physical structure and chemical composition of the particles and it generally decreases with increasing particle size. For insoluble CCN the critical supersaturation depends on the wettability of the surface (contact angle of liquid water) and for partially or fully soluble CCN it depends on the mass fraction, hygroscopicity and surfactant activity of the water-soluble particulate matter [10, 22, 24, 132, 133].

The nucleation of ice crystals on atmospheric aerosol particles can proceed via different pathways or modes. In the deposition mode, water vapor is adsorbed and immediately converted into ice on the surface of the IN (deposition or sorption

nuclei). In the condensation freezing mode, the aerosol particles act first as CCN and induce the formation of supercooled aqueous droplets which freeze later on (condensation freezing nuclei). In the immersion mode the IN are incorporated into pre-existing aqueous droplets and induce ice formation upon cooling (immersion nuclei). In the contact mode, freezing of a supercooled droplet is initiated upon contact with the surface of the IN (contact nuclei). Obviously, the IN activity of aerosol particles depends primarily on their surface composition and structure, but condensation and immersion freezing can also be governed by water-soluble bulk material [10, 130, 134–140].

Most water-soluble aerosol components are hygroscopic and absorb water to form aqueous solutions at RH <100%. The phase transition of dry particle material into a saturated aqueous solution is called deliquescence and occurs upon exceeding a substance-specific RH threshold value [deliquescence relative humidity (DRH)]. The reverse transition and its RH threshold value are called efflorescence and efflorescence relative humidity (ERH), respectively. The hygroscopic growth and CCN activation of aqueous solution droplets can be described by the so-called Köhler theory, which combines Raoult's law or alternative formulations for the activity of water in aqueous solutions and the Kelvin equation for the dependence of vapor pressure on the curvature and surface tension of a liquid droplet [10, 22, 24, 132, 141–146].

Figure 7.10a shows a typical example of the hygroscopic growth of water-soluble inorganic salts contained in air particulate matter: the hygroscopic growth curve (humidogram) of pure NaCl aerosol particles with dry particle diameters of ~100 nm measured in a hygroscopicity tandem differential mobility analyzer (H-TDMA) experiment at relative humidities up to 95%. Upon hydration (increase in RH), the crystalline NaCl particles undergo a deliquescence transition at DRH ≈ 75%. The water uptake and dependence of the aqueous solution droplet diameter on RH agree very well with Köhler theory calculations, which are based on a semi-empirical ion interaction parameterization of water activity and account for the effects of particle shape transformation (cubic crystals and spherical droplets; mobility and mass equivalent diameters) [141]. The hysteresis branch measured upon dehydration (decrease in RH) is due to the existence of solution droplets in a metastable state of NaCl supersaturation (ERH < RH < DRH). The efflorescence transition, i.e. the formation of a salt crystals and evaporation of the liquid water, occurs at ERH ≈ 40%.

Figure 7.10b displays the hygroscopic growth curve of aerosol particles composed of pure bovine serum albumin (BSA) as a model for globular proteins and similar organic macromolecules. The hygroscopic growth is much less pronounced than for inorganic salts but still significant, with deliquescence and efflorescence transitions at DRH ≈ ERH ≈ 40% (conversion of dry protein particles into saturated aqueous solution or gel-like droplets, v.v.) and no significant deviations between hydration and dehydration (no hysteresis effect). The dependence of the deliquesced particle diameter on RH is in good agreement with Köhler theory calculations based on a simple osmotic pressure parameterization of water activity, which has been derived under the assumption that the dissolved protein macromolecules behave like inert solid spheres [141].

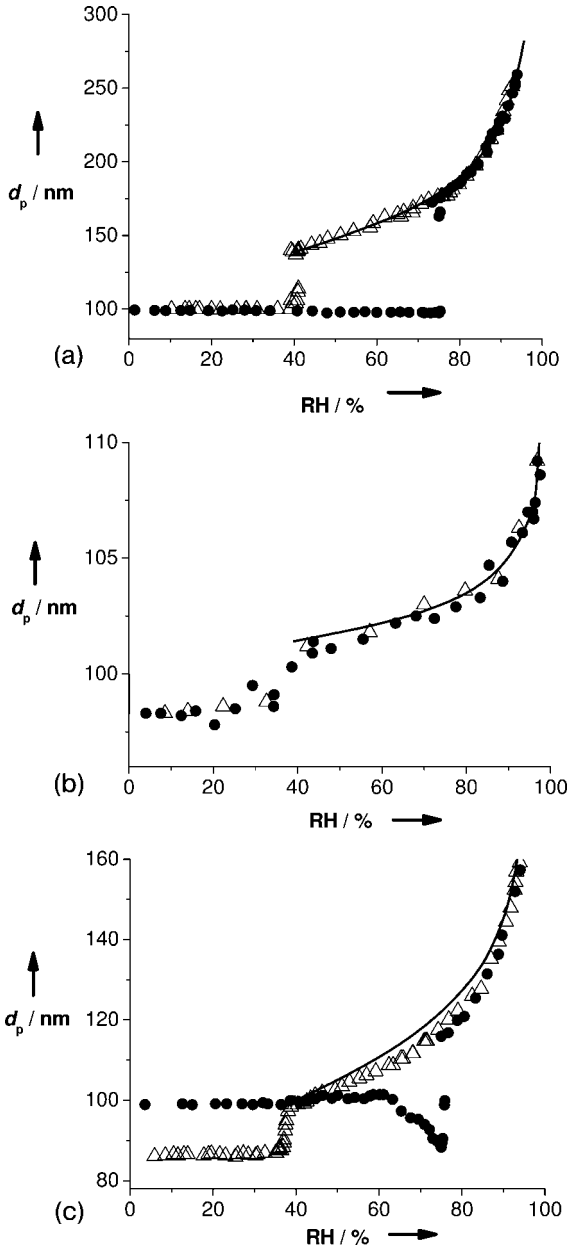


Figure 7.10 Hygroscopic growth curves for pure NaCl salt particles (a), pure BSA protein particles (b) and internally mixed BSA–NaCl protein–salt particles (c): data points measured upon hydration (full circles) and dehydration (open triangles) in H-TDMA aerosol experiments; solid lines represent Köhler theory calculations based on NaCl ion interaction and BSA osmotic pressure parameterizations for water activity [1, 141].

Figure 7.10c shows the hygroscopic growth curve of internally mixed NaCl–BSA particles (mass ratio 1:1) with dry particle diameter of ∼100 nm. The mixed aerosol particles have been generated in full analogy with the pure NaCl and pure BSA particles (nebulization of an aqueous solution). Upon hydration, however, the particles exhibit a significant decrease in the measured (mobility equivalent) diameter as the relative humidity approaches the deliquescence threshold (DRH ≈ 75%). The observed minimum diameter is ∼10% smaller than the initial diameter, indicating high initial porosity of the particles (envelope void fraction ∼30%) and strong restructuring upon humidification. Upon dehydration, the efflorescence threshold is lower than for pure NaCl (ERH ≈ 37% vs. 40%), indicating that the protein macromolecules inhibit the formation of salt crystals and enhance the stability of supersaturated salt solution droplets. The particle diameters observed after efflorescence essentially equal the minimum diameter observed prior to deliquescence. The hygroscopic growth of the deliquesced particles (aqueous solution droplets) is in fair agreement with Köhler theory calculations based on the observed minimum diameter rather than the initial diameter and on the assumption of simple solute additivity (linear combination of NaCl ion interaction and BSA osmotic pressure parameterizations of water activity) [141]. These and complementary measurement and modeling results can be explained by the formation of porous agglomerates due to ion–protein interactions and electric charge effects on the one hand and by compaction of the agglomerate structure due to capillary condensation and surface tension effects on the other.

Depending on their origin and conditioning, aerosol particles containing inorganic salts and organic (macro-) molecules can have complex and highly porous microstructures, which are influenced by electric charge effects and interaction with water vapor. Proteins and other surfactants tend to be enriched at the particle surface and form an envelope which can inhibit the access of water vapor to the particle core and lead to kinetic limitations of hygroscopic growth, phase transitions, CCN and IN activation. Formation and effects of organic surfactant films on sea salt particles have been discussed by O'Dowd *et al.* [147]. These and other effects of (non-linear) interactions between organic and inorganic aerosol components have to be further elucidated and considered for consistent analysis of measurement data from laboratory experiments and field measurements and reliable modeling of atmospheric aerosol processes (Figures 7.1 and 7.3).

Structural rearrangements, hygroscopic growth, phase transitions and CCN/IN activation of aerosol particles interacting with water vapor are not only important for the formation and properties of clouds and precipitation (number density and size of cloud droplets and ice particles; temporal and spatial distribution and intensity of precipitation). They also influence the chemical reactivity and aging of atmospheric particles (accessibility of particle components to reactive trace gases and radiation), their optical properties (absorption and scattering cross-sections) and their health effects upon inhalation into the human respiratory tract (deposition efficiency and bioavailability). Therefore, the water interactions of particles with complex chemical composition are widely and intensely studied in current aerosol, atmospheric and climate research. So far, however, their mechanistic and quantitative

understanding are still rather limited, especially with regard to carbonaceous components [5, 10, 21, 24, 25, 130, 146, 148–153].

7.4 Climate and Health Effects

Anthropogenic emissions are major sources of atmospheric aerosols. In particular, the emissions of particles and precursor gases from biomass burning and fossil fuel combustion have increased massively since pre-industrial times and account for a major fraction of fine air particulate matter in polluted urban environments and in the global atmosphere (carbonaceous components, sulfate, etc.) [3, 4, 6, 11, 154–160]. Numerous studies have shown that both natural and anthropogenic aerosols have a strong impact on climate and human health. Due to the limited knowledge of aerosol sources, composition, properties and processes outlined above, however, the actual effects of aerosols on climate and health are still far from being fully understood and quantified.

Aerosol effects on climate are generally classified as direct or indirect with respect to radiative forcing of the climate system. Radiative forcings are changes of the energy fluxes of solar radiation (maximum intensity in the spectral range of visible light) and terrestrial radiation (maximum intensity in the infrared spectral range) in the atmosphere, induced by anthropogenic or natural changes in atmospheric composition, the Earth's surface properties or solar activity. Negative forcings such as the scattering and reflection of solar radiation by aerosols and clouds tend to cool the Earth's surface, whereas positive forcings such as the absorption of terrestrial radiation by greenhouse gases and clouds tend to warm it (greenhouse effect) [4].

The optical properties relevant for the direct effects (scattering and absorption coefficient or extinction cross section and single scattering albedo, etc.) and the CCN, IN, chemical and biological activities relevant for indirect effects are determined by aerosol particle size, structure and chemical composition. Hence they are strongly influenced by the atmospheric processes outlined above (coagulation, chemical transformation, water interactions). Consequently, the actual climate system responses and feedbacks to natural or anthropogenic perturbations such as industrial and traffic-related greenhouse gas and aerosol emissions, volcanic eruptions, etc., are highly uncertain. In many cases, even the sign or direction of the feedback effect is unknown, i.e. it is not clear whether a perturbation will be reinforced (positive feedback) or damped (negative feedback) [1].

Numerous epidemiological studies have shown that fine air particulate matter and traffic-related air pollution are correlated with severe health effects, including enhanced mortality, cardiovascular, respiratory and allergic diseases [7, 161–164]. Moreover, toxicological investigations *in vivo* and *in vitro* have demonstrated substantial pulmonary toxicity of model and real environmental aerosol particles, but the biochemical mechanisms and molecular processes which cause the toxicological effects like oxidative stress and inflammatory response have not yet been resolved. Among the parameters and components potentially relevant for aerosol health effects

Table 7.3 Possible mechanisms by which aerosol particles and other air pollutants may cause adverse health effects (based on Bernstein *et al.* [7]).

(a)	Pulmonary inflammation induced by PM or O_3
(b)	Free radical and oxidative stress generated by transition metals or organic compounds (e.g. PAHs)
(c)	Covalent modification of key intracellular proteins (e.g. enzymes)
(d)	Inflammation and innate immune effects induced by biological compounds such as endotoxins and glucans
(e)	Stimulation of nocioreceptor and autonomic nervous system activity regulating heart rate variability and airway reactivity
(f)	Adjuvant effects in the immune system (e.g. DPM and transition metals enhancing responses to common environmental allergens)
(g)	Procoagulant activity by ultrafine particle accessing the systemic circulation
(h)	Suppression of normal defense mechanisms (e.g. suppression of alveolar macrophage functions)

are the specific surface, transition metals and organic compounds [7, 165–167]. Some of the possible mechanisms by which air particulate matter and other pollutants may affect human health are summarized in Table 7.3.

Ultrafine particles ($d_p < 100$ nm) are suspected to be particularly hazardous to human health, because they are sufficiently small to penetrate the membranes of the respiratory tract and enter the blood circulation or be transported along olfactory nerves into the brain [168–170]. Neither for ultrafine nor for larger aerosol particles, however, is it clear which physical and chemical properties actually determine their adverse health effects (particle size, structure, number and mass concentration, solubility, chemical composition and individual components, etc.).

Particularly little is known about the relations between allergic diseases and air quality. Nevertheless, traffic-related air pollution with high concentration levels of fine air particulate matter, nitrogen oxides and ozone is one of the prime suspects, besides unnatural nutrition and exaggerated hygiene, which may be responsible for the strong increase in allergies in industrialized countries over recent decades [7, 171–173]. The most prominent group of airborne allergens are protein molecules, which account for up to ~5% of urban air particulate matter. They are not only contained in coarse biological particles such as pollen grains (diameter >10 μm) but also in the fine fraction of air particulate matter, which can be explained by fine fragments of pollen, microorganisms or plant debris and by mixing of proteins dissolved in rain water with fine soil and road dust particles [69, 72, 174].

A molecular rationale for the promotion of allergies by traffic-related air pollution has been proposed by Franze and co-workers [72, 73], who found that proteins including birch pollen allergen Bet v 1 are efficiently nitrated by polluted urban air. The nitration reaction converts the natural aromatic amino acid tyrosine into nitrotyrosine and proceeds particularly fast at elevated NO_2 and O_3 concentrations (photo-smog or summer smog conditions), most likely involving nitrate radicals (NO_3) as reactive intermediates. From biomedical and immunological research, it is

known that protein nitration occurs upon inflammation of biological tissue, where it may serve to mark foreign proteins and guide the immune system. Moreover, conjugates of proteins and peptides with nitroaromatic compounds were found to evade immune tolerance and boost immune responses and post-translational modifications generally appear to enhance the allergenicity of proteins [72]. Thus the inhalation of aerosols containing nitrated proteins or nitrating reagents is likely to trigger immune reactions, promote the genesis of allergies and enhance the intensity of allergic diseases and airway inflammations. This hypothesis is supported by first results of ongoing biochemical experiments with nitrated proteins [175].

By means of newly developed enzyme immunoassays, nitrated proteins have been detected in urban road and window dust and fine air particulate matter, exhibiting degrees of nitration up to 0.1%. Upon exposure of birch pollen extract to heavily polluted air at a major urban traffic junction and to synthetic gas mixtures containing NO_2 and O_3 at concentration levels characteristic of intense summer smog, the degrees of nitration rose by up to 20%. The experimental results indicate that Bet v 1 is more easily nitrated than other proteins, which might be an explanation for why it is a particularly strong allergen [72]. If the ongoing biochemical experiments and further studies confirm that protein nitration by nitrogen oxides and ozone is indeed an important link between air pollution, airway inflammations and allergies, the spread and enhancement of these diseases could be encountered by the improvement of air quality and reduction of emission limits for nitrogen oxides and other traffic-related air pollutants. Moreover, it might be possible to develop pharmaceuticals against the adverse health effects of nitrated proteins.

Efficient control of air quality and related health effects requires a comprehensive understanding of the identity, sources, atmospheric interactions and sinks of hazardous pollutants. Without this understanding, the introduction of new laws, regulations and technical devices for environmental protection runs the risk of being ineffective or even of doing more harm than good through unwanted side-effects.

For example, epidemiological evidence for adverse health effects of fine and ultrafine particles has led to a lowering of present and future emission limits for soot and related DP [29, 113, 170, 176–178]. For compliance with these emission limits, different particle filter or trapping and exhaust aftertreatment technologies have been developed and are currently being introduced in diesel vehicles. Depending on the design of the particle filter or trap and catalytic converter systems, their operation can lead to substantial excess NO_2 emissions [29, 125]. If, however, elevated NO_2 concentrations and the nitration of proteins indeed promote allergies, such systems could reduce respiratory and cardiovascular diseases related to soot particles but at the same time enhance allergic diseases. Moreover, elevated NO_2 concentrations and incomplete oxidation of soot in exhaust filter systems could also increase the emissions of volatile or semi-volatile hazardous aerosol components such as nitrated PAH derivatives [31, 48]. Hence effective mitigation of the adverse health effects of diesel engine exhaust may require the introduction of advanced catalytic converter systems which minimize the emissions of both particulate and gaseous pollutants (soot, PAHs and PAH derivatives, nitrogen oxides, etc.) rather than simple particle filters.

In any case, comprehensive investigations and understanding and control of aerosol health effects need to consider both the particulate and gaseous components of aerosols in addition to their chemical reactivity and aging [48].

7.5 Summary and Outlook

Scientific investigations of and reports on atmospheric aerosols date back as long as to the 18th century and since then it has become increasingly clear that aerosol particles are of major importance for atmospheric chemistry and physics, the hydrological cycle, climate and human health [179]. Motivated by global change and adverse health effects of traffic-related air pollution, aerosol research activities have been increasingly extended and intensified over the past couple of decades [1].

These activities have led to a fairly comprehensive conceptual understanding of atmospheric aerosol sources, composition, properties, interactions and effects on climate. The parameters required for a quantitative description of the underlying physicochemical processes, however, are generally still uncertain by factors of two or more, which implies order of magnitude uncertainties for most effects involving multiple competitive or sequential processes. In some cases such as particle nucleation and reactive gas uptake, even the basic parameters are uncertain by orders of magnitude. Consequently, model calculations of atmospheric aerosol effects on future climate have to be regarded as sensitivity studies with more or less reliable qualitative and semi-quantitative results and implications rather than reliable quantitative predictions. In particular, interactions and feedback responses between aerosols and clouds, the hydrological cycle and the biosphere are difficult to quantify with the currently available information. Regardless of the rapid increase in numerical simulation capacities, this situation can hardly change before the basic physicochemical processes and properties of atmospheric aerosol particles have been elucidated to an extent comparable to the present state of knowledge of atmospheric gas-phase chemistry (universally applicable and validated master mechanisms, rate coefficients and structure–reactivity relationships, etc.).

Outstanding open questions and research aims for the elucidation of aerosol effects relevant for the science and policy of global change have been outlined in several recent monographs, reviews and research articles [4, 5, 23, 49, 50, 54]. Among these are the quantification, mechanistic elucidation and kinetic characterization of the following processes: formation of new particles and secondary organic aerosols; emission of primary organic aerosol components and black or elemental carbon; aging and deposition of aerosol particles; activation of cloud condensation and ice nuclei. As far as chemical transformation, heterogeneous and multiphase reactions and gas–particle interactions of aerosols and clouds are concerned, one of the most important prerequisites for efficient further investigation and scientific progress is the establishment of a common basis of consistent, unambiguous and universally applicable terminologies, model formalisms and kinetic and thermodynamic parameters.

With regard to atmospheric aerosol effects on human health, not only the quantitative but also the qualitative and conceptual understanding are very limited. Epidemiological and toxicological studies indicate strong adverse health effects of fine and ultrafine aerosol particles (nanoparticles) in addition to gaseous air pollutants, but the causative relations and mechanisms are hardly known [7, 164]. Their elucidation, however, is required for the development of efficient strategies of air quality control and medical treatment of related diseases, which permit the minimization of adverse aerosol health effects at minimum social and economic cost.

Particularly little is known about the relations between allergic diseases and air pollution and the interactions between natural aeroallergens and traffic-related pollutants. Several studies have shown synergistic and adjuvant effects of diesel particulate matter, O_3, NO_2 and allergenic pollen proteins, but the specific chemical reactions and molecular processes responsible for these effects have not yet been unambiguously identified. Recent investigations indicate that the nitration of allergenic proteins by polluted air may play an important role. Nitrated proteins are known to stimulate immune responses and they could promote the genesis of allergies, enhance allergic reactions and influence inflammatory processes, which is confirmed by the results of ongoing biochemical investigations [72, 175].

For efficient elucidation and abatement of adverse aerosol health effects, the knowledge of atmospheric and biomedical aerosol research should be integrated to formulate plausible hypotheses that specify potentially hazardous chemical substances and reactions on a molecular level. These hypotheses have to be tested in appropriate biochemical and medical studies, to identify the most relevant species and mechanisms of interaction and to establish the corresponding dose–response relationships. Ultimately, the identification and characterization of hazardous aerosol components and their sources and sinks (emission, transformation, deposition) should allow the optimization of air pollution control and the medical treatment of aerosol effects on human health.

List of Abbreviations, Acronyms and Symbols

asl	above sea level
BaP	benzo[*a*]pyrene
BC	black carbon
BSA	bovine serum albumin
CCN	cloud condensation nucleus
d_p	particle diameter
$dN/d\log d_p$	particle number size distribution function
DPM	diesel exhaust particulate matter
DRH	deliquescence relative humidity
EC	elemental carbon
ELPI	electrical low-pressure impactor
EPA	Environmental Protection Agency
ERH	efflorescence relative humidity

HC	hydrocarbon
H-TDMA	hygroscopicity tandem differential mobility analyzer
HULIS	humic-like substances
IN	ice nucleus
k_1	(pseudo-) first-order rate coefficient
LVOC	low-volatility organic compound
MWSOC	macromolecular water-soluble organic carbon
NVOC	non-volatile organic compound
OA	organic aerosol
OC	organic carbon
OPM	organic particulate matter
PAH	polycyclic aromatic hydrocarbon
PAH(5,6)	polycyclic aromatic hydrocarbons consisting of five or six aromatic rings
PM	particulate matter
$PM_{2.5}$ (1 or 10)	particulate matter of particles with aerodynamic diameters $\leq$2.5 μm (1 or 10 μm)
POA	primary organic aerosol
PRA	Pöschl, Rudich and Ammann
RH	relative humidity
RR	recovery ratio
SMPS	scanning mobility particle sizer
SOA	secondary organic aerosol
SVOC	semi-volatile organic compound
TC	total carbon
UV	ultraviolet
VMR	volume mixing ratio
VOC	volatile organic compound
WSOC	water-soluble organic carbon

References

1 U. Pöschl, *Angewandte Chemie International Edition* 2005, **44**, 7520.

2 S. Fuzzi, M. O. Andreae, B. J. Huebert, M. Kulmala, T. C. Bond, M. Boy, S. J. Doherty, A. Guenther, M. Kanakidou, K. Kawamura, V.-M. Kerminen, U. Lohmann, L. M. Russell, U. Pöschl, *Atmospheric Chemistry and Physics* 2006, **6**, 2017.

3 B. J. Finlayson-Pitts, J. N. Pitts, *Chemistry of the Upper and Lower Atmosphere*, Academic Press, San Diego, CA, 2000.

4 J. T. Houghton, Y. Ding, D. J. Griggs, M. Noguer, P. J. van der Linden, X. Dai, K. Maskell, C. A. Johnson, *Climate Change 2001: The Scientific Basis (Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change)*, Cambridge University Press, Cambridge, 2001.

5 U. Lohmann, J. Feichter, *Atmospheric Chemistry and Physics* 2005, **5**, 715.

6 J. H. Seinfeld, S. N. Pandis, *Atmospheric Chemistry and Physics*, Wiley, New York, 1998.

7 J. A. Bernstein, N. Alexis, C. Barnes, I. L. Bernstein, A. Nel, D. Peden, D. Diaz-Sanchez, S. M. Tarlo, P. B. Williams, *Journal of Allergy and Clinical Immunology* 2004, **114**, 1116.

8 B. J. Finlayson-Pitts, J. N. Pitts, *Science* 1997, **276**, 1045.

9 W. C. Hinds, *Aerosol Technology*, Wiley, New York, 1999.

10 H. R. Pruppacher, J. D. Klett, *Microphysics of Clouds and Precipitation*, Kluwer, Dordrecht, 1997.

11 F. Raes, R. Van Dingenen, E. Vignati, J. Wilson, J. P. Putaud, J. H. Seinfeld, P. Adams, *Atmospheric Environment* 2000, **34**, 4215.

12 J. Williams, M. de Reus, R. Krejci, H. Fischer, J. Strom, *Atmospheric Chemistry and Physics* 2002, **2**, 133.

13 R. Van Dingenen, F. Raes, J. P. Putaud, U. Baltensperger, A. Charron, M. C. Facchini, S. Decesari, S. Fuzzi, R. Gehrig, H. C. Hansson, R. M. Harrison, C. Huglin, A. M. Jones, P. Laj, G. Lorbeer, W. Maenhaut, F. Palmgren, X. Querol, S. Rodriguez, J. Schneider, H. ten Brink, P. Tunved, K. Torseth, B. Wehner, E. Weingartner, A. Wiedensohler, P. Wahlin, *Atmospheric Environment* 2004, **38**, 2561.

14 R. Krejci, J. Ström, M. de Reus, J. Williams, H. Fischer, M. O. Andreae, H.-C. Hansson, *Atmospheric Chemistry and Physics* 2005, **5**, 1527.

15 J. P. Putaud, F. Raes, R. Van Dingenen, E. Bruggemann, M. C. Facchini, S. Decesari, S. Fuzzi, R. Gehrig, C. Huglin, P. Laj, G. Lorbeer, W. Maenhaut, N. Mihalopoulos, K. Mulller, X. Querol, S. Rodriguez, J. Schneider, G. Spindler, H. ten Brink, K. Torseth, A. Wiedensohler, *Atmospheric Environment* 2004, **38**, 2579.

16 A. Zerrath, R. Niessner, U. Pöschl, *Journal of Aerosol Science* 2003, **34**, S649.

17 A. Zerrath, *Analytik von Cellulose und Glucose in atmosphärischen Aerosolproben und physikalische Aerosol-Charakterisierung mit einem elektrischen Niederdruckimpaktor*, PhD Thesis, Technical University Munich, 2005.

18 E. Matta, M. C. Facchini, S. Decesari, M. Mircea, F. Cavalli, S. Fuzzi, J. P. Putaud, A. Dell'Acqua, *Atmospheric Chemistry and Physics* 2003, **3**, 623.

19 J. P. Putaud, R. Van Dingenen, A. Dell'Acqua, F. Raes, E. Matta, S. Decesari, M. C. Facchini, S. Fuzzi, *Atmospheric Chemistry and Physics* 2004, **4**, 889.

20 M. Ebert, S. Weinbruch, P. Hoffmann, H. M. Ortner, *Atmospheric Environment* 2004, **38**, 6531.

21 A. Gelencser, *Carbonaceous Aerosol*, Springer, Dordrecht, 2004.

22 S. Henning, T. Rosenorn, B. D'Anna, A. A. Gola, B. Svenningsson, M. Bilde, *Atmospheric Chemistry and Physics* 2005, **5**, 575.

23 M. Kulmala, T. Suni, K. E. J. Lehtinen, M. Dal Maso, M. Boy, A. Reissell, U. Rannik, P. Aalto, P. Keronen, H. Hakola, J. B. Back, T. Hoffmann, T. Vesala, P. Hari, *Atmospheric Chemistry and Physics* 2004, **4**, 557.

24 R. Sorjamaa, B. Svenningsson, T. Raatikainen, S. Henning, M. Bilde, A. Laaksonen, *Atmospheric Chemistry and Physics* 2004, **4**, 2107.

25 M. Kanakidou, J. H. Seinfeld, S. N. Pandis, I. Barnes, F. J. Dentener, M. C. Facchini, R. Van Dingenen, B. Ervens, A. Nenes, C. J. Nielsen, E. Swietlicki, J. P. Putaud, Y. Balkanski, S. Fuzzi, J. Horth, G. K. Moortgat, R. Winterhalter, C. E. L. Myhre, K. Tsigaridis, E. Vignati, E. G. Stephanou, J. Wilson, *Atmospheric Chemistry and Physics* 2005, **5**, 1053.

26 U. Pöschl, *Analytical and Bioanalytical Chemistry* 2003, **375**, 30.

27 K. H. Homann, *Angewandte Chemie International Edition* 1998, **37**, 2435.

28 A. Sadezky, H. Muckenhuber, H. Grothe, R. Niessner, U. Pöschl, *Carbon* 2005,

29 N. P. Ivleva, A. K. Messerer, X. Yang, R. Niessner, U. Pöschl, Environmental Science and Technology 2007, **41**, 3702.

30 S. Szidat, T. M. Jenk, H. W. Gaggeler, H. A. Synal, R. Fisseha, U. Baltensperger, M. Kalberer, V. Samburova, L. Wacker,

M. Saurer, M. Schwikowski, I. Hajdas, *Radiocarbon* 2004, **46**, 475.
31 C. Schauer, R. Niessner, U. Pöschl, *Analytical and Bioanalytical Chemistry* 2004, **378**, 725.
32 T. Franze, *Analyse und Reaktivität von Proteinen in atmosphärischen Aerosolen und Enwicklung neuer Immunoassays zur Messung von Nitroproteinen*, PhD Thesis, Technical University Munich, 2004.
33 J. Hendricks, B. Karcher, A. Dopelheuer, J. Feichter, U. Lohmann, D. Baumgardner, *Atmospheric Chemistry and Physics* 2004, **4**, 2521.
34 A. Kirkevag, T. Iversen, A. Dahlback, *Atmospheric Environment. Aug* 1999, **33**, 2621.
35 M. Z. Jacobson, *Journal of Geophysical Research* 2002, **107**, 4410.
36 H. Schmid, L. Laskus, H. J. Abraham, U. Baltensperger, V. Lavanchy, M. Bizjak, P. Burba, H. Cachier, D. Crow, J. Chow, T. Gnauk, A. Even, H. M. ten Brink, K. P. Giesen, R. Hitzenberger, E. Hueglin, W. Maenhaut, C. Pio, A. Carvalho, J. P. Putaud, D. Toom-Sauntry, H. Puxbaum, *Atmospheric Environment* 2001, **35**, 2111.
37 R. Hitzenberger, S. G. Jennings, S. M. Larson, A. Dillner, H. Cachier, Z. Galambos, A. Rouc, T. G. Spain, *Atmospheric Environment* 1999, **33**, 2823.
38 K. Wittmaack, *Atmospheric Chemistry and Physics* 2005, **5**, 1905.
39 A. Petzold, H. Schloesser, P. J. Sheridan, W. P. Arnott, J. A. Ogren, A. Virkkula, *Aerosol Science and Technology* 2005, **39**, 40.
40 V. M. H. Lavanchy, H. W. Gaggeler, S. Nyeki, U. Baltensperger, *Atmospheric Environment* 1999, **33**, 2759.
41 H. Saathoff, K. H. Naumann, M. Schnaiter, W. Schock, E. Weingartner, U. Baltensperger, L. Kramer, Z. Bozoki, U. Poschl, R. Niessner, U. Schurath, *Journal of Aerosol Science* 2003, **34**, 1399.
42 W. P. Arnott, H. Moosmuller, C. F. Rogers, T. F. Jin, R. Bruch, *Atmospheric Environment* 1999, **33**, 2845.
43 P. J. Sheridan, W. P. Arnott, J. A. Ogren, E. Andrews, D. B. Atkinson, D. S. Covert, H. Moosmuller, A. Petzold, B. Schmid, A. W. Strawa, R. Varma, A. Virkkula, *Aerosol Science and Technology* 2005, **39**, 1.
44 D. S. Su, R. E. Jentoft, J. O. Muller, D. Rothe, E. Jacob, C. D. Simpson, Z. Tomovic, K. Mullen, A. Messerer, U. Poschl, R. Niessner, R. Schlogl, *Catalysis Today* 2004, **90**, 127.
45 S. Mertes, B. Dippel, A. Schwarzenbock, *Journal of Aerosol Science* 2004, **35**, 347.
46 A. Stratmann, G. Schweiger, *J. Aerosol Sci.* 2005, submitted.
47 L. M. Russell, *Environmental Science and Technology* 2003, **37**, 2982.
48 U. Pöschl, *Journal of Aerosol Medicine-Deposition Clearance and Effects in the Lung* 2002, **15**, 203.
49 M. Kulmala, L. Laakso, K. E. J. Lehtinen, I. Riipinen, M. Dal Maso, T. Anttila, V. M. Kerminen, U. Horrak, M. Vana, H. Tammet, *Atmospheric Chemistry and Physics* 2004, **4**, 2553.
50 M. Kulmala, H. Vehkamaki, T. Petajda, M. Dal Maso, A. Lauri, V. M. Kerminen, W. Birmili, P. H. McMurry, *Journal of Aerosol Science* 2004, **35**, 143.
51 T. Anttila, V. M. Kerminen, M. Kulmala, A. Laaksonen, C. D. O'Dowd, *Atmospheric Chemistry and Physics* 2004, **4**, 1071.
52 R. Y. Zhang, I. Suh, J. Zhao, D. Zhang, E. C. Fortner, X. X. Tie, L. T. Molina, M. J. Molina, *Science* 2004, **304**, 1487.
53 L. Laakso, T. Anttila, K. E. J. Lehtinen, P. P. Aalto, M. Kulmala, U. Horrak, J. Paatero, M. Hanke, F. Arnold, *Atmospheric Chemistry and Physics* 2004, **4**, 2353.
54 U. Pöschl, Y. Rudich, M. Ammann, *Atmospheric Chemistry and Physics Discussions* 2005, **5**, 2111.
55 M. P. Tolocka, M. Jang, J. M. Ginter, F. J. Cox, R. M. Kamens, M. V. Johnston, *Environmental Science and Technology* 2004, **38**, 1428.
56 M. S. Jang, N. M. Czoschke, S. Lee, R. M. Kamens, *Science* 2002, **298**, 814.
57 M. Kalberer, D. Paulsen, M. Sax, M. Steinbacher, J. Dommen, A. S. H. Prevot, R. Fisseha, E. Weingartner, V. Frankevich,

R. Zenobi, U. Baltensperger, *Science* 2004, **303**, 1659.
58 A. Limbeck, M. Kulmala, H. Puxbaum, *Geophysical Research Letters* 2003, **30**.
59 M. E. Jenkin, *Atmospheric Chemistry and Physics* 2004, **4**, 1741.
60 K. Tsigaridis, M. Kanakidou, *Atmospheric Chemistry and Physics* 2003, **3**, 1849.
61 C. Marcolli, B. P. Luo, T. Peter, F. G. Wienhold, *Atmospheric Chemistry and Physics* 2004, **4**, 2593.
62 M. C. Jacobson, H. C. Hansson, K. J. Noone, R. J. Charlson, *Reviews of Geophysics* 2000, **38**, 267.
63 A. H. Falkovich, E. R. Graber, G. Schkolnik, Y. Rudich, W. Maenhaut, P. Artaxo, *Atmospheric Chemistry and Physics* 2005, **5**, 781.
64 B. Graham, O. L. Mayol-Bracero, P. Guyon, G. C. Roberts, S. Decesari, M. C. Facchini, P. Artaxo, W. Maenhaut, P. Koll, M. O. Andreae, *Journal of Geophysical Research – Atmospheres* 2002, 107.
65 O. L. Mayol-Bracero, P. Guyon, B. Graham, G. Roberts, M. O. Andreae, S. Decesari, M. C. Facchini, S. Fuzzi, P. Artaxo, *Journal of Geophysical Research – Atmospheres* 2002, **107.**
66 J. F. Hamilton, P. J. Webb, A. C. Lewis, J. R. Hopkins, S. Smith, P. Davy, *Atmospheric Chemistry and Physics* 2004, **4**, 1279.
67 M. Claeys, B. Graham, G. Vas, W. Wang, R. Vermeylen, V. Pashynska, J. Cafmeyer, P. Guyon, M. O. Andreae, P. Artaxo, W. Maenhaut, *Science* 2004, **303**, 1173.
68 S. Zappoli, A. Andracchio, S. Fuzzi, M. C. Facchini, A. Gelencser, G. Kiss, Z. Krivacsy, A. Molnar, E. Meszaros, H. C. Hansson, K. Rosman, Y. Zebuhr, *Atmospheric Environment. Aug* 1999, **33**, 2733.
69 Q. Zhang, C. Anastasio, *Atmospheric Environment* 2003, **37**, 2247.
70 H. Puxbaum, M. Tenze-Kunit, *Atmospheric Environment* 2003, **37**, 3693.
71 N. Havers, P. Burba, J. Lambert, D. Klockow, *Journal of Atmospheric Chemistry* 1998, **29**, 45.
72 T. Franze, M. G. Weller, R. Niessner, U. Pöschl, *Environmental Science and Technology* 2005, **39**, 1673.
73 T. Franze, M. G. Weller, R. Niessner, U. Pöschl, *Analyst* 2003, **128**, 824.
74 T. Franze, M. G. Weller, R. Niessner, U. Pöschl, *Analyst* 2004, **129**, 589.
75 W. Walcher, T. Franze, M. G. Weller, U. Pöschl, C. G. Huber, *Journal of Proteome Research* 2003, **2**, 534.
76 G. Schkolnik, A. H. Falkovich, Y. Rudich, W. Maenhaut, P. Artaxo, *Environmental Science and Technology* 2005, **39**, 2744.
77 P. H. McMurry, *Atmospheric Environment* 2000, **34**, 1959.
78 C. A. Noble, K. A. Prather, *Mass Spectrometry Reviews* 2000, **19**, 248.
79 H. J. Tobias, P. M. Kooiman, K. S. Docherty, P. J. Ziemann, *Aerosol Science and Technology* 2000, **33**, 170.
80 D. Y. H. Pui, R. C. Flagan, S. L. Kaufman, A. D. Maynard, J. F. de la Mora, S. V. Hering, J. L. Jimenez, K. A. Prather, A. S. Wexler, P. J. Ziemann, *Journal of Nanoparticle Research* 2004, **6**, 314.
81 S. F. Maria, L. M. Russell, M. K. Gilles, S. C. B. Myneni, *Science* 2004, **306**, 1921.
82 M. R. Alfarra, H. Coe, J. D. Allan, K. N. Bower, H. Boudries, M. R. Canagaratna, J. L. Jimenez, J. T. Jayne, A. A. Garforth, S. M. Li, D. R. Worsnop, *Atmospheric Environment* 2004, **38**, 5745.
83 R. Atkinson, D. L. Baulch, R. A. Cox, R. F. Hampson, J. A. Kerr, M. J. Rossi, J. Troe, *Journal of Physical and Chemical Reference Data* 1997, **26**, 1329.
84 M. Ammann, U. Poschl, Y. Rudich, *Physical Chemistry Chemical Physics* 2003, **5**, 351.
85 J. T. Jayne, U. Poschl, Y. M. Chen, D. Dai, L. T. Molina, D. R. Worsnop, C. E. Kolb, M. J. Molina, *Journal of Physical Chemistry A* 1997, **101**, 10000.
86 U. Pöschl, M. Canagaratna, J. T. Jayne, L. T. Molina, D. R. Worsnop, C. E. Kolb, M. J. Molina, *Journal of Physical Chemistry a* 1998, **102**, 10082.

87 A. R. Ravishankara, *Science* 1997, **276**, 1058.

88 J. P. Reid, R. M. Sayer, *Chemical Society Reviews* 2003, **32**, 70.

89 Y. Rudich, *Chemical Reviews* 2003, **103**, 5097.

90 U. Pöschl, T. Letzel, C. Schauer, R. Niessner, *Journal of Physical Chemistry A* 2001, **105**, 4029.

91 R. Sander, *Surveys in Geophysics* 1999, **20**, 1.

92 S. P. Sander, B. J. Finlayson-Pitts, R. R. Friedl, D. M. Golden, R. E. Huie, C. E. Kolb, M. J. Kurylo, M. J. Molina, G. K. Moortgat, V. L. Orkin, A. R. Ravishankara, *Chemical Kinetics and Photochemical Data for Use in Atmospheric Studies, Evaluation Number 14. JPL Publication 02-25*, Jet Propulsion Laboratory, Pasadena, CA, 2002.

93 A. E. Waibel, T. Peter, K. S. Carslaw, H. Oelhaf, G. Wetzel, P. J. Crutzen, U. Pöschl, A. Tsias, E. Reimer, H. Fischer, *Science* 1999, **283**, 2064.

94 D. J. Stewart, P. T. Griffiths, R. A. Cox, *Atmospheric Chemistry and Physics* 2004, **4**, 1381.

95 J. Austin, D. Shindell, S. R. Beagley, C. Bruhl, M. Dameris, E. Manzini, T. Nagashima, P. Newman, S. Pawson, G. Pitari, E. Rozanov, C. Schnadt, T. G. Shepherd, *Atmospheric Chemistry and Physics* 2003, **3**, 1.

96 S. K. Meilinger, B. Karcher, T. Peter, *Atmospheric Chemistry and Physics* 2002, **2**, 307.

97 S. K. Meilinger, B. Karcher, T. Peter, *Atmospheric Chemistry and Physics* 2005, **5**, 533.

98 M. O. Andreae, P. J. Crutzen, *Science* 1997, **276**, 1052.

99 R. Sander, W. C. Keene, A. A. P. Pszenny, R. Arimoto, G. P. Ayers, E. Baboukas, J. M. Cainey, P. J. Crutzen, R. A. Duce, G. Honninger, B. J. Huebert, W. Maenhaut, N. Mihalopoulos, V. C. Turekian, R. Van Dingenen, *Atmospheric Chemistry and Physics* 2003, **3**, 1301.

100 A. A. P. Pszenny, J. Moldanov, W. C. Keene, R. Sander, J. R. Maben, M. Martinez, P. J. Crutzen, D. Perner, R. G. Prinn, *Atmospheric Chemistry and Physics* 2004, **4**, 147.

101 R. C. Sullivan, T. Thornberry, J. P. D. Abbatt, *Atmospheric Chemistry and Physics* 2004, **4**, 1301.

102 S. Tilmes, R. Muller, J. U. Grooss, J. M. Russell, *Atmospheric Chemistry and Physics* 2004, **4**, 2181.

103 J.-U. Grooß, G. Günther, R. Müller, P. Konopka, S. Bausch, H. Schlager, C. Voigt, C. M. Volk, G. C. Toon, *Atmospheric Chemistry and Physics* 2005, **5**, 1437.

104 K. Broekhuizen, P. P. Kumar, J. P. D. Abbatt, *Geophysical Research Letters* 2004, **31**, 1107.

105 A. Asad, B. T. Mmereki, D. J. Donaldson, *Atmospheric Chemistry and Physics* 2004, **4**, 2083.

106 A. Gelencser, A. Hoffer, G. Kiss, E. Tombacz, R. Kurdi, L. Bencze, *Journal of Atmospheric Chemistry* 2003, **45**, 25.

107 P. P. Kumar, K. Broekhuizen, J. P. D. Abbatt, *Atmospheric Chemistry and Physics* 2003, **3**, 509.

108 T. Moise, Y. Rudich, *Journal of Physical Chemistry A* 2002, **106**, 6469.

109 N. O. A. Kwamena, J. A. Thornton, J. P. D. Abbatt, *Journal of Physical Chemistry A* 2004, **108**, 11626.

110 C. Schauer, R. Niessner, U. Pöschl, *Environmental Science and Technology* 2003, **37**, 2861.

111 A. Messerer, R. Niessner, U. Poschl, *Journal of Aerosol Science* 2003, **34**, 1009.

112 A. Messerer, H. J. Schmid, C. Knab, U. Poschl, R. Niessner, *Chemie Ingenieur Technik* 2004, **76**, 1092.

113 A. Messerer, D. Rothe, U. Poschl, R. Niessner, *Topics in Catalysis* 2004, **30–31**, 247.

114 M. Ammann, U. Pöschl, *Atmospheric Chemistry and Physics Discussions* 2005, **5**, 2193.

115 R. Atkinson, D. L. Baulch, R. A. Cox, J. N. Crowley, R. F. Hampson, R. G. Hynes, M. E. Jenkin, M. J. Rossi, J. Troe, *Atmospheric Chemistry and Physics* 2004, **4**, 1461.

116 M. E. Jenkin, S. M. Saunders, V. Wagner, M. J. Pilling, *Atmospheric Chemistry and Physics* 2003, **3**, 181.

117 S. M. Saunders, M. E. Jenkin, R. G. Derwent, M. J. Pilling, *Atmospheric Chemistry and Physics* 2003, **3**, 161.

118 U. Pöschl, R. von Kuhlmann, N. Poisson, P. J. Crutzen, *Journal of Atmospheric Chemistry* 2000, **37**, 29.

119 R. von Kuhlmann, M. G. Lawrence, U. Poschl, P. J. Crutzen, *Atmospheric Chemistry and Physics* 2004, **4**, 1.

120 B. Ervens, C. George, J. E. Williams, G. V. Buxton, G. A. Salmon, M. Bydder, F. Wilkinson, F. Dentener, P. Mirabel, R. Wolke, H. Herrmann, *Journal of Geophysical Research – Atmospheres* 2003, **108.**

121 R. Sander, A. Kerkweg, P. Jockel, J. Lelieveld, *Atmospheric Chemistry and Physics* 2005, **5**, 445.

122 M. J. Molina, A. V. Ivanov, S. Trakhtenberg, L. T. Molina, *Geophysical Research Letters* 2004, **31.**

123 Y. Katrib, S. T. Martin, H. M. Hung, Y. Rudich, H. Z. Zhang, J. G. Slowik, P. Davidovits, J. T. Jayne, D. R. Worsnop, *Journal of Physical Chemistry A* 2004, **108**, 6686.

124 Y. Katrib, S. T. Martin, Y. Rudich, P. Davidovits, J. T. Jayne, D. R. Worsnop, *Atmospheric Chemistry and Physics* 2005, **5**, 275.

125 A. Messerer, D. Rothe, R. Niessner, U. Pöschl, *Chemie Ingenieur Technik* 2005, **77**, 881.

126 P. Spichtinger, K. Gierens, H. G. J. Smit, J. Ovarlez, J. F. Gayet, *Atmospheric Chemistry and Physics* 2004, **4**, 639.

127 P. Spichtinger, K. Gierens, H. Wernli, *Atmospheric Chemistry and Physics* 2005, **5**, 973.

128 P. Spichtinger, K. Gierens, A. Dornbrack, *Atmospheric Chemistry and Physics* 2005, **5**, 1243.

129 E. J. Jensen, J. B. Smith, L. Pfister, J. V. Pittman, E. M. Weinstock, D. S. Sayres, R. L. Herman, R. F. Troy, K. Rosenlof, T. L. Thompson, A. M. Fridlind, P. K. Hudson, D. J. Cziczo, A. J. Heymsfield, C. Schmitt, J. C. Wilson, *Atmospheric Chemistry and Physics* 2005, **5**, 851.

130 B. Kärcher, T. Koop, *Atmospheric Chemistry and Physics* 2005, **5**, 703.

131 K. Gierens, *Atmospheric Chemistry and Physics* 2003, **3**, 437.

132 B. Svenningsson, J. Rissler, E. Swietlicki, M. Mircea, M. Bilde, M. C. Facchini, S. Decesari, S. Fuzzi, J. Zhou, J. Mønster, T. Rosenørn, *Atmospheric Chemistry and Physics Discussions* 2005, **5**, 2833.

133 A. M. L. Ekman, C. Wang, J. Wilson, J. Strom, *Atmospheric Chemistry and Physics* 2004, **4**, 773.

134 T. J. Fortin, K. Drdla, L. T. Iraci, M. A. Tolbert, *Atmospheric Chemistry and Physics* 2003, **3**, 987.

135 C. M. Archuleta, P. J. DeMott, S. M. Kreidenweis, *Atmospheric Chemistry and Physics Discussions* 2005, **5**, 3391.

136 C. H. Twohy, M. R. Poellot, *Atmospheric Chemistry and Physics Discussions* 2005, **5**, 3723.

137 M. Seifert, J. Strom, R. Krejci, A. Minikin, A. Petzold, J. F. Gayet, U. Schumann, J. Ovarlez, *Atmospheric Chemistry and Physics* 2003, **3**, 1037.

138 J. Strom, M. Seifert, B. Karcher, J. Ovarlez, A. Minikin, J. F. Gayet, R. Krejci, A. Petzold, F. Auriol, W. Haag, R. Busen, U. Schumann, H. C. Hansson, *Atmospheric Chemistry and Physics* 2003, **3**, 1807.

139 M. Seifert, J. Strom, R. Krejci, A. Minikin, A. Petzold, J. F. Gayet, H. Schlager, H. Ziereis, U. Schumann, J. Ovarlez, *Atmospheric Chemistry and Physics* 2004, **4**, 1343.

140 M. Seifert, J. Strom, R. Krejci, A. Minikin, A. Petzold, J. F. Gayet, H. Schlager, H. Ziereis, U. Schumann, J. Ovarlez, *Atmospheric Chemistry and Physics* 2004, **4**, 293.

141 E. Mikhailov, S. Vlasenko, R. Niessner, U. Pöschl, *Atmospheric Chemistry and Physics* 2004, **4**, 323.

142 S. M. Kreidenweis, K. Koehler, P. J. DeMott, A. J. Prenni, C. Carrico,

B. Ervens, *Atmospheric Chemistry and Physics* 2005, **5**, 1357.

143 D. O. Topping, G. B. McFiggans, H. Coe, *Atmospheric Chemistry and Physics* 2005, **5**, 1205.

144 D. O. Topping, G. B. McFiggans, H. Coe, *Atmospheric Chemistry and Physics* 2005, **5**, 1223.

145 C. Marcolli, T. Peter, *Atmospheric Chemistry and Physics* 2005, **5**, 1545.

146 T. Raatikainen, A. Laaksonen, *Atmospheric Chemistry and Physics Discussions* 2005, **5**, 3641.

147 C. D. O'Dowd, M. C. Facchini, F. Cavalli, D. Ceburnis, M. Mircea, S. Decesari, S. Fuzzi, Y. J. Yoon, J. P. Putaud, *Nature* 2004, **431**, 676.

148 S. T. Martin, H. M. Hung, R. J. Park, D. J. Jacob, R. J. D. Spurr, K. V. Chance, M. Chin, *Atmospheric Chemistry and Physics* 2004, **4**, 183.

149 M. Gysel, E. Weingartner, S. Nyeki, D. Paulsen, U. Baltensperger, I. Galambos, G. Kiss, *Atmospheric Chemistry and Physics* 2004, **4**, 35.

150 D. V. Spracklen, K. J. Pringle, K. S. Carslaw, M. P. Chipperfield, G. W. Mann, *Atmospheric Chemistry and Physics Discussions* 2005, **5**, 179.

151 D. V. Spracklen, K. J. Pringle, K. S. Carslaw, M. P. Chipperfield, G. W. Mann, *Atmospheric Chemistry and Physics Discussions* 2005, **5**, 3437.

152 C. E. L. Myhre, C. J. Nielsen, *Atmospheric Chemistry and Physics* 2004, **4**, 1759.

153 A. Massling, A. Wiedensohler, B. Busch, C. Neususs, P. Quinn, T. Bates, D. Covert, *Atmospheric Chemistry and Physics* 2003, **3**, 1377.

154 J. S. Reid, R. Koppmann, T. F. Eck, D. P. Eleuterio, *Atmospheric Chemistry and Physics* 2005, **5**, 799.

155 J. S. Reid, T. F. Eck, S. A. Christopher, R. Koppmann, O. Dubovik, D. P. Eleuterio, B. N. Holben, E. A. Reid, J. Zhang, *Atmospheric Chemistry and Physics* 2005, **5**, 827.

156 C. Michel, C. Liousse, J. M. Gregoire, K. Tansey, G. R. Carmichael, J. H. Woo, *Journal of Geophysical Research – Atmospheres* 2005, **110.**

157 F. Cousin, C. Liousse, H. Cachier, B. Bessagnet, B. Guillaume, R. Rosset, *Atmospheric Environment* 2005, **39**, 1539.

158 M. Z. Jacobson, *Journal of Geophysical Research – Atmospheres* 2004, **109.**

159 D. G. Streets, T. C. Bond, T. Lee, C. Jang, *Journal of Geophysical Research-Atmospheres* 2004, **109.**

160 T. C. Bond, D. G. Streets, K. F. Yarber, S. M. Nelson, J. H. Woo, Z. Klimont, *Journal of Geophysical Research-Atmospheres* 2004, **109.**

161 W. J. Gauderman, E. Avol, F. Gilliland, H. Vora, D. Thomas, K. Berhane, R. McConnell, N. Kuenzli, F. Lurmann, E. Rappaport, H. Margolis, D. Bates, J. Peters, *New England Journal of Medicine* 2004, **351**, 1057.

162 K. Katsouyanni, G. Touloumi, E. Samoli, A. Gryparis, A. Le Tertre, Y. Monopolis, G. Rossi, D. Zmirou, F. Ballester, A. Boumghar, H. R. Anderson, B. Wojtyniak, A. Paldy, R. Braunstein, J. Pekkanen, C. Schindler, *J. Schwartz, Epidemiology* 2001, **12**, 521.

163 C. A. Pope, R. T. Burnett, G. D. Thurston, M. J. Thun, E. E. Calle, D. Krewski, J. J. Godleski, *Circulation* 2004, **109**, 71.

164 J. Samet, R. Wassel, K. J. Holmes, E. Abt, K. Bakshi, *Environmental Science and Technology* 2005, **39**, 209A.

165 H. Bömmel, M. Haake, P. Luft, J. Horejs-Hoeck, H. Hein, J. Bartels, C. Schauer, U. Pöschl, M. Kracht, A. Duschl, *International Immunopharmacology* 2003, **3**, 1371.

166 K. Donaldson, V. Stone, P. J. A. Borm, L. A. Jimenez, P. S. Gilmour, R. P. F. Schins, A. M. Knaapen, I. Rahman, S. P. Faux, D. M. Brown, W. MacNee, *Free Radical Biology and Medicine* 2003, **34**, 1369.

167 R. P. F. Schins, J. H. Lightbody, P. J. A. Borm, T. M. Shi, K. Donaldson, V. Stone, *Toxicology and Applied Pharmacology* 2004, **195**, 1.

168 G. Oberdörster, Z. Sharp, V. Atudorei, A. Elder, R. Gelein, W. Kreyling, C. Cox, *Inhalation Toxicology* 2004, **16**, 437.

169 A. Nemmar, P. H. M. Hoet, B. Vanquickenborne, D. Dinsdale, M. Thomeer, M. F. Hoylaerts, H. Vanbilloen, L. Mortelmans, B. Nemery, *Circulation* 2002, **105**, 411.

170 G. Oberdörster, E. Oberdörster, J. Oberdörster, *Environmental Health Perspectives* 2005, **113**, 823.

171 B. Brunekreef, J. Sunyer, *Eur. Respir. J.* 2003, **21**, 913.

172 J. Ring, U. Krämer, T. Schäfer, H. Behrendt, *Curr. Opin. Immunol.* 2001, **13**, 701.

173 J. Ring, B. Eberlein-Koenig, H. Behrendt, *Ann. Allergy Asthma Immunol.* 2001, **87**, 2.

174 A. G. Miguel, G. R. Cass, M. M. Glovsky, J. Weiss, *Environmental Science and Technology* 1999, **33**, 4159.

175 Y. K. Gruijthuijsen, I. Grieshuber, A. Stöcklinger, U. Tischler, T. Fehrenbach, M. G. Weller, L. Vogel, S. Vieths, U. Pöschl, A. Duschl, *International Archives of Allergy and Immunology* 2006, **141**, 265.

176 A. Ibald-Mulli, H. E. Wichmann, W. Kreyling, A. Peters, *Journal of Aerosol Medicine – Deposition Clearance and Effects in the Lung* 2002, **15**, 189.

177 A. Peters, H. E. Wichmann, T. Tuch, J. Heinrich, J. Heyder, *American Journal of Respiratory and Critical Care Medicine* 1997, **155**, 1376.

178 D. Rothe, F. I. Zuther, E. Jacob, A. Messerer, U. Pöschl, R. Niessner, C. Knab, M. Mangold, C. Mangold, *SAE Technical Papers* 2004, 3024.

179 O. Preining, J. E. Davis, *History of Aerosol Science*, Österreichische Akademie der Wissenschaften, Vienna, 2000.

8
Measurement and Detection of Nanoparticles Within the Environment

Thomas A.J. Kuhlbusch, Heinz Fissan, and Christof Asbach

8.1
Introduction

Nanotechnology opens up opportunities for new and improved products and needs tailored production tools. Nanoparticles are one of the most important building blocks, for example to develop new or improved materials, allow for higher catalytic efficiencies or reduce energy and material consumption.

Nanoparticles (NPs) are defined in this chapter as intentionally produced particles for use in products either as single particles or as agglomerates with diameters below 100 nm. Another term often used is ultrafine particles (UFPs, Figure 8.1). Ultrafine particles in this chapter denote particles in the environment which at least partially consist of unintentionally produced and/or naturally formed particles. Nanoparticles are normally solid particles whereas naturally and unintentionally manmade particles may be of solid or liquid nature. Even though particles in the sub-100-nm range can be differentiated into nanoparticles and UFPs, the measurement and detection techniques are fundamentally the same.

Figure 8.2 shows the relative scale of natural and manmade materials. This figure shows that nanoparticles are about one hundredth the size of a blood cell. The small size of nanoparticles is actually one of the important characteristics which give the free, unbound particles a generally high mobility. Particles of sizes below 100 nm are within the size of the pores of cells and hence may penetrate into cells, the blood or even end organs such as the brain [1]. Whether this may happen or is of relevance to health is still under debate and not a topic of this chapter.

Nanoparticles are mainly defined by their particle diameter being <100 nm. Other specific properties related to nanoparticles and possible concentration metrics are listed in Table 8.1.

The overview of possible particle properties of interest in view of nanoparticle effects on humans, animals and plants in Table 8.1 is not exclusive and only indicates the complexity with which nanoparticles may interact with the environment. The list of concentration metrics (Table 8.1) again is not exclusive but indicates that it is

Nanotechnology. Volume 2: Environmental Aspects. Edited by Harald Krug

ISBN: 978-3-527-31735-6

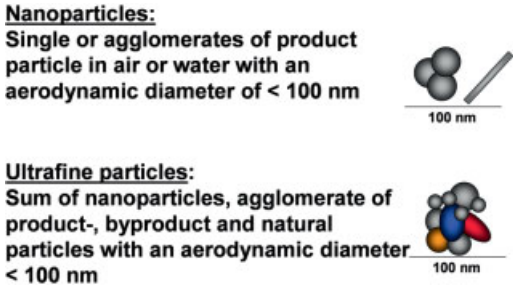


Figure 8.1 Particle definitions.

possible to use different metrics for the quantitative investigation of exposure, dose and effects. More detailed information on how the above properties are related to particle-induced negative health effects and the detection of particles in different tissues is given elsewhere [2–4].

The specific properties listed in Table 8.1 may also be used for the measurement and detection of nanoparticles in the environment.

Another issue illustrated in Figure 8.2 also is that natural and manmade nanosized particles co-exist. The properties listed in Table 8.1 may be used to determine, measure and quantify nanoparticles, but the methods generally do not differentiate manmade nanoparticle and other nanosized particles. One of the major tasks of measuring and detecting nanoparticles in the environment is the differentiation between UFPs and nanoparticles. A simple differentiation may even not be enough since a nanoparticle being attached to a larger particle will lose some of its properties, especially its size and hence its mobility. Therefore, generally three kinds of particles below 100 nm have to be differentiated.

- nanoparticles and their agglomerates (type 1)
- by-product and unintentionally produced (type 2), in addition to
- natural particles (type 3)
- mixtures of type 1 nanoparticles with type 2 or type 3 particles.

One further issue when discussing nanoparticles in the environment is the stability of nanoparticle agglomerates and agglomerates of nanoparticles with type 2 or 3 particles being larger than 100 nm. The nanoparticles will have lost their

Table 8.1 Possible nanoparticle properties and concentration metrics of interest.

Particle property	Concentration metric
Shape/morphology	Mass concentration
Chemical composition	Surface area concentration
Hygroscopicity	Number concentration
Solubility	Size distribution <100 nm
Charge, mobility	
Particle reactivity (radical formation)	

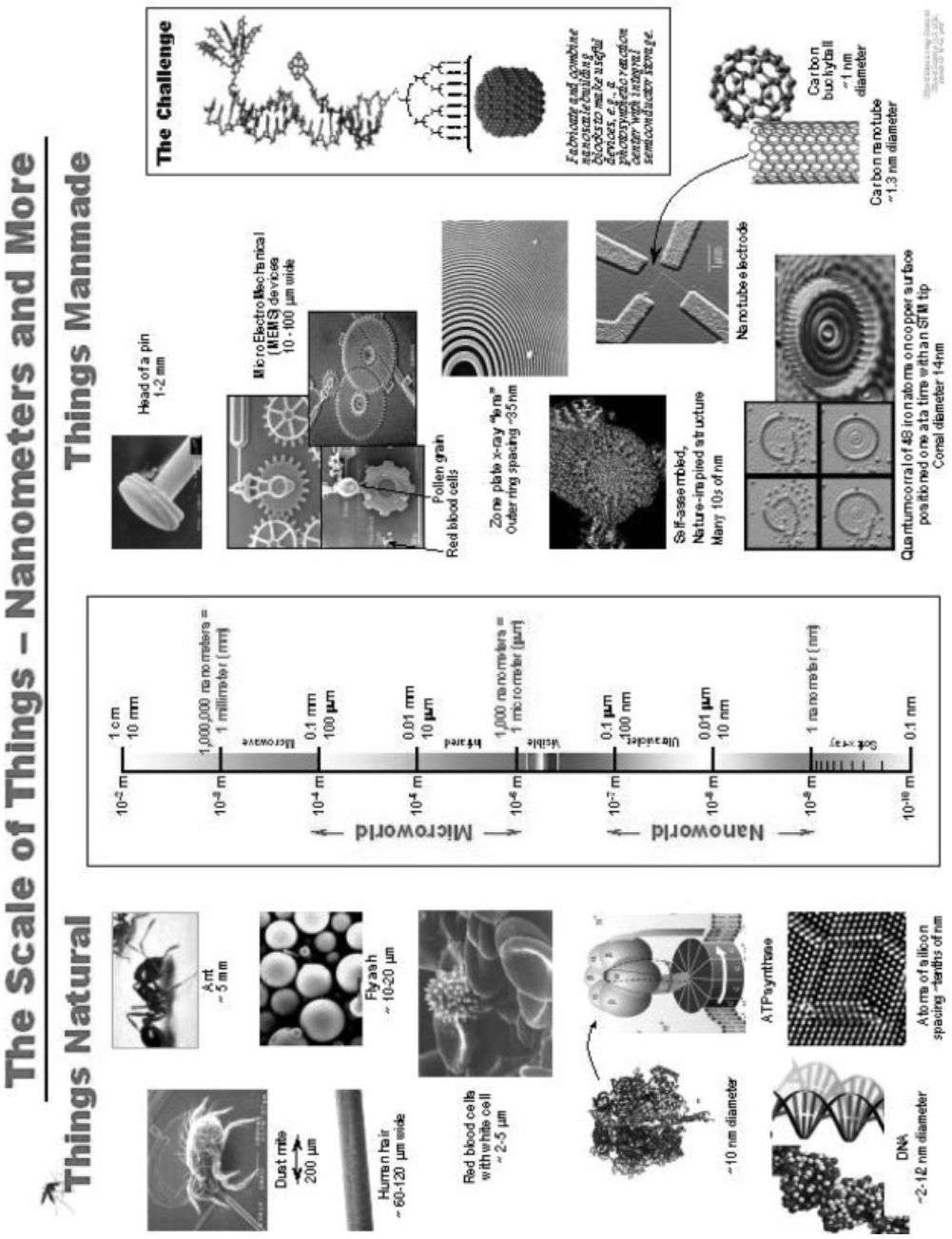


Figure 8.2 Relative scales from centimeter to nanometer (From NNI webpage. Courtesy Office of Basic Energy Sciences, Office of Science, US Department of Energy).

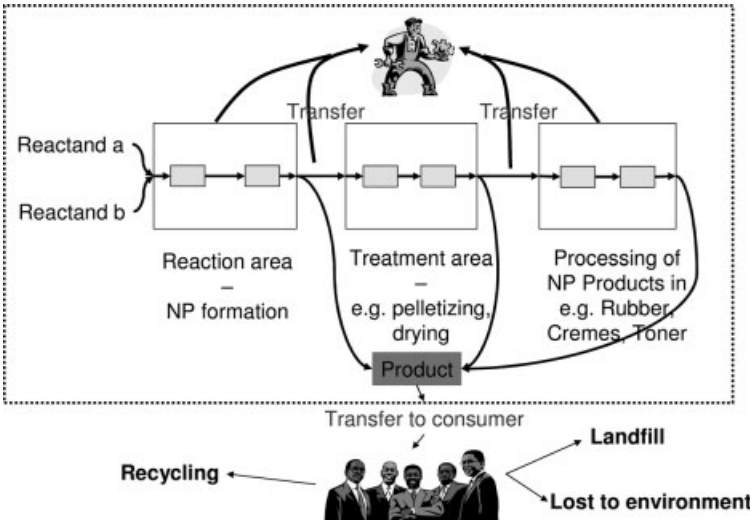


Figure 8.3 The lifecycle of nanoparticles [4].

specific nanoparticle properties but may recover some of their properties when released from the agglomerate after uptake in plants, animals or humans. This issue being specifically related to toxicology will not be further discussed here but leads to the discussion of the measurement techniques that may also include particle sizes up to about 400 nm. Still, it should be noted that to our knowledge no release of nanoparticles from agglomerates has yet been demonstrated under conditions after, for example, uptake of agglomerates in lungs.

A further point of discussion related to the detection and measurement of nanoparticles in the environment is shown in Figure 8.3. This figure shows the lifecycle of nanoparticles from production, handling and processing to the stage where they are used in products, for example for consumers, and finally to recycling, deposition in landfills or other processes. The life cycle is shown to indicate possible release of nanoparticles into the environment. They may be released during the production, handling (e.g. packing), handling during nanoparticle processing (e.g. use of carbon black nanoparticles during tire production) or use of products (e.g. paint containing nanoparticles sprayed onto surfaces).

Emissions from products containing bound nanoparticles in a fixed matrix such as carbon black in plastics of computers or in tires can generally be imagined but are not likely and cannot yet be detected. Hence the release of nanoparticles from products containing nanoparticles in a fixed matrix can be neglected, as was announced, for example, in Californian Proposition 65 [5] for carbon black. Detection of nanoparticles in most of these fixed matrices in the environment is nearly impossible.

The environment and environmental matrix containing nanoparticles are also of crucial importance for the detection and measurement of nanoparticles. Generally three matrices can be differentiated in the environment: soils, water and air. These three matrices differ not only in their physical state (solid, liquid and gas) but also in their mobility, increasing from soils to air. This mobility and the

mobility of nanoparticles dissolved in these matrices are of great importance when assessing the risk.

Figure 8.3 also differentiates workplace and public environment exposure. Possible emissions and consequently higher concentrations of nanoparticles in the environmental matrices will more likely be in work and plant areas of nanoparticle production than in the public environment. Another difference between public and workplace environments are the possible temporal changes in nanoparticle concentrations. The variance over time will be higher closer to the sources, and hence in workplace and plant areas. Therefore, the workplace and public environments are discussed separately for some cases.

8.2 Occurrence of Nanoparticles in Environmental Media

Mainly two different environments can be differentiated when discussing nanoparticles: the ambient environment and plants or workplaces in nanoparticle production, handling and processing. The environmental media in which nanoparticles may occur are the same, but the media soil and water will only be discussed for the ambient, public environment since the information given there is also valid for the work environment.

8.2.1 Ambient Environment

8.2.1.1 Water and Soils

Nanoparticles may reach waters, either by intentional use for remediation [6] or after unintentional release during/after production and subsequent wash-off. nanoparticles may also be released to water when they are produced via the liquid phase and leaks occur or the effluent is not sufficiently cleaned up. Once suspended in the liquid phase, nanoparticles will likely attach to other particles or to surfaces such as those of sand grains. No information on transport distances and the spatial distribution of nanoparticles in waters is currently available, to our knowledge.

Soils may become contaminated by nanoparticles by airborne deposition, deposition from the water phase or by direct deposition of powders and fluids, either intentionally or unintentionally. No systematic differences are made between soils and sediments in this chapter, since no information on differences in the behavior and the determination of nanoparticles is currently available.

Once nanoparticles are attached to soil surfaces, three ways of possible (re)mobilization of nanoparticles can be differentiated: (a) transport by water through the soils to the groundwater, (b) uptake by plants via roots and (c) wind erosion of soil particles containing nanoparticles.

No studies investigating the uptake of nanoparticles by plants are currently available. Lecoanet and Wiesner [7] stated that nanosized particles do not move far under environmental conditions. It was shown that the mobility of nanoparticles

correlates with size, with smaller nanosized particles being easily adsorbed on surfaces of sand grains and therefore immobilized. Biological transport may still occur from ingested sediments, but the physical movement of nanosized materials is restricted by their small size and propensity to adsorb on surfaces.

Nevertheless, taking the example of fullerenes, it could be demonstrated that the mobility of nanoparticles in water and soils is also strongly dependent on their physical–chemical properties. The mobility and behavior of three different fullerene solutions and four different oxidized materials was shown to be highly variable when changing ionic strength and pH values [7, 8]. While common models of particle transport through porous media described the behavior of mineral nanoparticles fairly well, the behavior of the fullerenes could not be modeled. Especially the latter component was found in brains of artificially exposed fish, where it induced oxidative stress [9]. Fullerenes, on the other hand, showed the lowest mobility of the substances tested, reducing the likelihood of exposure. This example demonstrates the need for knowledge of the mobility of nanoparticles.

So far, no studies have been conducted, to our knowledge, to determine nanoparticles in environmental media outside plants or experimental sites. This lack of current knowledge may be due to the low concentrations of nanoparticles in waters and soils, which make detection and quantification nearly impossible. Still, with increasing use of mobile nanoparticles, it may become important and should not be neglected in the future.

8.2.1.2 **Air**

The main matrix studied for the transport of nanoparticles so far is the air. This focus on air has several reasons, all related to the implications and use of particles in general. Some of the most important uses and effects of particles include the industrial production of particles (carbon black, titanium dioxide, etc.), horizontal dispersion in the atmosphere [10], climatic implications [11], cloud nucleation [12, 13], transport of nutrients [14] and health effects [15].

Another important reason is the high mobility of nanoparticles and UFPs in air, leading to a wide dispersion of these particles in our atmosphere. UFP number concentrations in ambient air may vary from a few hundred to over 10^5 particles per cubic centimeter, depending on the distance to sources such as incomplete combustion [16], but are determined anywhere in the Earth's atmosphere.

These UFPs are not due to nanoparticles but are mainly due to unintentionally manmade emissions in addition to natural emission. The occurrence of UFPs long before the intentional production of nanoparticles, their high mobility and ubiquitous presence and their implications for the environment and health led to the development of various measurement devices and to innumerable studies.

8.2.2
Workplace Environment

No studies investigating nanoparticle concentrations in soils and/or waters in workplaces and plant areas are known to the authors. Up to now, only very few

studies, measuring airborne nanoparticles and UFPs in workplace environments, have been published [17–19]. Sources of these small particles have to be differentiated at workplaces when detecting and measuring nanoparticles. High concentrations of particles smaller than 100 nm are commonly detected in welding and soldering workplaces or other workplaces where diesel forklifts are running. However, these particles are not termed nanoparticles since they are not intentionally produced. These are unintentional byproducts of a process or a work activity and are therefore classified as UFPs.

Nanoparticle release may occur during the production, handling and/or processing of nanoparticles. First measurements in work areas showed no detectable release of nanoparticles during normal handling and processing [17]. Nevertheless, on one occasion a leak in the production line occurred and high concentrations of nanoparticles (>100 000 cm^{-3}) were released into the work environment [19].

8.3 Nanoparticle Detection and Measurement Techniques

8.3.1 Soil

Basically no specific detection techniques for nanoparticles in soils are known. The only applicable method is the preparation of soil samples for analysis by transmission electron microscopy (TEM) coupled with, for example, energy-dispersive X-ray (EDX) analysis for the detection of nanoparticles of known chemical composition.

8.3.2 Water and Liquids

Nanoparticles can be produced in water but the detection, identification and size measurements of nanoparticles in water are still not as advanced as for airborne particles. One generally feasible way of determining nanoparticles is the filtration of waters and fluids with subsequent analysis of the deposited particles by microscopy [TEM, scanning electron microscopy (SEM), atomic force microscopy (AFM)]. This off-line technique permits the determination of nanoparticles down to diameters of a few nanometers.

No on-line technique for measurements down to a few nanometers is currently available.

8.3.2.1 Coulter Counter

The Coulter principle (electrical sensing zone method) is a widely used method for particle size analysis in liquids and is the recommended limit test for particulate matter in large-volume solutions. It sizes and counts particles based on measurable changes in electrical resistance produced by nonconductive particles suspended in an electrolyte. Suspended particles pass through the sensing zone consisting of a small

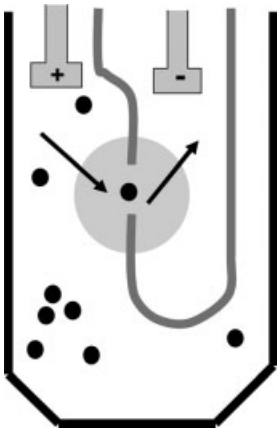


Figure 8.4 Principle setup of a Coulter Counter (Adapted from Beckman Coulter [52]).

opening (aperture) between the electrodes (Figure 8.4). Particles displace their own volume of electrolyte in the sensing zone, which is then measured as a voltage pulse. The height of each pulse is proportional to the particle volume and the number of pulses per unit time is proportional to the concentration during constant-volume flow. Several thousand particles per second can individually be counted and sized and this information is converted to particle size distributions. This method is according to its principle independent of particle shape, color and density. The detectable particle size range of the Coulter Counter extends from about 300 nm to 1200 μm, and hence not applicable to nanoparticles but to their larger agglomerates.

8.3.2.2 **Light Scattering**

Another principle in use for the detection of particles in liquids is based on light scattering. Gas molecules and atmospheric particles are smaller than the wavelengths of visible light. When light hits a gas molecule, the molecule absorbs and scatters the light in different directions. This is why a beam of a torch can be seen, for example, at night even from outside the light's path. The different colors of light are scattered differently after collision. The scattering is called Rayleigh scattering.

One of the scattering measurement techniques used for liquids is dynamic light scattering (DLS), which permits measurements of particle sizes from 2 nm to 6 μm. DLS measures the intensity fluctuations of scattered light by Brownian motion. The Brownian motion of the particles causes a Doppler shift in the incident light frequency. The magnitude of the frequency shift is related to the frequency of the Brownian motion, which on the other hand is directly related to the size of the particles.

The backscattered light (angle about 160–180° of incident light) is most commonly used for the detection and measurement of nanoparticles and their size distribution, since multiple scattering effects can be reduced and solutions with higher concentrations measured because the light does not have to go through the whole sample (Figure 8.5).

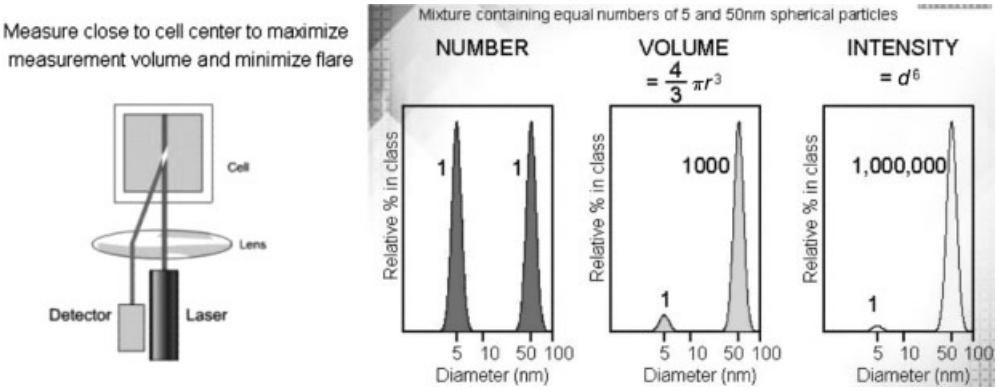


Figure 8.5 Instrumentation setup for the detection of nanoparticles in liquids by dynamic light scattering and example size distribution from Zetasizer Nano ZS90 (Rework, taken from Malvern Instruments).

A slightly different technique is also based on the illumination of particles. The viewing module of the instrument (Nanosight LM 10), in combination with a standard light microscope, uses an He–Ne laser light source to illuminate nanoscale particles. Enhanced by a near-perfect black background, particles appear individually as point scatterers, moving under Brownian motion. The tracking of individual nanoparticles and their Brownian motion in addition to the independently recorded scattering intensity is used for particle sizing by this instrument (diameter 15–500 nm). An example of the particle images is given in Figure 8.6.

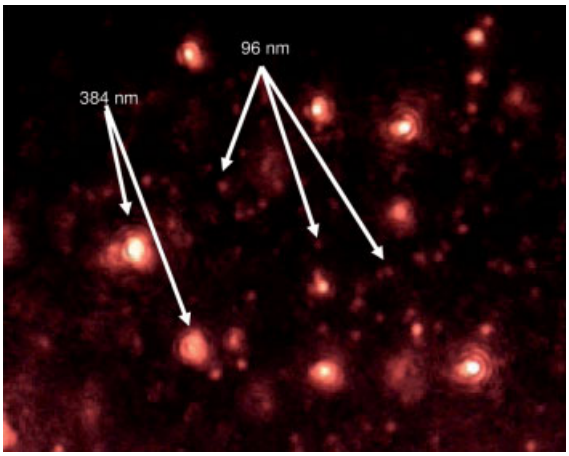


Figure 8.6 A bimodal system of 96- and 384-nm polystyrene reference spheres in water (picture: courtesy of Nanosight, UK).

8.3.3 Air

8.3.3.1 Basics

This section explains some of the basic particle and aerosol properties which are necessary to understand the measurement techniques discussed thereafter.

Mechanical Mobility The terminal settling velocity of a particle can be viewed as the velocity of a particle being released in still air undergoing gravitational settling (e.g. [20]). The terminal velocity can be calculated by a force balance, where the sum of drag force F_D and gravitational force F_G is zero:

$$F_D = F_G = mg \tag{8.1}$$

$$\frac{3\pi\eta V d}{C_C} = \frac{(\rho_p - \rho_g)\pi d^3 g}{6} \tag{8.2}$$

where η is the viscosity of air, V the particle velocity, d the particle diameter, ρ_p the particle density, ρ_g the density of the gas and g the acceleration relative to the gas due to gravity. C_C is the dimensionless Cunningham slip correction factor, which takes into account that the motion of submicrometer particles is affected by interaction with single molecules [21, 22]. C_C increases with decreasing particle diameter and approaches 1 for particle diameters above 1 µm:

$$C_C = 1 + \frac{2\lambda}{d_p}\left[\alpha + \beta \exp\left(-\frac{\gamma d_p}{2\lambda}\right)\right] \tag{8.3}$$

where λ is the mean free path of the gas molecules, $\alpha = 1.165$, $\beta = 0.483$ and $\gamma = 0.997$ are empirical constants [23].

Generally ρ_g is much smaller than ρ_p and hence can be neglected. Equation (8.2) can now be solved for the terminal settling velocity V_{TS} as used, for example, for the definition of equivalent diameters.

$$V_{TS} = \frac{\rho_p d^2 g C_C}{18\eta} \tag{8.4}$$

Stokes law, which describes the total resisting force on a spherical particle due to its velocity V relative to the fluid, can also be transformed to

$$F_D = \frac{3\pi\eta V d}{C_C} \Rightarrow B = \frac{V}{F_D} = \frac{C_C}{3\pi\eta d} \tag{8.5}$$

where B denotes the mechanical mobility of a particle. The equation now expresses the particle mobility or velocity per unit force B.

The terminal settling velocity V_{TS} [Equation (8.4)] can now be rewritten as

$$V_{TS} = F_G B \tag{8.6}$$

Equivalent Diameter The morphology of airborne particles can vary over a wide range, from spherical to needle-like in shape, from single particles to agglomerates. The latter denote small particles attached to each other by strong or weak bonds. Agglomerates usually exhibit irregular shapes with any fractal dimensions. This variance in morphology prevents easy comparisons and necessitates the use of particle models. The most commonly used models in aerosol measurement are based on equivalent spheres. One concept is the assumption of equal settling velocity of particles. In this model, each irregularly shaped particle is assigned an equivalent diameter of a sphere with the same terminal settling velocity as given in Equations (8.4) and (8.6). If the equivalent sphere is assumed to have the same density ρ_p as the irregularly shaped particle, it is referred to as the Stokes diameter d_{st}, whereas if unit density (1 g cm^{-3}) is assumed as density for the sphere, the diameter is referred to as the aerodynamic diameter d_{ae}.

This concept of equivalent diameters now allows the comparison of particles of different shape and density (Figure 8.7). Stokes and aerodynamic particle diameter can be exchanged using the following equation:

$$d_{ae} = d_{st}\sqrt{\frac{\rho_p}{\rho_0}} \tag{8.7}$$

Inertia (Impactor, Cyclone) One common principle used in aerosol measurements is the fractionation of particles according to their size by using the differences of the particle inertia. For the sake of simplicity, only impaction is discussed in more detail.

Impactors are used to remove particles of a given size from an aerosol; for example, impaction as a separation technique is used in environmental standards as a separator to remove particles larger than 10 µm d_{ae} or 2.5 µm d_{ae} (EN12341, EN14907) from the aerosol. The impaction principle is based on the acceleration of the aerosol in a nozzle and the direction of the outflow out of the nozzle onto a flat plate, called impaction plate. The flow is deflected by 90° by the impaction plate and particles above a certain size cannot follow the gas flow due to inertia and are deposited on the impaction plate. This is also exemplarily shown in Figure 8.8. It can be noted that the

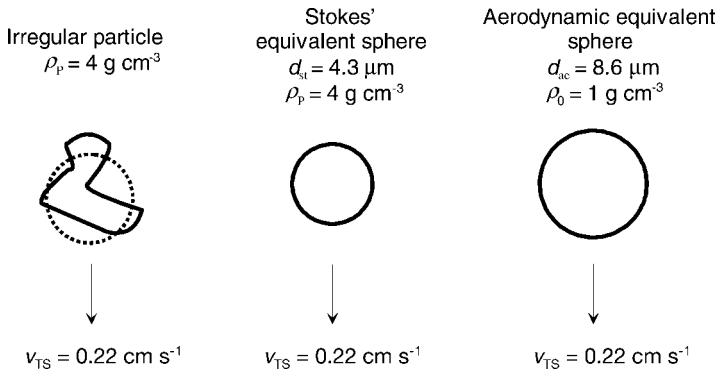


Figure 8.7 An irregular particle and its equivalent spheres (Adapted from [20]).

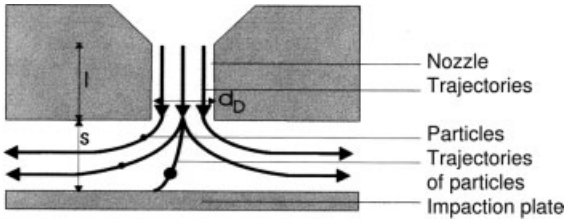


Figure 8.8 Schematic of an impactor (Adapted from [20]).

deposition of particles by impaction in the human lungs are in accordance to the aerodynamic particle equivalent diameter.

The cut-off sizes of any separating device, as for the impactor, is not a step function meaning that all particles >10 µm d_{ae} will be impacted and that all particles <10 µm d_{ae} will pass the impactor. The cut-off curve more or less has an S-shape (Figure 8.9) which is for example due to the finite dimensions of the impactor since the impaction probability is also dependent on whether the particle was released from the middle of the nozzle or the edge. The cut-off, also often called $d_{p,50}$, is therefore generally defined as the diameter at which 50% of the particles are deposited and 50% will pass the impactor.

The following equation can be used to calculate the cut-off size for a single-stage round nozzle impactor:

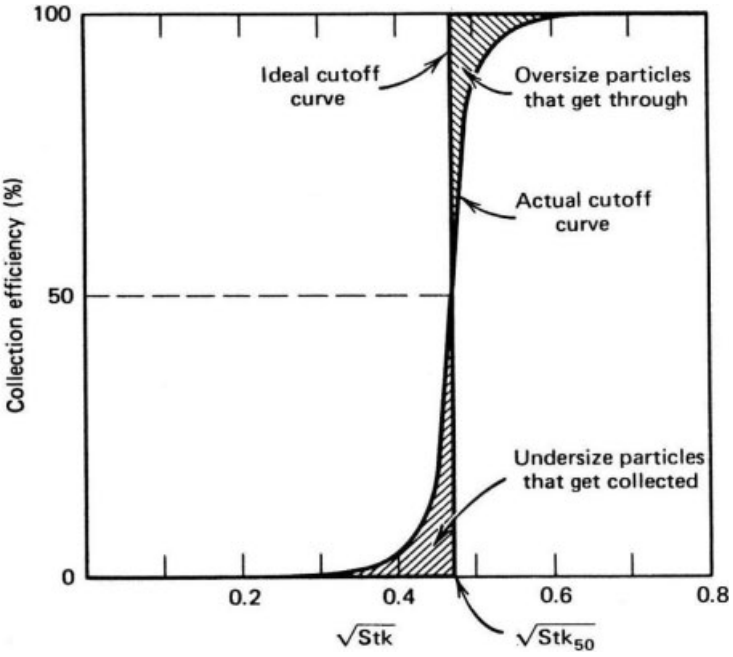


Figure 8.9 Collection efficiency curves (Adapted from [20]).

$$d_{P,50} = \sqrt{\frac{9\pi Stk_{50}\eta N d^3}{4\rho_P C_C \bar{V}}} \tag{8.8}$$

where η = viscosity of air, ρ_P = particle density, Stk_{50} = Stokes number, d = nozzle diameter, $\bar{V}$ = volumetric flow rate through impactor, N = number of nozzles and C_C = Cunningham correction factor (see Equation 8.3), also slip correction factor.

Equation (8.8) is only applicable within following limits:

- *Ratio of nozzle–impaction plate distance to nozzle diameter.* The ratio of the distance from the nozzle to the impaction plate (s) to the nozzle diameter (d_n) should be in the range $0.5 \leq s/d_n \leq 5.0$.
- *Ratio nozzle length to nozzle diameter.* The ratio of nozzle length (l) to nozzle diameter (d_D) should be in the range $0.25 \leq l/d_D \leq 2.0$. These limits ensure a homogeneous flow pattern at the exit of the nozzle, meaning the same velocity anywhere at the nozzle exit. The flow pattern will not have been well developed if the ratio is too small ($l/d_D < 0.25$) whereas distinct velocity profiles have developed with lower velocities at the edge compared with the middle of the nozzle if ratio is to large ($l/d_D > 2.0$).
- *Reynolds number.* The Reynolds number should be in the range 40–3000 to ensure laminar flow conditions.

Low-pressure conditions are necessary to allow sampling of nanoscale particles with the principle of impaction [24].

Electrical Mobility When an electric field acts upon a charged particle suspended in a gas, the particle becomes accelerated until it reaches its terminal electric migration velocity. The terminal velocity is a characteristic value for particles within a given electric field strength, which is exploited in several measurement techniques [25], such as in electrostatic classification. The ratio of the terminal migration velocity v_e and the electric field strength E is usually referred to as the 'electrical mobility' Z_p of a particle [20]:

$$Z_p = \frac{v_e}{E} \tag{8.9}$$

The electrical mobility thus expresses the ability of a charged particle to move within an electric field. This ability is dependent upon the particle charge, expressed as multiples n of the elementary charge e (1.6×10^{-19} A s) and the mechanical mobility B of the particle which represents the general ability of the particle to move in a gas:

$$Z_p = neB \tag{8.10}$$

The mechanical mobility B is a function of particle and gas properties and is explained in (8.5).

Particle Size Distributions (Mass, Surface, Number) One basic piece of information on airborne particles is their size distribution. Three modes can generally be differentiated: the nucleation mode (diameter <30 nm), accumulation mode (diameter 200–700 nm) and the coarse mode (diameter >1000 nm). The nucleation mode is

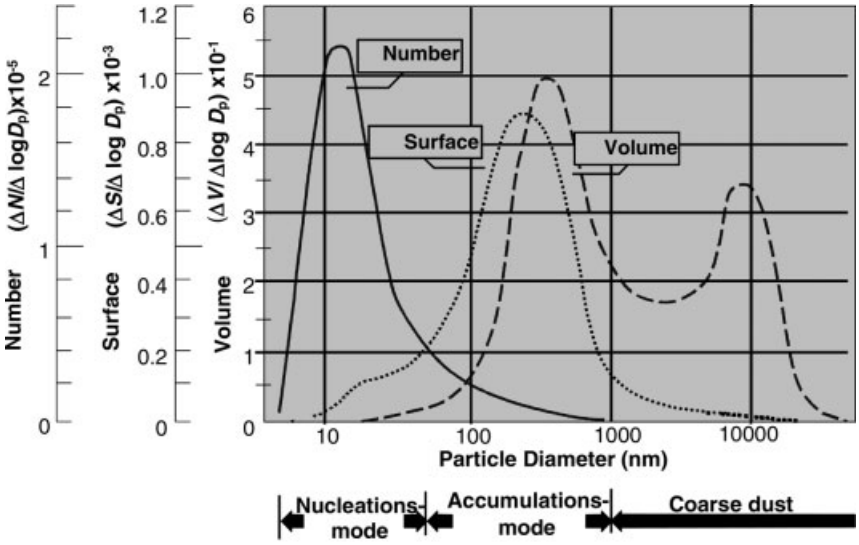


Figure 8.10 Number, surface and mass size distributions (Adapted from [56]).

determined by particles recently being formed by homogeneous nucleation. Hence nucleation mode particles are normally formed via gas to liquid or solid-phase processes. These particles, being normally <30 nm in diameter, either agglomerate with particles of their size, especially when particle concentrations exceed approximately 10^6 cm^{-3}, or with particles of the accumulation mode. The latter is due to the high surface area of accumulation mode particles. The accumulation mode particles tend to grow and their upper size range is normally limited to about 1 µm in diameter due to sedimentation. Coarse mode particles mainly stem from mechanical processes such as wind erosion and eruptions in contrast to the nucleation particles. Coarse mode particles have only a limited lifetime in the atmosphere due to sedimentation. An example of a particle size distribution is given in Figure 8.10.

Figure 8.10 also gives an example of how the modes of a given particle size distribution change when weighted according to the number concentration (d_p^0), surface area concentration (d_p^2) or mass concentration (d_p^3). In summary, the most sensitive concentration value for the determination of nucleation mode particles according to Figure 8.10 is the number size distribution, for the accumulation mode particles the surface area and for the accumulation and coarse mode particles the mass concentration.

Figure 8.11 shows the parabolic curve of the atmospheric lifetimes of airborne particles with respect to the particle size. Particles around 0.2 µm have the longest atmospheric lifetime. Atmospheric particle lifetimes are mainly governed by agglomeration (mainly small on to larger particles), sedimentation and wet deposition (wash-out by, e.g., rain). The latter process affects particles independent of their size; sedimentation is the main dry deposition process of larger particles, whereas small particles have higher agglomeration rates than larger particles. Another removal process, particle deposition by diffusion, is of lesser importance in ambient

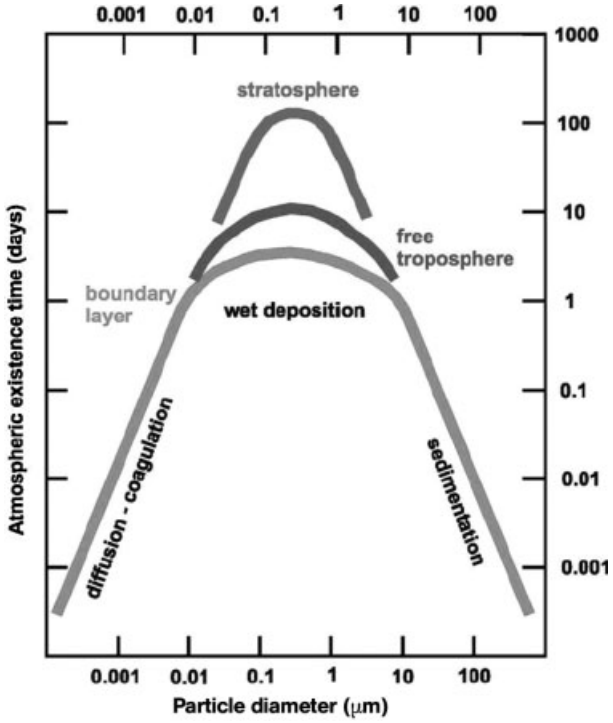


Figure 8.11 Atmospheric residence time of particles in days (Source: Jaenicke, R. (1982) Physical aspects of the atmospheric aerosol, in *Chemistry of the unpolluted and polluted troposhere* (Eds Georgii, H.W. and Jaeschke, W.), D. Reidel Publishing, Dordrecht, 341–373.).

air due to the low surface to volume ratio but becomes important when assessing indoor air and possibly workplaces (especially for nanoparticles and UFPs).

A standard way to quantify the longevity of a substance in the atmosphere is its 'lifetime' – the time that it takes for an initial amount to be cut by about two-thirds. Particle lifetimes in the atmosphere are strongly size dependent.

Figure 8.12 gives an example of particle number size distribution. Figure 8.12a shows the variance between different site types, from rural background to urban traffic sites. A clear increase in number concentrations and a shift towards UFP particles from the rural background to the urban traffic site can be seen. Figure 8.12b shows the same average urban particle number size distribution as in (a), with the only difference being that the *y*-axis is linear in scale. This example demonstrates the influence of changing scales on the perception of particle size distributions.

8.3.3.2 Online Physical Characterization

Condensation Particle Counter Condensation particle counters (CPC)s, sometimes also referred to as condensation nucleus counters (CNC)s or Aitken nucleus counters

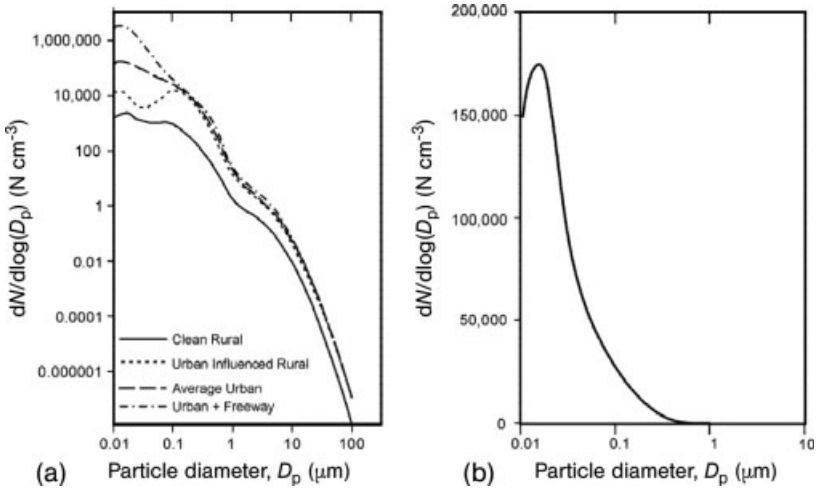


Figure 8.12 (a) Particle number size distributions for various site types in a log–log plot. (b) 'Average urban' in (a) with linear concentration axis [57].

(ANC)s, are used to measure the total number concentration of gas-borne particles larger than some minimum detectable size. No upper size limit is given for CPCs other than the collection efficiencies of the inlet system. CPCs are often used as either stand-alone instruments to measure the total concentration, for example in room or ambient air, or as detectors with other instruments, such as electrostatic classifiers (DMAs) to detect the number concentration of size-selected particles. Condensation particle counters are available as either hand-held or stationary instruments. The latter usually offer a larger liquid reservoir and can thus be used over longer periods, for example in conjunction with a scanning mobility particle sizer (SMPS). Hand-held CPCs are more mobile and are used, for instance, for the mapping of total particle concentrations in, for example, workplace environments.

In CPCs, particles are enlarged by vapor condensation and then are detected optically. Without additional preconditioning, the lowest detectable particle size would be limited to the range of the wavelength of light. Such particle counters would therefore not be suitable to detect nanoparticles. In order to grow small particles to optically detectable sizes, they are exposed to supersaturated vapor that condenses on the particle surfaces. Commonly used vapors are *n*-butanol [26] and water [27]. Diameter growth factors of 100–1000 are common [28], resulting in lower detection limits of 3 nm for *n*-butanol-based and 2.5 nm for water-based CPCs.

The fact that particles in air act as condensation nuclei was first described by Coullier in 1875 [53], who found that if air expands adiabatically the condensation effect is stronger in unfiltered compared with filtered air.

A first version of a condensation particle counter was developed by Aitken in the late 1880s [29, 30]. In his 'dust counter' as illustrated in Figure 8.13, he first flushed a test receiver (A) with particle free air, before introducing a known amount ($\sim$1 cm^3) of aerosol into the receiver. An air pump (B) was used to produce a known expansion.

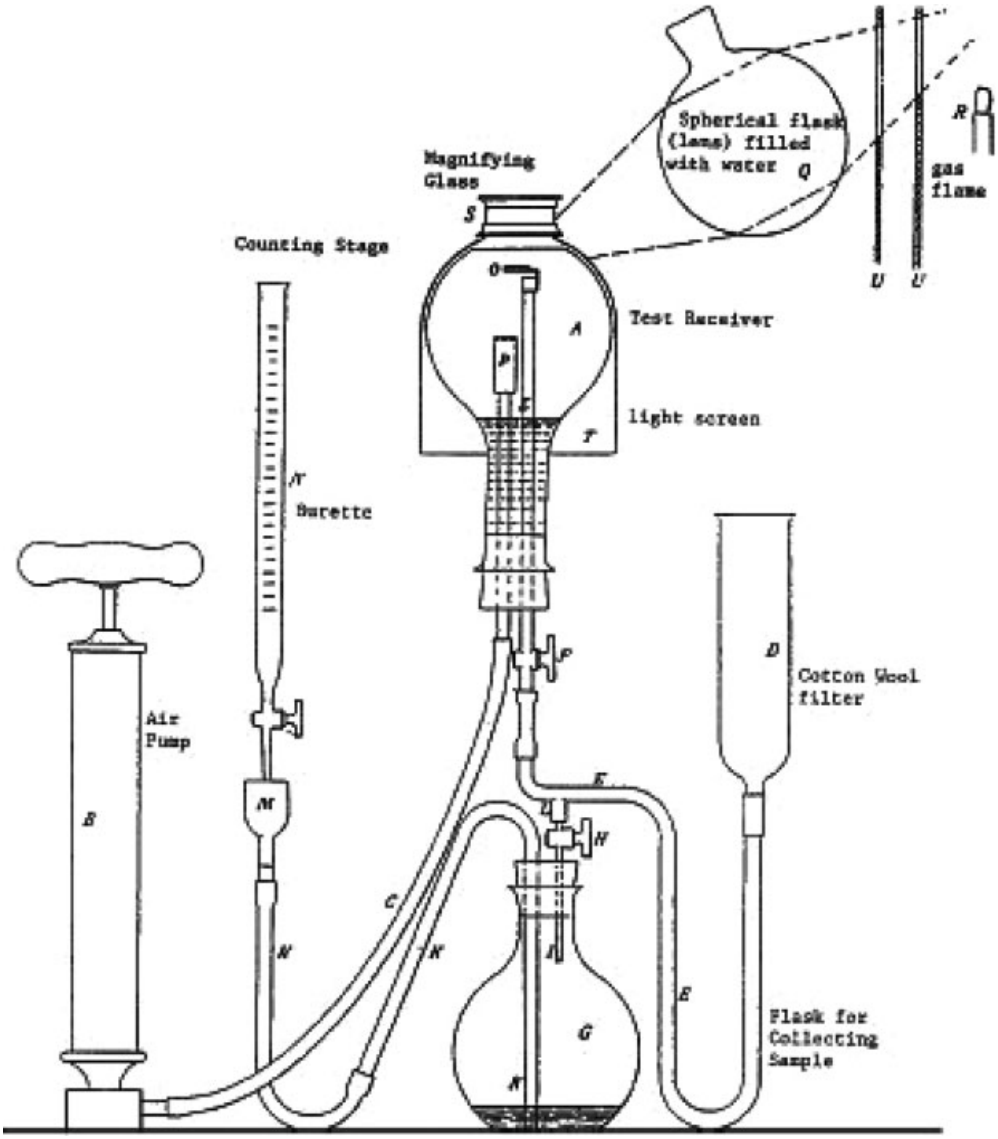


Figure 8.13 'Dust counter' as developed by Aitken [29].

The particles in the test receiver grew due to condensation, resulting in increased settling of the particles. Aitken used a magnifying glass (S) to count manually the particles settled on a deposition stage (O). Based on the assumption that the stage contained a representative sample of the particles in the receiver flask, he concluded that the total number of particles could be determined by means of the ratio of the volume above the stage and the total volume of the flask. Aitken used his dust counter and an improved portable dust counter [31] for intensive studies on atmospheric

particles and found that the particle concentrations were significantly higher when the wind was blowing from industrial sources and that the concentrations were affected by sunlight-driven photochemical reactions [32]. Based on these studies, he concluded [34]: 'Though this investigation clearly shows that the sun produces certain kinds of fogs, yet it is by no means here contended that it is to be censured for their appearance. It would rather appear that it is doing its best to show us the state of pollution into which our modern civilization has brought our atmosphere, as it only inflicts these fogs on the areas upon which man has thrown the waste products of his industries and converted the atmosphere into a vast sewer, as a penalty for something wrong in his methods'. Aitken's dust counter was thus an early instrument to help understand air pollution. Even now, well over a century later, the principle of enlarging particles by condensation remains an important tool for the determination of particle numbers. The main difference is that today particles are not grown in order to increase their settling, but to change their optical properties. CPCs can generally be distinguished into direct and indirect detection instruments [32]. While direct instruments determine the total number by counting individual droplets formed by condensation, indirect instruments measure the attenuation through or the light scattered by a population of droplets. Direct instruments usually have a lower upper concentration limit, whereas indirect instruments generally require relatively high concentrations. Modern instruments include both direct and indirect counting, depending on the particle concentration. When a certain concentration limit is exceeded, the instrument automatically switches from direct to indirect mode.

The method of producing supersaturation has changed since the Aitken's time. Whereas early instruments still used the discontinuous expansion method, modern CPCs use steady flow, forced-convection heat transfer that allows particle counting in real time. In these instruments, the saturated aerosol at ~35–40 °C enters a laminar flow condenser. The condenser walls are typically maintained at ~10 °C, causing a forced heat transfer from the warm aerosol to the cool walls [33]. Figure 8.14 shows a schematic of a modern, butanol-based ultrafine particle counter (UCPC, TSI Model 3025A) that can detect particles down to 3 nm.

Electrometer Electrometers are used to measure a current, induced by particle-borne charges. The results are integral values of the total electrical particle charge as current. Figure 8.15 shows a schematic diagram of an electrometer. The instrument consists of an absolute filter inside a grounded metal housing. The filter is connected to an electrometer and all charges are removed from the particles and led to the electrometer. The housing creates a Faraday cup that shields the electrometer input from stray electric fields. The total current measured by the electrometer can be expressed as

$$I_e = N\bar{n}_p e Q_e \tag{8.11}$$

where N is the total number of particles deposited on the filter, $\bar{n}_p$ is the average number of elementary charges on a particle, e is the elementary charge (1.602×10^{-19} A s) and Q_e is the flow rate through the electrometer.

Aged particles in ambient air are usually nearly neutralized, that is, the sum of all particle charges is close to zero, because approximately the same amount of particles

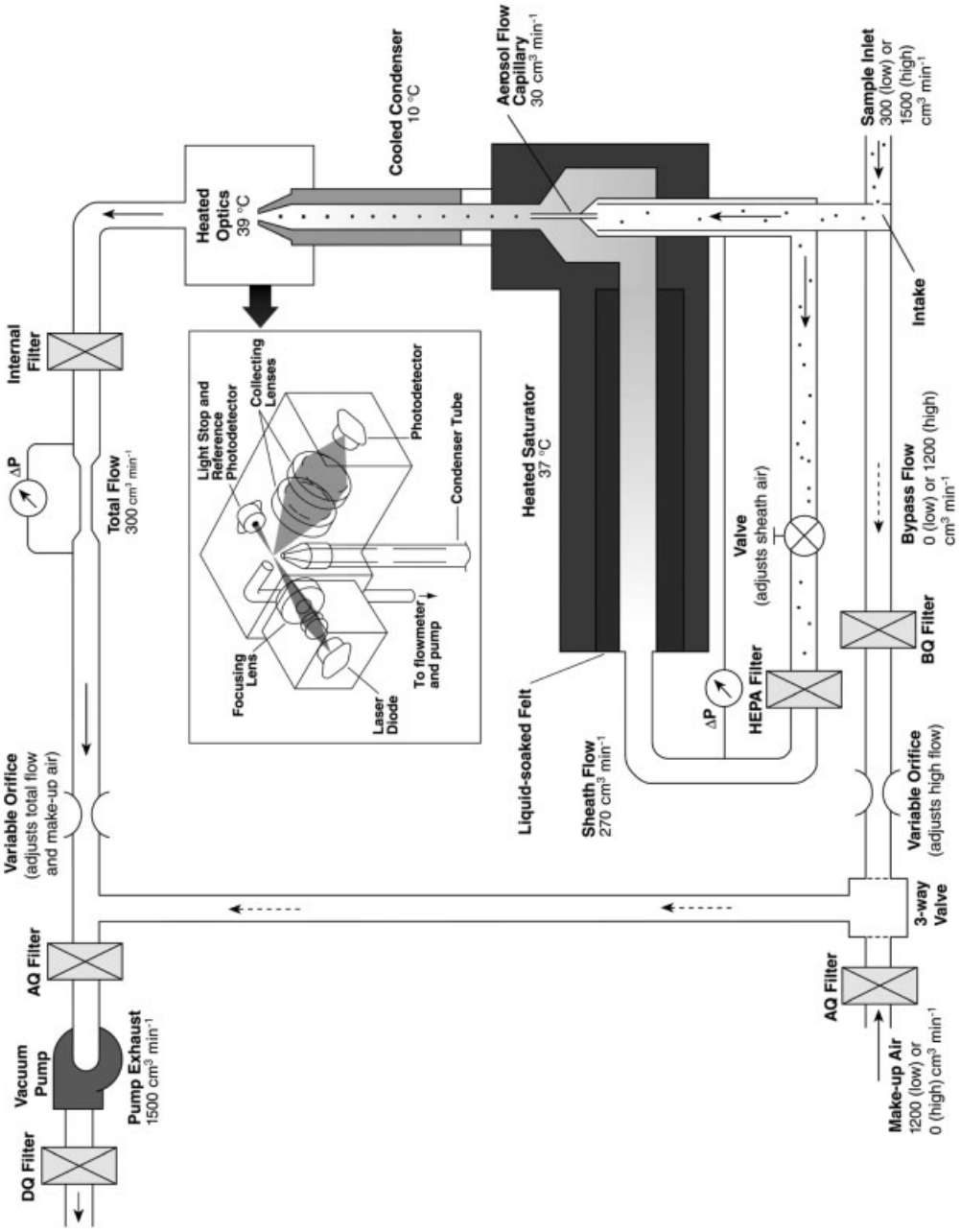


Figure 8.14 Schematic of a modern ultrafine condensation particle counter (TSI, Model 3025A).

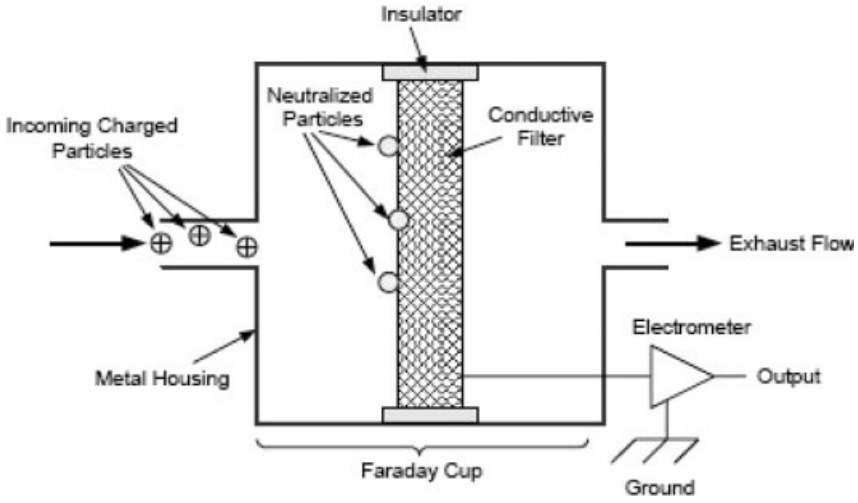


Figure 8.15 Schematic of an aerosol electrometer (TSI, Model 3068B).

is positively and negatively charged. In this case the electrometer current would be close to zero. If, however, the particles are freshly generated, they usually bear mainly unipolar charges with the polarity dependent on the generation process. The polarity of the measured current could therefore yield insight into the origin of the deposited particles. If the particles become intentionally charged prior to deposition in an electrometer and the relationship between particle size and charge is known, the integral current allows (limited) interpretation with respect to particle size and concentration. If the particles are monodisperse or mono-mobile, as for example delivered by an electrostatic classifier, the current can be directly related to the number concentration of particles in the aerosol. This fact is used in several applications, where condensation particle counters cannot be used, for example due to pressure or time resolution restrictions. Electrometers can detect arbitrarily small particles; however, they require a certain minimum current, caused by the deposited particles. Therefore, they require higher minimum particle concentrations than CPCs (Figure 8.15).

Overview of Aerosol Particle Sizing Instrumentation Various techniques for the determination and characterization of airborne particle size distributions (e.g. number, surface or size) exist. Three physically different basic principles for particle size determination can be differentiated: the electrical mobility, mechanical mobility and optical scattering. Figure 8.13 gives an overview of particle size ranges covered by the different techniques (abbreviations are explained in Table 8.2).

The optical scattering of light is dependent on the particle size in addition to the refractive index and hence is used directly for particle size classification and counting. The electrical and mechanical mobility are used for the fractionation of particle sizes with subsequent measurement of the number concentrations per size class by a separate particle counting device such as CPC or electrometer. Particle number size distributions are hence calculated based on the known investigated volume and the

Table 8.2 Instrumentation fort he determination of particle size distributions.

Particle sizer	Abbreviation	Sampling interval	Earliest References
Nano-scanning mobility particle sizer	Nano-SMPS	Continuous	Chen *et al.* [59]
Scanning mobility particle sizer	SMPS	Continuous	Wang and Flagan [37]
Differential mobility spectrometer	DMS	Continuous	Reavell *et al.* [60]
Fast mobility particle sizer	FMPS	Continuous	Mirme *et al.* [43]
Electrical diffusion battery	EDB	Continuous	Fierz *et al.* [61]
Nano-Moudi	Nano-Moudi	Discontinuous	Marple and Olson [62]
Moudi	Moudi	Discontinuous	Marple *et al.* [63]
Electrical low-pressure impactor	ELPI	(Dis)Continuous	Berner *et al.* [64]; Keskinen *et al.* [65]
Optical particle counter (sizer)	OPC	Continuous	Sinclair and La Mer [66]
Aerodynamic particle sizer	APS	Continuous	[68]

counted particles per size class. Table 8.2 presents an overview of the available techniques and corresponding earliest references.

A very good overview of particle sizers based on optical and time-of-flight techniques can be found in a dissertation by Cole [34]. Further good overviews were given by Klaus and Baron [35], Chow [36] and McMurry [28].

The principles of the first two techniques (APS and OPS) will be briefly described in this section and the other techniques are presented in separate sections hereafter.

Particles are accelerated in an APS by passing the aerosol through a nozzle. Particles experience acceleration in that nozzle in accordance with their aerodynamic diameter and their speed after the nozzle is directly proportional to it. This speed is determined by a split laser beam after the nozzle, which simultaneously counts the particles. The main drawback of the APS is that it can only detect particles larger than about 500 nm aerodynamic diameter. Hence it can only be applied for the detection of large nanoparticle agglomerates and not for nanoparticles.

The OPC has a similar drawback since it normally is limited to particle sizes in the size range of the wavelength used. The detectable lower particle diameter is normally around 200 nm (Figure 8.16).

Differential and Scanning Mobility Particle Sizer To determine the number size distribution of submicron airborne particles, differential mobility particle sizers (DMPSs) and SMPSs are very commonly used. Depending on their configuration, they can cover a size range between 3 nm and 1 μm. The hardware of both instruments is substantially the same. The aerosol first passes through a size-selective inlet (impactor), before being neutralized in a neutralizer. The neutralized aerosol is then

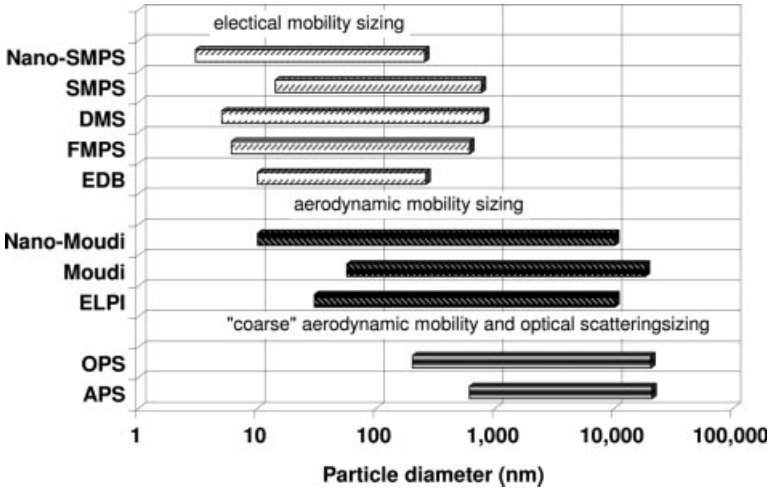


Figure 8.16 Size ranges of particle size distribution analyzers.

fractionated in a DMA, before the classified particles are counted in particle counting equipment (usually a CPC, sometimes an electrometer). Computer software is used to control the voltages applied to the DMA and to read the counts from the CPC. A schematic of a DMPS/SMPS is shown in Figure 8.17.

The general principle of a DMPS/SMPS is that a range of voltages is applied to the DMA. Each voltage corresponds to a certain electrical mobility. The concentration of mono-mobile particles is measured with the CPC and along with the known charge distribution evaluated to obtain the concentration of particles with this mobility diameter. In a DMPS, the voltages are applied sequentially and the number of particles counted after the concentrations have stabilized, before the next voltage is applied. In the more recently developed SMPS [37], the voltage is continuously ramped and an algorithm used to relate the measured concentrations to the applied voltages, that is to the according electrical mobility. While a DMPS still needed for ~15–20 min for a full scan, an SMPS can accomplish full scans within ~2 min.

The inversion of the measured mobility data into size distributions was described by Hoppel [39]. Since the particles were neutralized prior to classification in the DMA, the counted numbers not only contain singly but also larger, multiply charged particles, as sketched in Figure 8.18a. In order to determine the number size distribution from the mobility distribution, the concentration of multiply charged particles has to be subtracted for each channel (Figure 8.18b) and the resulting concentration of singly charged particles divided by the probability of singly charged particles for that particular size (Figure 8.18c), as for example given by Wiedensohler [40] (see Table 8.1). To understand the inversion technique, it is easiest to start with the channel of lowest electrical mobility, that is, with the largest particles. An appropriate size-selective inlet for a DMPS/SMPS is designed such that its cut-off diameter is below the diameter of doubly charged particles of the lowest detectable electrical mobility. Therefore, all particles in the lowest mobility channel are singly charged, because all particles bearing higher charge levels are captured in the

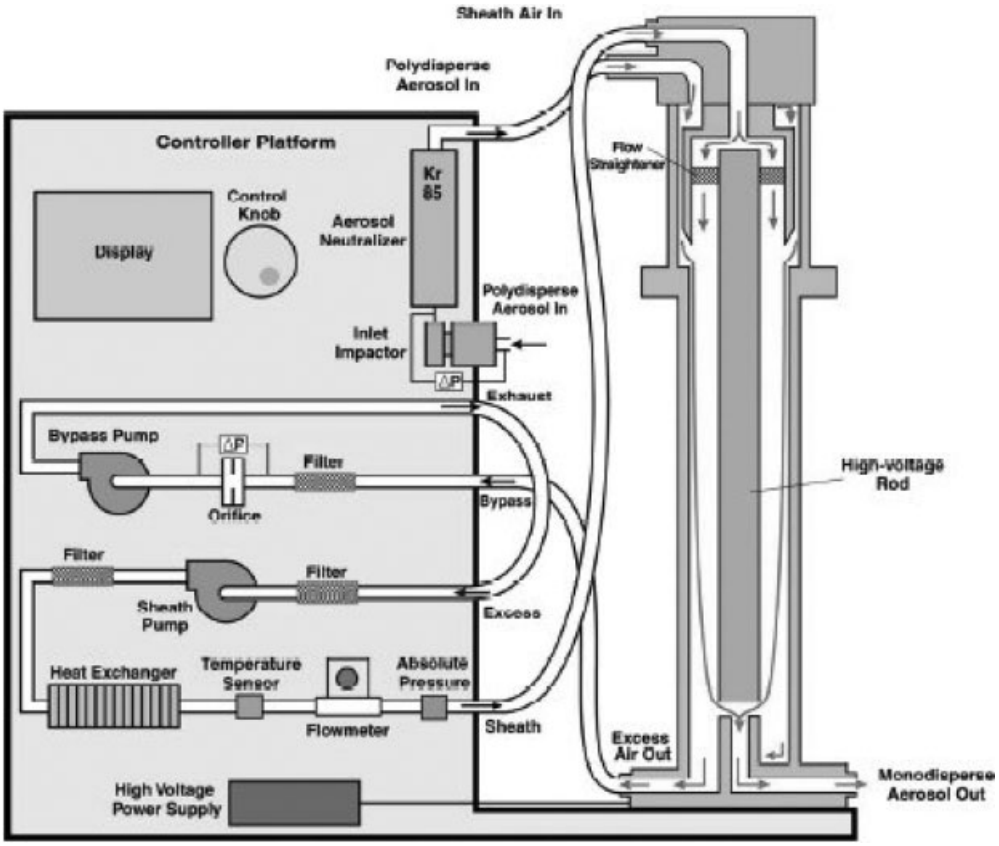


Figure 8.17 Schematic of a DMPS/SMPS (Courtesy of TSI Inc.).

pre-selector. The measured number concentration in the lowest mobility channel can thus without further correction be directly divided by the probability of singly charged particles in order to obtain the airborne number concentration. The same is true for all lower mobility channels, where the size of doubly charged particles is large enough to be captured in the pre-selector. In higher mobility channels, doubly (or higher) charged particles are present. Their size can be calculated by means of Equation 8.10 in the section on electrical mobility. Their concentration within the total measured concentration in the channel can be determined by multiplying the airborne concentration of particles of this (larger) size with the probability of doubly charged particles. The concentration of doubly charged particles is then subtracted from the measured concentration in order to obtain the concentration of only singly charged particles, which then needs to be divided by the probability of singly charged particles of that size in order to obtain the airborne concentration. Once the channels also contain triply or higher charged particles, they need to be subtracted from the measured data as described for the doubly charged particles and the result divided by the probability of singly charged particles.

After the correction described above (Figure 8.18a–c), the result is the number concentration for each channel where the channels now represent the average diameter of the singly charged particles. The magnitudes of the concentrations, however, might not necessarily represent the actual size distribution, as the width of the channels can vary with respect to the particle diameter. The number concentration for each channel (Figure 8.18c) therefore needs to be weighted with the channel width (Figure 8.18d). In SMPSs, the channel widths are based on either common or natural logarithms. Therefore, the number concentrations dN per channel are weighted with either the common logarithm [dN/d log(d_p)] or the natural logarithm [dN/d ln(d_p)].

All of the above-mentioned data inversion was based on the assumption that no particle losses occur in a DMPS/SMPS system. This is obviously not true. The data therefore need to be post-processed to be corrected for particle losses. Since the DMPS/SMPS systems are designed for submicron particles, losses can mainly be attributed to diffusion, whereas other loss mechanisms can be neglected. Corrections for diffusion losses are not generally covered in DMPS/SMPS evaluation software packages and therefore in some cases need to be done manually. If the size

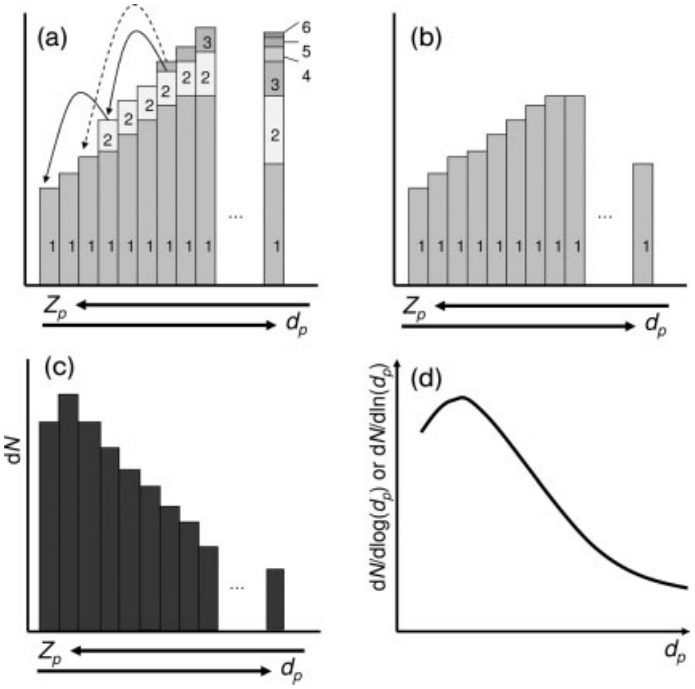


Figure 8.18 DMPS/SMPS data inversion: (a) measured mobility distribution, including multiply (1–6×) charged particles; arrows indicate the size of multiply charged particles; (b) multiply charged particles subtracted, only singly charged particles; (c) size distribution after division by probabilities of singly charged particles; (d) size distribution dN/d log(d_p) or dN/d ln(d_p), weighted with channel width.

distributions are not corrected, they tend to under-predict significantly the concentrations of particularly small particles with sizes mainly below 50 nm. In a DMPS/SMPS, diffusion losses occur in

- the size-selective inlet
- the neutralizer and internal plumbing
- the tubing to the DMA and CPC,
- the DMA
- the CPC.

If the DMPS/SMPS does not sample directly into the size-selective inlet, but with tubing connected to the inlet, losses inside this upstream tubing also need to be considered. Furthermore, the CPC counting efficiency needs to be taken into consideration. While the losses in the DMA require special treatment as described Reineking and Porstendörfer [67], the losses inside all tubing can be quantified as described by Gormley and Kennedy [40].

Based on the assumption that all sampled particles are spherical, the measured number size distribution can be converted into a surface or volume size distribution as illustrated in Figure 8.19. To compute the surface distribution, the number concentrations for each channel i need to be multiplied by the surface area of the particles in that channel:

$$\mathrm{d}S_i = \mathrm{d}N_i \pi \bar{d}_{\mathrm{p},i}^{\,2} \tag{8.12}$$

Similarly, the number concentration has to be multiplied by the particle volume to obtain the volume size distribution:

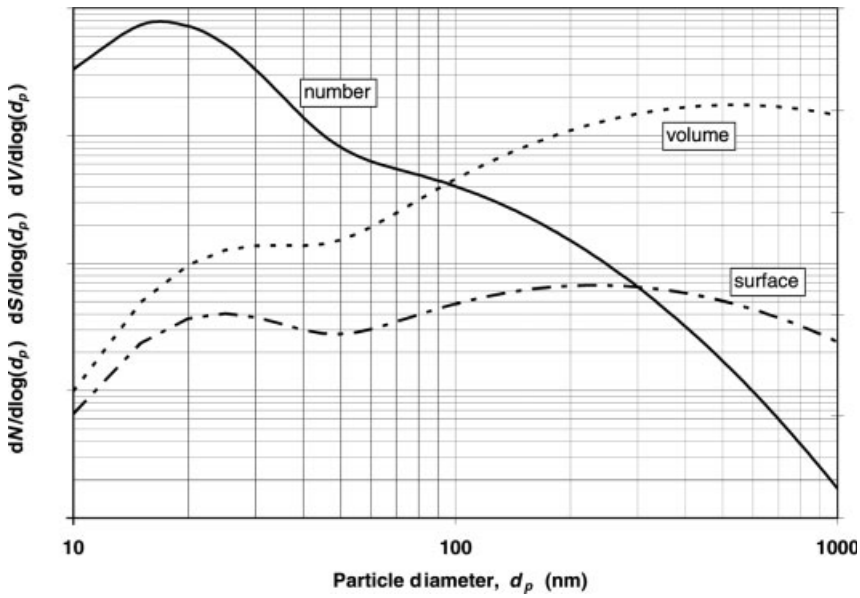


Figure 8.19 Number, surface and volume size distributions for spherical particles.

$$dV_i = dN_i \frac{\pi}{6} \bar{d}_p^3 \tag{8.13}$$

If the particles all have the same density, the volume distribution can also be transferred into a mass size distribution. If the particles are not spherical, the surface and volume distribution in Figure 8.19 can be understood as the surface and volume of electrically equivalent spheres.

Under certain assumptions, the surface area and volume distributions of chain-like ultrafine aggregates, such as soot particles, can be estimated, based on electrical mobility measurements as described by Lall and co-workers ([41, 42]). Their theory, however, is based on several assumptions:

- All aggregates are composed of primary particles, all of which have the same known diameter.
- The primary particles are much smaller than the mean free path of the surrounding gas molecules.
- The connections between the primary particles do not exhibit necks, that is, the surface area can be obtained by summing over the surface areas of the single primary particles.
- Fractal dimensions of the aggregates are smaller than two.

They take into account a different charging efficiency for aggregates compared with spheres, resulting in a shifted size distribution, and calculate the surface area of the particles based on the number and size of primary particles of which the agglomerates are composed.

Fast Mobility Particle Sizer SMPSs offer a time resolution of ~2 min. If the size distributions are quickly changing, such as due to the influence of a car driving by or accidental release of nanoparticles in workplaces, this time resolution may be too short. Tammet *et al.* [55], 1998 of Tartu University in Estonia developed an electrical aerosol spectrometer that measures number size distributions based on electrical mobility measurements. However, the different mobility channels are not measured sequentially as in an SMPS, but simultaneously, resulting in significantly higher time resolution. The instrument has been commercialized by TSI in two versions, the fast mobility particle sizer (FMPS) with a time resolution of 1 s and the engine exhaust particle sizer (EEPS) with a time resolution of 0.1 s. The two instruments are fundamentally the same. Whereas the FMPS is designed for ambient or workplace measurements, the EEPS is tailored for measuring engine exhausts and additionally includes means for the recording of engine data. It is only the higher particle concentration in engine exhaust that allows the EEPS to offer higher time resolution than the FMPS. Both instruments cover the same size range from 5.6 to 560 nm. The principle of the two instruments is shown in Figure 8.20.

The aerosol first passes a two-stage diffusion charger. The first stage puts a negative net charge on the particles in order to remove any potential high positive charge levels on the particles. The second stage puts a predictable net positive charge on the particles. The main instrument functions very similarly to a DMA, just that the

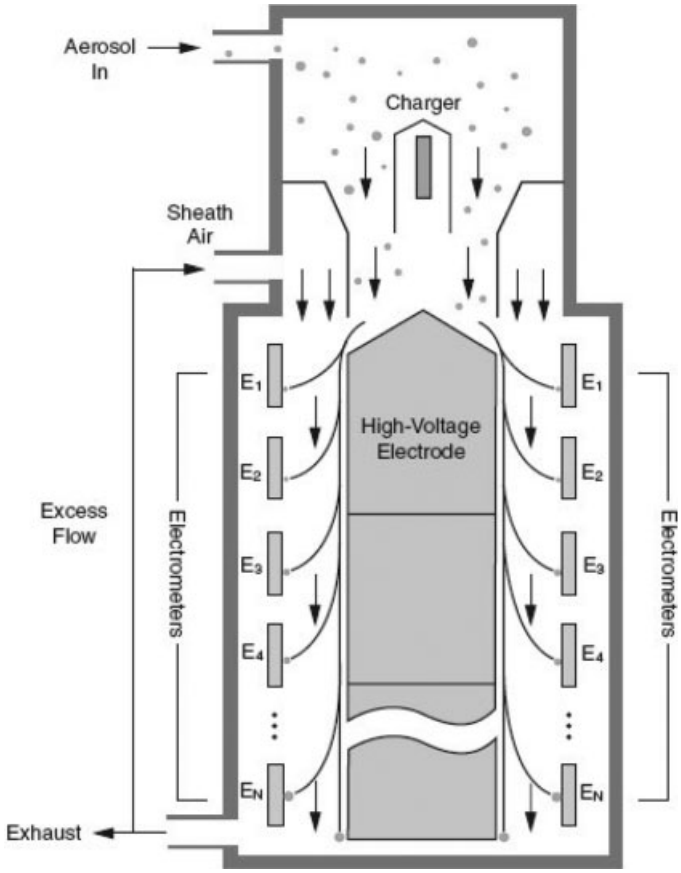


Figure 8.20 Schematic of the fast mobility particle sizer/engine exhaust particle sizer analyzer (Courtesy: TSI Inc.).

aerosol is introduced near the inner rod, with the sheath air surrounding the aerosol flow. The center rod is divided into three sections with different fixed voltages (increasing from top to bottom) applied to them. Since the voltages are fixed, the trajectories of the particles depend only on their electrical mobility. The outer cylinder contains a series of 22 electrodes, each separately connected to an electrometer. The current induced by deposited particles is continuously measured and used to determine the particle size distribution. The data inversion takes into account a number of parameters that affect the electrometer reading and the time resolution. The electrometer current can be affected by image charges that are induced if charged particles flow past a detection stage without being deposited. Additionally, there are time delays between the detection of small particles in an upper stage and larger particles in a lower stage.

A complex inversion algorithm is used to deconvolute the measured data and take into account image charges and time delays.

8.3.3.3 Online Physical–Chemical Characterization

Online, size dependent, single or bulk chemical analysis became practically available with the development of aerosol mass spectrometer only during recent years. They enable studies of e.g. particle composition, reaction, and source apportionment which were not possible before.

Aerosol Mass Spectrometers Two different aerosol mass spectrometers (AMSs) can be differentiated today – the real-time single-particle mass spectrometer (RTSPMS) and the thermal desorption chemical ionization mass spectrometer (TDCIMS).

The advantages of real-time, size-dependent determination of single particles were already seen in the 1970s. A schematic diagram of the first AMS is shown in Figure 8.21. The principle of this method is based on the ionization of particles directly in the mass spectrometer followed by ion mass and hence chemical analysis in the spectrometer. In the literature, a variety of general names have been applied to RTSPMS methods, including direct-inlet mass spectrometry, online laser microprobe mass spectrometry, particle-inlet mass spectrometry, particle beam mass spectrometry and rapid single-particle mass spectrometry [46].

The TDCIMS is a new instrument that can perform online measurements of the molecular composition of ultrafine aerosols at a time resolution of 5–10 min [47].

The TDCIMS will be outlined here with emphasis on recent major enhancements, while more details on the underlying measurement principle have been presented elsewhere [47]. Figure 8.22 shows a schematic of the instrument. The TDCIMS can be divided into three parts: the aerosol charger, the electrostatic precipitator and the

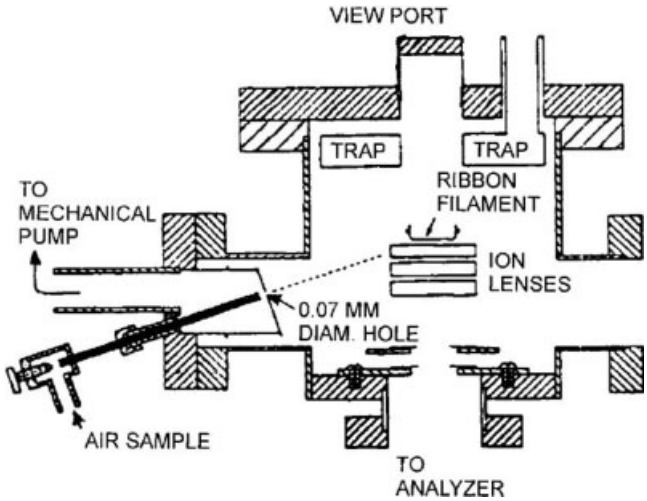


Figure 8.21 Schematic diagram of the original real-time single-particle mass spectrometer. The diagram shows the direct sample inlet directing an air flow to the filament for surface desorption/ionization of aerosol particles. Ions are mass analyzed in a magnetic sector mass analyzer (Adapted from [58] in [46]. Copyright 1977 American Chemical Society).

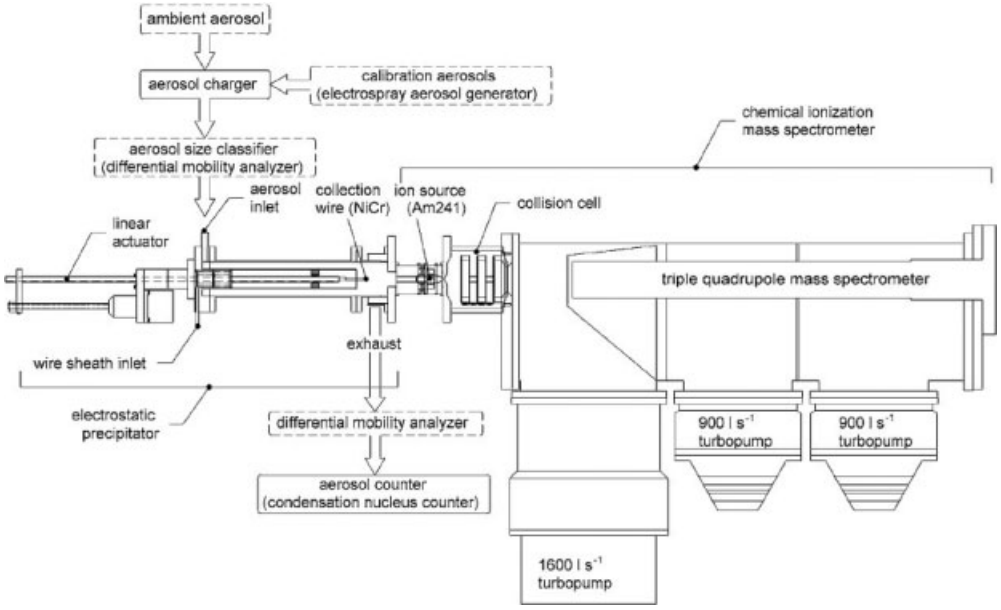


Figure 8.22 Schematic of an TDCIMS, with optional elements shown in dashed boxes (Adapted from [48]).

chemical ionization mass spectrometer. The aerosol charger generates singly charged aerosol from sampled ambient particles at flow rates of up to 6.6 (slpm). An optional differential mobility analyzer placed downstream of the charger can be used to achieve size selectivity of the aerosol. Charged particles are then introduced into the electrostatic precipitator, a cylindrical chamber that contains a collection wire mounted on a ceramic rod that is biased to a high voltage (usually 4200 V) and located on the center line of the chamber (see Figure 8.22). Prior to collection, the wire is cleaned by applying a current pulse to it, resistively heating it to 500 °C. After allowing the wire to cool to ambient temperature, a high voltage is applied so that these charged particles pass through a clean buffer gas that isolates the collection wire from contamination from the ambient gas and deposit on the tip of the wire. Once a sufficient number of particles have been collected, usually 1–10 pg over a period of 5–10 min, the wire is inserted into the ion source region of the chemical ionization mass spectrometer. A current is applied to the wire to thermally desorb the aerosol at a temperature of 300 °C. The third part of the instrument is the chemical ionization mass spectrometer and consists of the ion source region, a declustering collision cell and a mass spectrometer. The ion source consists of a ^{241}Am α-source that ionizes the reactant gas mixture at atmospheric pressure to form H_3O^+, O^{2-} and CO^{3-} and their higher clusters as the primary stable ions. The reactant gas is cryogenic nitrogen that has passed through a liquid nitrogen trap at slightly elevated pressure to remove some impurities, but which nonetheless is able to generate these ions in abundance. These regent ions react with the compounds evaporated from the aerosol to ionize them according to typical chemical ionization

mass spectrometry (CIMS) procedures [49] and electrostatic lenses direct these ions into the collision cell, where the ions are declustered from neutral compounds that may be present in the gas, most commonly water. The ions are detected using selected ion monitoring with a triple quadrupole mass spectrometer (paragraph adapted from [48]).

8.3.3.4 Offline Physical Characterization

Not all particle characteristics can be determined online, since either no online methods exist, as e.g. for particle morphology, or detection limit constrains. Therefore particles can be sampled on suitable substrates to overcome these limitations.

Electrostatic Precipitations and Impaction Electrostatic precipitators (ESPs) provide a simple mean for the collection of samples for offline analysis of, for example, chemical composition or morphology of airborne particles. Common applications of ESPs include samples for electron microscopy (SEM/TEM), for example coupled with EDX analysis, samples for atomic force microscopy (AFM), samples for total reflection X-ray fluorescence (TXRF) analysis, nanomaterial evaluation and atmospheric particle sampling.

Inside an ESP, the particles are exposed to an electric field, which directs particles of one polarity towards a sample electrode, whereas those particles of other polarity are deposited on the ESP wall. Uncharged particles are not affected by the electric field and therefore follow the gas streamlines. For effective particle sampling, unipolarly charged particles should be introduced into an ESP. Unipolar charging of particles can for example be achieved by a corona charger upstream of the ESP.

If connected directly to an outlet of a DMA, an ESP can be used to sample size selected particles without additional charging.

Figure 8.23 shows an ESP as developed by Dixkens and Fissan [50] which has been commercialized (TSI Model 3089). In this ESP, particles are homogenously distributed within a deposition spot. The spot size can be varied by controlling the flow aerosol flow rate and voltage applied to the electrode system.

Nanoparticles may also be sampled by using the impaction process. Commercially available cascade impactors separating and sampling particles even down to about 20 nm are the Nano-Moudi, the Berner low-pressure impactor and the electrical low-pressure impactor (ELPI) (Figures 8.24 and 8.25, see also [51]).

These cascade impactors are all based on the same principle, the inertia of particles. This principle is explained in Section 8.3.3.1. The main extension to the principle explained therein is the low-pressure conditions. Low pressure is necessary actually to use the principle of inertia to differentiate particles in the nanometer size range since gas–particle interactions otherwise interfere with the process and make a particle size separation for those size classes impossible.

Once the particles have been deposited on the substrates, either by ESP or by impaction, particles may be chemically analyzed in the bulk or as single particles. ESP is the better option for the latter chemical analysis.

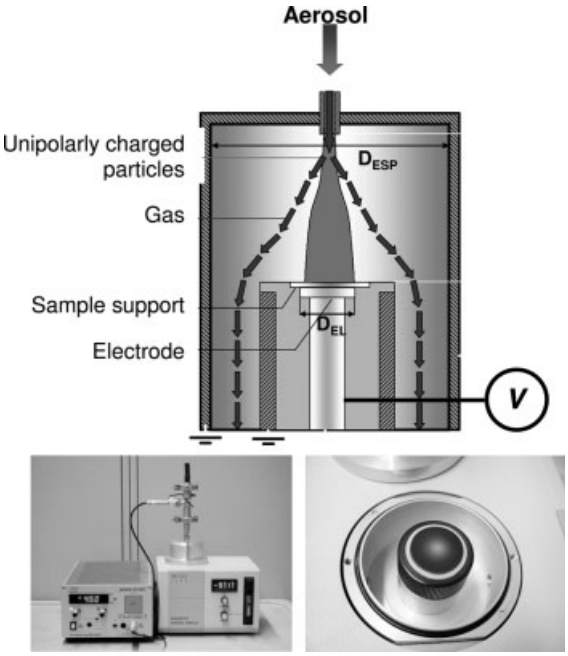


Figure 8.23 Schematic of an electrostatic precipitator (Dixkens and Fissan, 1999) and picture of commercial electrostatic precipitator with glassy carbon carrier.

8.4 Nanoparticle Detection and Measurement Strategies

No clear differentiation between natural and manmade ultrafine particles and nanoparticles can easily be done when measuring particles in the nanometer size range, as was shown in Section 8.3. Normally only a combination of methods will allow for a clear differentiation of nanoparticles from others. This will be explained in some detail for airborne nanoparticles.

The detection and measurement strategies to be pursued depend mainly on the nanoparticle characteristics and on the detection limits. Nanoparticle characteristics can be divided into physical–chemical and morphological characteristics. The physical–chemical characteristics determine, for example, which detection technique can be used for the identification of the nanoparticles. If, for example, TiO_2, Al_2O_3 or Fe nanoparticles are to be detected, a combination of TEM with EDX can be used. The microscopic techniques (TEM, SEM, ESEM (Environmental Scanning Electron Microscope)) also allow for the determination of the morphology/shape which can be of interest for the identification of nanoparticles of known shape and/or primary particles size. However, these techniques are less suitable for carbon nanoparticles since they cannot determine the chemical composition. Techniques such as single-particle AMS and/or particle size distribution measurement along with the online

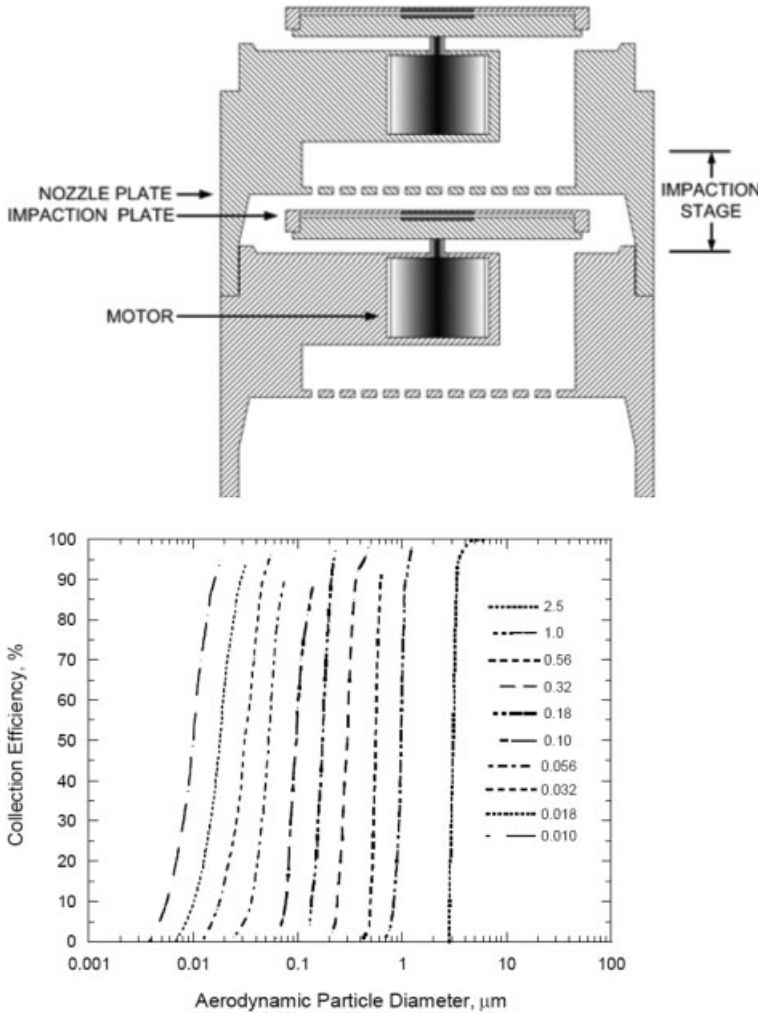


Figure 8.24 Schematic of Nano-Moudi II setup and particle separation curves (Courtesy of MSP Corporation).

detection of submicrometer elemental carbon (e.g. with an Aethalometer) may be more suitable. However, the microscopic methods are still an important tool actually to identify the source of carbon (e.g. diesel, nanotubes, carbon black).

Particle number concentrations, particle size distributions and surface area concentration may also be used for the measurements of nanoparticles. However, these methods are of more general character and do not differentiate between nanoparticles and other particles of diameter <100 nm.

The above-mentioned possible setups of measurement devices and techniques are important for the actual particle characterization. Nevertheless, these techniques are time consuming and expensive. Still, measurements as described above will be necessary if single nanoparticles in a complex matrix are to be detected and described

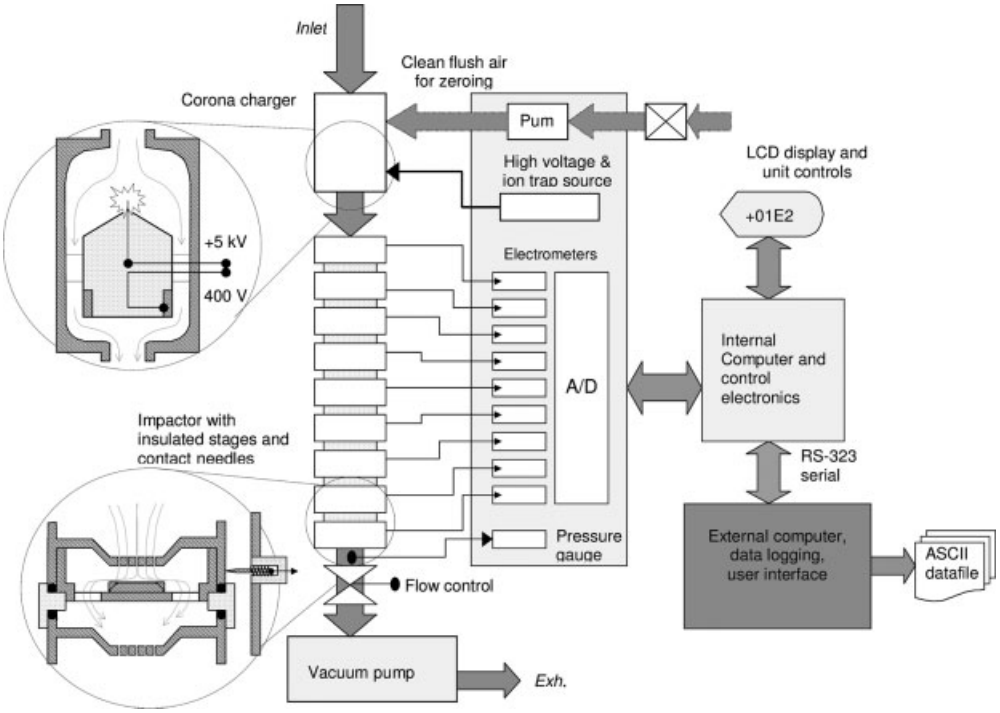


Figure 8.25 Schematic of an ELPI (Courtesy of Dekati Ltd.).

to be traceable to the product or are of such high toxicity that even single particles can cause health effects (e.g. asbestos).

The measurement techniques to be employed may be different if only larger quantities of nanoparticles have to be detected. Even here two different levels may be differentiated, exemplarily discussed here for number concentrations of particles of diameter <100 nm: nanoparticle contributions to airborne particles from 1000–100 000 cm^{-3} and contributions >100 000 cm^{-3}. Measurement techniques as used by Möhlmann [18] are applicable to identify source contributions >100 000 cm^{-3}. In his study, a single SMPS measuring particle number size distributions from 10 to 700 nm was used. This measurement allows for the calculation of the particle number concentration of particles of diameter <100 nm and for the determination of the mode (particle size with the highest number concentration). Still, since particle number concentrations of up to 100 000 cm^{-3} may occur naturally or by contributions from outside, this kind of measurement strategy is only applicable for higher number concentrations.

A different approach must be pursued if particle contributions by processes, leaks and work activities at levels down to about a few thousand particles per cubic meter are to be determined. Kuhlbusch *et al.* [17] and Kuhlbusch and Fissan [19] showed that particle contributions from sources outside the plant can be significant and have to be taken into account. By choosing a setup of instruments with simultaneous measurements inside the working area and at a comparison site in the direct vicinity of the work area but outside, they demonstrated the influence of outside

Table 8.3 Concentration-dependent measurement strategies.

Number concentration range (cm^{-3})	Strategy
>100 000	Determination of particle number concentrations and/or number size distributions only at the location of interest
1000–100 000	Determination of particle number concentrations and/or number size distributions simultaneously at the location of interest and a corresponding comparison location
<1000	Determination of particle number concentrations and/or number size distributions, along with particle samplers for single particle analysis for nanoparticle identification

Source: T. A. J. Kuhlbusch *et al.*, in preparation [54].

contributions to indoor measurements. By choosing this kind of setup they were able to differentiate (a) sources from outside of the work area, (b) continuous source contributions and (c) discontinuous source contributions. This possibility of differentiation of continuous and discontinuous source activities can be important when the continuous sources are active during the assessment of a discontinuous activity. The particles from the continuous source would have been attributed to the discontinuous source activity if no differentiation were to have been made.

Still, even when determining source activities by the methods describe above, single particle analysis may still be necessary to avoid data misinterpretation, such as attributing welding particles to nanoparticle bagging activity (Table 8.3).

References

1 Oberdörster, G., Sharp, Z., Atudorei, V., Elder, A., Gelein, R., Kreyling, W. and Cox, C. 2004 Translocation of inhaled ultrafine particles to the brain. *Inhalation Toxicology*, **16** (6–7), 437–445.

2 Donaldson, K., Li, X.Y. and MacNee, W. (1998) Ultrafine (nanometer) particle mediated lung injury. *Journal of Aerosol Science*, **29** (5–6), 553–560.

3 Oberdorster, G., Oberdorster, E. and Oberdorster, J. (2005) Nanotoxicology: an emerging discipline evolving from studies of ultrafine particles. *Environmental Health Perspectives*, **113** (7), 823–839.

4 Borm, P.J.A., Robbins, D., Haubold, S., Kuhlbusch, T., Fissan, H., Donaldson, K., Schins, R., Stone, V., Kreyling, W., Lademann, J., Krutmann, J., Warheit, D., and Oberdorster, E., (2006) The potential risks of nanomaterials: a review carried out for ECETOC. *Particle and Fibre Toxicology*, **3** (14 August), 11.

5 Californian Proposition 65 (2003) Californian Proposition 65: Office of Environmental Health Hazard Assessment (OEHHA) of the California Environmental Protection Agency is adding carbon black (airborne, unbound particles of respirable size) to the list of chemicals known to the State to cause cancer for purposes of the Safe Drinking Water and Toxic Enforcement Act of 1986 (Proposition 65). The listing of carbon black is effective 21 February 2003.

6 Tungittiplakorn, W., Lion, L.W., Cohen, C. and Kim, J.Y. (2004) Engineered polymeric nanoparticles for soil remediation.

Environmental Science and Technology, **38**, 1605–1610.

7 Lecoanet, H. and Wiesner, M.R. (2004) Velocity Effects on the deposition of fullerene and oxide nanoparticles in porous media. *Environmental Science and Technology*, **38** (16), 4377–4382.

8 Lecoanet, H., Bottero, J.Y. and Wiesner, M.R. (2004) Laboratory assessment of the mobility of several commercial nanomaterials in porous media. *Environmental Science and Technology*, **38** (19), 5164–5169.

9 Oberdörster, E. (2004) Manufactured nanomaterials (fullerenes, C60) induce oxidative stress in the brain of juvenile largemouth bass. *Environmental Health Perspectives*, **112**, 1058–1062.

10 Junge, C.E. (1963) *Air Chemistry and Radioactivity*, Academic Press, New York.

11 IPCC (2001) Third Assessment Report, Climate Change 2001, IPCC, Geneva, Switzerland. http://www.ipcc.ch/pub/reports.htm.

12 Twomey, S. (1977) The influence of pollution on the shortwave albedo of clouds. *Journal of the Atmospheric Sciences*, **34**, 1149–1152.

13 Twomey, S. (1980) Cloud nucleation in the atmosphere and the influence of nucleus concentration levels in atmospheric physics. *Journal of Physical Chemistry*, **84**, 1459–1463.

14 Swap, R., Garstang, S., Greco, S., Talbot, R. and Kallberg, P. (1992) Saharan dust in the Amazon basin. *Tellus*, **44**, 133–149.

15 WHO (2006) Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide, Global update 2005, http://www.euro.who.int/air, WHO, Geneva, Switzerland.

16 Van Dingenen, R., Raes, F., Putaud, J.-P.,*et al*. (2004) A European aerosol phenomenology – 1: physical characteristics of particulate matter at kerbside, urban, rural and background sites in Europe, *Atmospheric Environment*, **38**, 2561–2577.

17 Kuhlbusch, T.A.J., Neumann, S. and Fissan, H. (2004) Number size distribution, mass concentration and particle composition of PM1, PM2.5 and PM10 in bagging areas of carbon black production, *Journal of Occupational and Environmental Health*, **1**, 660–671.

18 Möhlmann, C., (2004) German Activity on Ultra-fine Particles in the Workplace. First International Symposium on Occupational Health Implications of Nanomaterials, 12–14 October 2004, Palace Hotel, Buxton, Derbyshire, UK, http://www.hsl.gov.uk/capabilities/nanosymrep_final.pdf.

19 Kuhlbusch, T.A.J. and Fissan, H. (2006) Particle characteristics in the reactor and pelletizing areas of carbon black production. *Journal of Occupational and Environmental Health*, **3** (10), 558–567.

20 Hinds, W. (1999) *Aerosol Technology: Properties, Behavior and Measurement of Airborne Particles*, 2nd edn, Wiley-Interscience, New York.

21 Cunningham, E. (1910) On the velocity of steady fall of spherical particles through fluid medium. *Proceedings of the Royal Society of London, Series A*, **83**, 357–365.

22 Knudsen, M. and Weber, S. (1911) Air resistance against the slow movement of small spheres. *Annalen der Physik*, **36**, 983–996.

23 Kim, J.H., Mulholland, G.W., Kukuck, S.R. and Pui, D.Y.H. (2005) Slip correction measurements of certified PSL nanoparticles using a nanometer differential mobility analyzer (Nano-DMA) for Knudsen numbers from 0.5 to 83. *Journal of Research of the National Institute of Standards and Technology*, **110**, 31.

24 Berner, A. (1972) Praktische Erfahrungen mit einem 20-Stufen-Impaktor. *Staub – Reinhaltung der Luft*, **32**, 315–320.

25 Knutson, E.O. and Whitby, K.T. (1975) Aerosol classification by electric mobility: apparatus, theory and applications. *Journal of Aerosol Science*, **6**, 443–451.

26 Stolzenburg, M.R. and McMurry, P.H. (1991) An ultrafine aerosol condensation nucleus counter. *Aerosol Science and Technology*, **14**, 48–65.

27 Hering, S.V., Stolzenburg, M.R., Quant, F.R., Oberreit, D.R. and Keady, P.B. (2005) A laminar-flow, water-based condensation particle counter (WCPC). *Aerosol Science and Technology*, **39**, 659–672.

28 McMurry, P. (2000) A review of atmospheric aerosol measurements. *Atmospheric Environment*, **34**, 1959–1999.

29 Aitken, J. (1888) On the number of dust particles in the atmosphere. *Transactions of the Royal Society of Edinburgh*, **35**, 1–20.

30 Aitken, J. (1888–1889) On improvements in the apparatus for counting the dust particles in the atmosphere. *Proceedings of the Royal Society of Edinburgh*, **16**, 207–235.

31 Aitken, J. (1890–1891) On a simple pocket dust counter. *Proceedings of the Royal Society of Edinburgh*, **18**, 39–53.

32 McMurry, P. (2000) The history of condensation nucleus counters. *Aerosol Science and Technology*, **33**, 297–322.

33 Aitken, J. (1911) The sun as a fog producer. *Proceedings of the Royal Society of Edinburgh*, **32**, 193–325.

34 Cole, R. (1999) High Resolution Micro-motion Aerosol Spectrometry. PhD Dissertation, University of Arkansas at Little Rock.

35 Klaus, W. and Baron, P. (1993) *Aerosol Measurement Principles, Techniques and Applications*, Van Nostrand Reinhold, New York.

36 Chow, J.C. (1995) Measurement methods to determine compliance with ambient air quality standards for suspended particles. *Journal of the Air and Waste Management Association*, **45** (5), 320–382.

37 Wang, S.C. and Flagan, R. (1990) Scanning electrical mobility spectrometer. *Aerosol Science and Technology*, **13**, 230–240.

38 Hoppel, W.A. (1978) Determination of the aerosol size distribution from the mobility distribution of charged fraction of aerosols. *Journal of Aerosol Science*, **9**, 41–54.

39 Wiedensohler, A. (1988) An approximation of the bipolar charge distribution for particles in the submicron range. *Journal of Aerosol Science*, **19**, 387–389.

40 Gormley, P.G. and Kennedy, M. (1949) Diffusion from a stream flowing through a cylindrical tube. *Proceedings of the Royal Irish Academy*, **52A**, 163–169.

41 Lall, A.A. and Friedlander, S.K. (2006) On-line measurement of ultrafine aggregate surface area and volume distributions by electrical mobility analysis: I. Theoretical analysis. *Journal of Aerosol Science*, **37**, 260–271.

42 Lall, A.A., Seipenbusch, M., Rong, W. and Friedlander, S.K. (2006) On-line measurement of ultrafine aggregate surface area and volume distributions by electrical mobility analysis: II. Comparison of measurements and theory. *Journal of Aerosol Science*, **37**, 272–282.

43 Mirme, A., Tamm, E. and Tammet, H. (1981) Electrical granulomater of aerosol particles with wide measurement range. *Acta et Commentationes Universitatis Tartuensis*, **588**, 84–92 (in Russian).

44 Mirme, A., Noppel, M., Peil, I., Salm, J., Tamm, E. and Tammet, H. (1984) Multi-channel electric aerosol spectrometer. In Proceedings of the 11th International Conference on Atmospheric Aerosols, Condensation and Ice Nuclei Budapest, Vol. 2, 155–159.

45 Mirme, A. and Tamm, E. (1993) Comparison of sequential and parallel measurement principles an aerosol spectrometry. *Journal of Aerosol Science*, **22** (S1), S211–S212

46 Noble, C.A. and Prather, K.A. (2000) Real time single particle mass spectrometry: a historical review of a quarter century of the

chemical analysis of aerosols. *Mass Spectrometry Reviews*, **19**, 248–274.

47 Voisin, D., Smith, J.N., Sakurai, H., McMurry, P.H. and Eisele, F.L. (2003) Thermal desorption chemical ionization mass spectrometer for ultrafine particle chemical composition. *Aerosol Science and Technology*, **37**, 471–475.

48 Smith, J.N., Moore, K.F., McMurry, P.H. and Eisele, F.L. (2004) Atmospheric measurements of sub-20 nm diameter particle chemical composition by thermal desorption chemical ionization mass spectrometry. *Aerosol Science and Technology*, **38**, 100–110.

49 Eisele, F.L. and Tanner, D.J. (1993) Measurement of the gas-phase concentration of H_2SO_4 and methane sulfonic-acid and estimates of H_2SO_4 production and loss in the atmosphere. *Journal of Geophysical Research-Atmospheres*, **98** (D5), 9001–9010.

50 Dixkens, H. and Fissan, H. (1999) Development of an electrostatic precipitator for off-line particle analysis. *Aerosol Science & Technology*, **30**, 438–453.

51 Chow, J.C. and Watson, J.G. (2007) Review of measurement methods and compositions for ultrafine particles, *Aerosol and Air Quality Research*, **7** (2), 121–173.

52 Beckman Coulter , (2006) http://www.beckmancoulter.com,

53 Coullier, M. (1875) Note sur une nouvelle propiété de l'air. *Journal de Pharmacie at de Chimie*, **22**, 165–173.

54 Kuhlbusch, T.A.J., Fissan, H. and Asbach, C. (2006) A measurement strategy for the determination of nanoparticles at workplaces at lower concentration levels, *JOEH*, 2008, in preparation.

55 Tammet, H., Mirme, A. and Tamm, E. (1998) Electrical aerosol spectrometer of Tartu University. *Journal of Aerosol Science*, **29** (S1), S427–S428

56 National Research Council (1979) *Airborne Particles*, University Park Press, Baltimore, MD.

57 Whitby, K.T. and Sverdrup, G.M. (1980) California aerosols: their physical and chemical characteristics. in *The Character and Origins of Smog Aerosols: a Digest of Results from the California Aerosol Characterization Experiment (ACHEX)* (eds G.M. Hidy, P.K. Mueller, D. Grosjean, B.R. Appel and J.J. Wesolowske), John Wiley & Sons, Inc., New York, pp. 477–517.

58 Davis, W.D. (1977) Continuous mass spectrometric analysis of particulates by use of surface ionization. *Environmental Science and Technology*, **11**, 587–592.

59 Chen, D.-R., Pui, D.Y.H., Hummes, D., Fissan, H., Quant, F.R. and Sem, G.J. (1998) Design and evaluation of a nanometer aerosol differential mobility analyzer (Nano-DMA) *Journal of Aerosol Science*, **29** (5), 497–509.

60 Reavell, K., Hands, T. and Collings, N. (2002) A Fast Response Particulate Spectrometer for Combustion Aerosols. SAE Paper 2002-01-2714.

61 Fierz, M., Scherrer, L. and Burtscher, H. (2002) Real-time measurement of aerosol size distributions with an electrical diffusion battery. *Journal of Aerosol Science*, **33**, 1049–1060.

62 Marple, V.A. and Olson, B.A. (1999) A Micro-orifice Impactor with Cut Sizes Down to 10 Nanometers for Diesel Exhaust Sampling, Final Report, Generic Center for Respirable Dust, Pennsylvania State University, University Park, PA.

63 Marple, V.A., Rubow, K.L. and Behm, S.M. (1991) Microorifice Uniform Deposit Impactor (MOUDI) – description, calibration and use. *Aerosol Science and Technology*, **14**, 434–446.

64 Berner, A., Lurzer, C., Pohl, F., Preining, O. and Wagner, P. (1979) The size distribution of the urban aerosol in Vienna. *Science of the Total Environment*, **13**, 245–261.

65 Keskinen, J., Pietarinen, K., and Lehtimäki, M. (1992) Electrical low-pressure impactor. *Journal of Aerosol Science*, **23**, 353–360.

66 Sinclair, D. and La Mer, V.K. (1949) Light scattering as a measure of particle size in aerosols. *Chemical Reviews*, **44**, 245–267.

67 Reineking, A., Porstendörfer (1986) Measurement of Particle lodd functions in a differential mobility analyzer (TSI, model 3071) for different flow rates, *Aerosol Science & Technology*, **27**, 483–486.

68 P.A. Baron, P.A. (1986) Calibration and use of the aerodynamic particle sizer (APS 3300), *Aerosol Science & Technology*, **5**, 55–67.

9
Epidemiological Studies on Particulate Air Pollution

Irene Brüske-Hohlfeld and Annette Peters

9.1
Introduction

This chapter presents an overview of the main results stemming from epidemiological research on the health effects of exposure to particulate air pollution in the environment and at the workplace. Over the past two decades, evidence has accumulated that airborne particles are correlated with the incidence of respiratory and cardiovascular disease. Although remarkably consistent between numerous epidemiological studies in different geographic areas, these findings were at first received with some skepticism, as there appeared to be no plausible biological mechanism to explain the observed association between respiratory and cardiovascular mortality and the level of airborne particles. As epidemiology is an observational rather than an experimental science, it cannot establish causality on its own and is a rather blunt tool for elucidating biological mechanisms. However, in the meantime complementary information from controlled *in vivo* and *in vitro* experimental studies has supplied supporting evidence, which will be presented, for example, in Chapter 10.

9.1.1
Outline of the Chapter

We will start with a short definition of particle sizes, a brief comment on epidemiological study design and a description of what is known about the main potential entry routes for nanoparticles, focusing on the inhalation and metabolism of airborne particles. Thereafter, a summary of the observed adverse health effects of particulate air pollution from environmental epidemiological studies is given, with special emphasis on studies that have particularly investigated the effects of ultrafine particles. Then, looking – so to speak - backwards from the adverse health effects to particle exposure, we will bundle the evidence for diseases related to three organs: the

Nanotechnology. Volume 2: Environmental Aspects. Edited by Harald Krug

ISBN: 978-3-527-31735-6

heart, the lungs and the central nervous system. Finally, we will summarize epidemiological studies that have looked into dusty workplace environments and have investigated the impact on health of exposed workers.

9.1.2 A Short Definition of Particle Sizes

Environmental air pollution consists of a complex mixture of compounds in gaseous, liquid and solid phases, the latter usually referred to as particulate matter (PM). Some particles are introduced from the source into the air in solid or liquid form, whereas others are formed in the air by gas to particle conversion. In general, ambient levels of particulate matter are characterized as particulate matter with an effective aerodynamic diameter of less than 10 µm (PM_{10}) or 2.5 µm ($PM_{2.5}$); see Figure 9.1.

These PM size cuts generally represent different sources and display different physical and chemical properties. They are generated by a large number of sources: motor vehicles, power plants, wind blown dust, photochemical processes, cigarette smoking, nearby quarry operation, etc. Ultrafine particles (UFP) as a component of ambient particulate air pollution are largely the result of combustion processes, such as automobile traffic and heating of homes, and composed of a core of elemental carbon covered by organic carbon compounds and secondary sulfates and nitrates. Although comparable in size (<100 nm) to technically produced nanoparticles, ambient ultrafine particles lack their potential specific toxicity. They have a high tendency to react chemically or coagulate. The number concentration of these very small particles exceeds by far that of larger particles in the urban area, but their contribution to the total

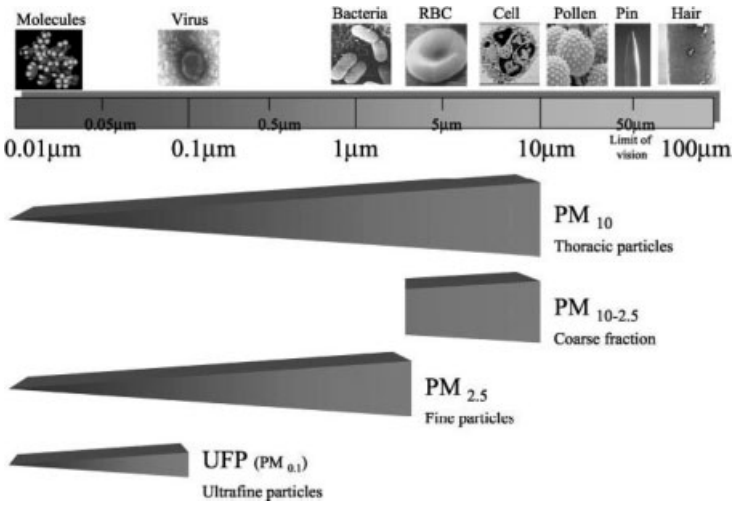


Figure 9.1 Particulate matter air pollution size distribution. (Reproduced with permission from [115]).

mass concentration is relatively low. With regard to ultrafine particles, the number concentration (n cm^{-3}) or surface area concentration ($\mu m^2\, m^{-3}$) or particle length concentration (mm cm^{-3}) is more relevant than particle mass. Such data on exposure are not routinely available by monitoring stations, but have to be collected independently. It has been proposed that the adverse health effect of airborne particles was mainly associated with the number concentrations of ultrafine particles [1–3] rather than the mass concentrations $PM_{2.5}$ or PM_{10}.

9.1.3
A Brief Comment on Epidemiological Study Design

As epidemiology often takes advantage of already existing data, most of the concurrent research relies on measurements of particle mass ($\mu g\, m^{-3}$) of PM_{10} or $PM_{2.5}$, which is typically recorded for regulatory purposes at central-site community-based monitoring stations. Although these studies do not specifically concentrate on nanosized particles, they contribute valuable information as it seems that it is the inherent fraction of combustion-derived ultrafine particles within PM_{10} that is actually responsible for the adverse health effects. There is sufficient reason to believe that ultrafine particles are important because compared with larger particles, they have a much larger surface area and higher concentrations of adsorbed or condensed toxic air pollutants (oxidant gases, organic compounds, transition metals) per unit mass.

The epidemiological approach to measuring associations between air pollution and health varies with the available data and the hypothesized health effects being investigated. Usually, air pollution epidemiological studies are classified as acute or chronic exposure studies. The acute exposure studies use short-term temporal changes in pollution as their source of exposure variability and evaluate short-term changes in health measures in a so called time-series analysis. The study design is that of a longitudinal panel study, in which a population is followed for a certain time and measurements are taken for the same person repeatedly. Thus, every person acts as his or her own control. Panel studies are free of confounding by personal characteristics and are most effective for studying short-term health effects (e.g. changes in lung function tests, inflammatory blood markers and immediate impact on mortality rate) of air pollution.

The development of lung cancer or other chronic diseases may be related to air pollution but will not correlate with short-term changes of exposure and a time series analysis will not be the appropriate form of statistical analysis. Instead, cohort studies offer the opportunity to compare the mortality of lung cancer, for example by using spatial or long-term temporal differences in pollution as a source of exposure variability. If confounding factors, such as smoking, can be measured in a cohort study, their influence can be removed in the statistical analyses.

Results in epidemiological research are usually presented as relative risk (RR) or odds ratio (OR). Without going into the underlying statistics and just to give the reader a rough idea, these numbers can be interpreted as the ratio of two probabilities for an event. The factor indicates the relative magnitude with which a risk is different

between two groups. The relative risk expresses the ratio of the probability of the event occurring in the exposed group versus the control (non-exposed) group. A 95% confidence interval (CI) is defined as the interval between two numbers with an associated probability p which is generated from a random sample of an underlying population, such that if the sampling was repeated numerous times and the CI recalculated from each sample according to the same method, a proportion p of the CIs would contain the population parameter in question. It must be noted that this is not equivalent to a (Bayesian) credible interval. We will cite OR and RR estimates from epidemiological studies along with their CIs to give the reader a perception of the magnitude of the measured association and the precision of this estimate mirrored by the width of the CI.

9.2 Potential Entry Routes for Nanoparticles into the Human Body

In principle, there are three main contact sites of the human organism with the environment: skin, lungs and intestinal tract. The skin provides a relatively thick (10 μm) first barrier against hazardous compounds that is difficult to pass, as opposed to the lungs, where in the gas exchange region the barrier between the alveolar wall and the capillaries is very thin. The air in the lumen of the alveoli is on average only 0.5 μm away from the blood. Epidemiological studies with their main focus on environmental air pollution can only contribute scientific findings to the effect of inhaling particles. The dermal or oral uptake of particles, although probably important in the context of manufactured nanomaterial, has so far not been the objective of epidemiology.

9.2.1 Inhalation and Metabolism of Airborne Particles

Particles can be inhaled when their aerodynamic diameter is less than 10 μm; larger ones will be trapped in the nose. In general, as particle size decreases, the access to the lower respiratory tract and the alveolar region increases [4]. This rule does not apply, however, for particles smaller than 100 nm, as the deposition of nanosized particles becomes governed by diffusional processes rather than gravity. For example, 20 nm UFP are predicted to be deposited in the alveolar region up to 50% and only about 10% each in the nasopharyngeal and tracheobronchial regions; in contrast, about 90% of inhaled UFP around 1 nm in size deposit in the nasopharyngeal region [5], whereas only about 10% of this size deposit in the tracheobronchial and essentially none in the alveolar region.

Inhaled particles will be cleared by various human defense mechanisms. The mucociliary escalator dominates clearance from the upper airways, where particles in the size range 2.5–10 μm deposit. The mucociliary escalator is an efficient transport system pushing the mucus, which covers the airways, together with trapped solid materials towards the mouth. Smaller particles ($PM_{2.5}$ and particles <100 nm)

reach the alveoli of the lung and can only be cleared by alveolar macrophages. The uptake of particles and fibers in the alveoli, not only results in activation of macrophages, but also stimulates the release of chemokines and pro-inflammatory cytokines into the circulation and the production of reactive oxygen species. Although the inflammatory response is a key component of host defense, it can also contribute to persistent inflammation and the pathogenesis of disease [6]. Independently of particle size, there are specific particle characteristics of manufactured nanoparticles such as shape (fibers versus crystals), surface (coated versus uncoated) and surface charges (hydrophilic versus hydrophobic properties), which affect deposition and clearance. Even physiological features of the host organism, such as blood circulation during strenuous physical activity [7, 8] or changed air flow due to pre-existing lung diseases [9], determine the extent and site of deposition of particles.

The impact of inhaled particles on extra pulmonary organs has only recently been recognized. Nemmar *et al.* found in five healthy volunteers that inhaled ultrafine ^{99m}Tc–carbon particles passed rapidly into the systemic blood circulation [10]. The literature on the translocation of very small particles from the lungs into the blood circulation is limited and still conflicting. In experimental animal studies, several authors have reported extra pulmonary translocation of ultrafine particles [11–13] after intratracheal installation or inhalation. However, the amount of ultrafine particles that translocate into blood and extra pulmonary organs differed among these studies.

The difference in deposition characteristics is very important to understand why nanoparticles can probably gain access into the human central nervous system (CNS) directly from deposits on the nasal mucosa via the olfactory epithelium and the olfactory nerves. This pathway has been well demonstrated for inhaled or nasally instilled compounds in animal experiments [14] and – if it also exists in humans – would be very important, as it circumvents the tight blood–brain barrier. Although the olfactory system of rodents requires 50% of the nasal mucosa as compared with only 5% in humans, Elder *et al.* suggest that the direct access of nanoparticles to the brain via the olfactory epithelium and the olfactory nerves is also relevant in humans [15].

9.3 Studies of Environmental Air Pollution in the USA and Europe

9.3.1 PM_{10} and $PM_{2.5}$

Based on health statistics and smog episodes in the past, a temporal correlation between high levels of air pollution and acute increases in morbidity and mortality was observed already in the 1950s (e.g. [16]). Since then, numerous epidemiological studies have shown that the association of daily deaths with daily air pollution is not confined to smog episodes, but exists at levels commonly observed in cities and rural areas.

9.3.1.1 Short-Term Studies

From 1988 to 1993, the averages of the annual mean PM_{10} concentrations at 799 sites monitored by the US EPA declined by 20%. Despite these improvements in air quality, Samet *et al.* [17] reported associations between particle concentrations and the number of deaths per day in 20 of the largest cities and metropolitan areas in the USA from 1987 to 1994 with mean 24-hour PM_{10} concentrations well below the standard. Analyses of the daily number of deaths occurring within an urban region have shown that 10 $\mu g\,m^{-3}$ PM_{10} were associated with an increase of 0.2%. The result is based on recent reanalyses of the National Mortality Morbidity Air Pollution Study (NMMAPS) that included 90 urban areas of the USA [18].

The APHEA (Air Pollution and Health: a European Approach) project was a large multicenter European study investigating the short-term effects of air pollution on health [19]. Twenty-nine European cities provided data on mortality from respiratory and cardiovascular diseases and data on daily ambient air pollution. An increase in PM_{10} by 10 $\mu g\,m^{-3}$ was associated with increases of 0.76% (95% CI: 0.47 to 1.05%) in cardiovascular deaths and 0.58% (95% CI: 0.35 to 0.90%) in respiratory deaths [20].

9.3.1.2 Long-Term Studies

In a prospective cohort study, Dockery *et al.* [21] estimated the effects of air pollution on mortality with data from a 14–16-year mortality follow-up of 8111 adults in six US cities, while controlling for individual risk factors. The adjusted mortality–rate ratio for the most polluted of the cities as compared with the least polluted was 1.26 (95% CI: 1.08 to 1.47). Air pollution was positively associated with death from lung cancer and cardiopulmonary disease but not with death from other causes considered together. A follow-up to the Six Cities Study shows that an overall reduction in $PM_{2.5}$ levels results in reduced long-term mortality risk [22]. Another study [23] dealt with the effect of air pollution control measures and compared for 72 months before and after the banning of coal sales in Dublin, Ireland, on age-standardized death rates. Average black smoke concentrations in Dublin declined by 35.6 $mg\,m^{-3}$ (70%) after the ban on coal sales. Adjusted non-trauma death rates decreased by 5.7% (95% CI: 4 to 7%, $p < 0.0001$), respiratory deaths by 15.5% (95% CI: 12 to 19%, $p < 0.0001$) and cardiovascular deaths by 10.3% (95% CI: 8 to 13%, $p < 0.0001$).

In 1982, the American Cancer Society enrolled approximately 1.2 million adults as part of the Cancer Prevention II study, a large cohort study. Participants completed a questionnaire detailing individual risk factor data (age, sex, race, weight, height, smoking history, education, marital status, diet, alcohol consumption and occupational exposures). These data were linked with air pollution data for metropolitan areas throughout the USA and combined with vital status and cause of death data through 31 December 1998. Fine particulate and sulfur oxide-related pollution were associated with all-cause, lung cancer and cardiopulmonary mortality. Each 10 $\mu g\,m^{-3}$ elevation in fine particulate air pollution was associated with approximately a 4, 6 and 8% increased risk of all-cause, cardiopulmonary and lung cancer mortality, respectively [24].

The impact of air pollution on potentially susceptible patients with pre-existing disease was evaluated by defining cohorts hospitalized for certain diseases: chronic

obstructive pulmonary disease [25], congestive heart disease [26, 27], myocardial infarction [28, 29] or diabetes [29, 30]. All of these studies showed an increased risk of experiencing acute exacerbation of their disease on days with a high concentration of air pollution or shortly afterwards. However, physiological responses with potentially negative effects such as an increase in plasma viscosity [31], in fibrinogen [32] and in C-reactive protein [4] were not restricted to frail populations, but were also observed in samples of randomly collected healthy subjects. Small increases in blood pressure may occur in association with elevated concentrations of ambient particles [33, 34].

9.3.2 Ultrafine Particles (UFP)

A study conducted in Erfurt, Germany, on daily mortality [35] showed comparable and independent increases in mortality associated with fine and ultrafine particles. All particles had a strong seasonality with maximum concentrations in winter. The ultrafine particle concentrations showed a pronounced day of the week effect with concentrations during the weekend 40% lower than during the week. This and a clear increase in the ultrafine particle concentrations during the rush hours suggested that the main source of ultrafine particles was automobile traffic. Associations between health effects and particle number and particle mass concentrations have been observed in different size classes and both immediate effects (lags 0 or 1 day) and delayed effects (lags 4 or 5 days) were found. The effects could be found for total mortality but also for respiratory and cardiovascular causes. There was a tendency for more immediate effects on respiratory causes and more delayed effects for cardiovascular causes. Mortality increased in association with ambient particles after adjustment for season, influenza epidemics, day of the week and meteorology and sensitivity analyses showed the results to be stable.

The first epidemiological evidence of effects of UFP on morbidity was collected in Erfurt on 27 adult asthmatics [36]. A stronger decrease in the peak expiratory flow was observed upon correlation with UFP number concentrations than with fine particle mass concentrations ($PM_{2.5}$). A decrease in respiratory functions, e.g. peak expiratory flow [37], and an increase in symptoms and medication use [38] were associated with elevated particle concentrations of ultrafine particles, independently of fine particles. Inflammatory events in the lungs took several days to develop. It was considered likely that a lag time existed between exposure to ultrafine particles and the acute respiratory health effects of the exposed population. Cumulative effects over 5 days seemed to be stronger than same-day effects. There was an indication that the acute effects of the number of ultrafine particles on respiratory health were stronger than that of the mass of the fine particles [39, 40].

To improve the knowledge on human exposure to particulate matter of different sizes and of different chemical composition in Europe and to develop standards for air quality in Europe, the ULTRA project was initiated. Specifically, the aims of the project were to improve exposure assessment to fine particles by assessing the size distributions, including ultrafine particles and elemental compositions of

fine particles in ambient air in three European cities with different sources of particulate air pollution. Three panel studies were carried out in Amsterdam, The Netherlands, Erfurt, Germany, and Helsinki, Finland, during winter and spring 1998–1999 [34, 41, 42]. In all three cities, about 50 elderly persons with coronary heart disease were followed up for 6 months with biweekly intensive examinations, which included measurements of the function of the heart and lungs, blood pressure and of biomarkers for lung damage from urine. The subjects also kept daily symptom diaries. These studies were limited to the investigation of the acute health effects of short-term exposure by evaluating the impact of day-to-day variation in ambient pollution on health through correlating mortality and morbidity with daily pollution levels. There was an association between exposure to ultrafine particles and cardiovascular morbidity in the population with chronic heart diseases. In Helsinki [43], independent associations between both fine and ultrafine particles and the risk of ST-segment depression in their ECG were observed among subjects with coronary heart disease. ST-segment depression is regarded as an indicator of myocardial ischemia. The study reports increased odds ratios for 45 subjects ranging from 1.03 to 3.29 with an estimated 2-day lag (95% CI: 0.54 to 6.32).

For the Erfurt Panel, the following ECG parameters reflecting myocardial substrate and vulnerability were measured: QT and QTc duration, T-wave amplitude, T-wave complexity and variability of T-wave complexity. Fixed-effect regression analysis was used, adjusting for subject, trend, weekday and meteorology. The analysis showed a significant increase in QT duration in response to exposure to organic carbon, a significant decrease in T-wave amplitude with exposure to ultrafine, accumulation mode and $PM_{2.5}$ particles (particles <2.5 μm in aerodynamic diameter) and a corresponding significant increase of T-wave complexity in association with $PM_{2.5}$ particles for the 24 h before ECG recordings. Variability of T-wave complexity showed a significant increase with organic and elemental carbon in the same time interval. The study provided evidence suggesting an immediate effect of air pollution on repolarization duration, morphology and variability representing myocardial substrate and vulnerability, key factors in the mechanisms of cardiac death [44].

In summary, both fine and ultrafine particles are associated with respiratory and cardiovascular morbidity and mortality and appear to do so independently of each other. There is also epidemiological evidence of similar responses to fine and ultrafine particles, although the size of the effects is often larger for ultrafine than for fine particles (at least on a per mass basis). In general, the relative effects of particulate air pollution are greater for respiratory than cardiovascular mortality. Nevertheless, due to the higher background rate of cardiovascular mortality, the absolute number of deaths attributable to particulate air pollution is much higher for cardiovascular than for respiratory deaths [45–47].

Studies on particles mass concentration indicate that there is linear relationship between PM_{10} and $PM_{2.5}$ and various health indicators (such as cough, symptom exacerbation, bronchodilator use, hospital admissions and mortality) [48] for concentration levels between 0 and 200 $\mu g\,m^{-3}$ and no threshold in particle concentrations below which health would not be jeopardized.

9.4
Cardiovascular Disease

Repeated exposures to elevated ambient air pollution concentrations might not only transiently deteriorate risk factor profiles. Several mechanisms have been hypothesized to contribute to deaths from cardiovascular diseases [49], as shown in Figure 9.2. The inhalation of particles provokes oxidative stress and triggers alveolar and systemic inflammation [2], the linchpin of further patho-physiological mechanisms leading to (1) altered blood rheology favoring coagulation [31, 50], (2) vascular dysfunction [43, 51] and (3) enhanced atherosclerosis, all increasing the risk of a subsequent myocardial infarction; and (4) the alteration of the autonomic nervous control of the heart increases the likelihood of ischemic events and cardiac arrhythmias [52]. Patients with implanted cardioverter defibrillators were more likely to receive interventions with high ambient air pollution 2 days before [53].

An association was found between exposure to traffic and the onset of a myocardial infarction within 1 h afterwards (OR 2.92; 95% CI: 2.22 to 3.83, $p < 0.001$). The time the subjects spent in cars, on public transportation or on motorcycles or bicycles was consistently linked with an increase in the risk of

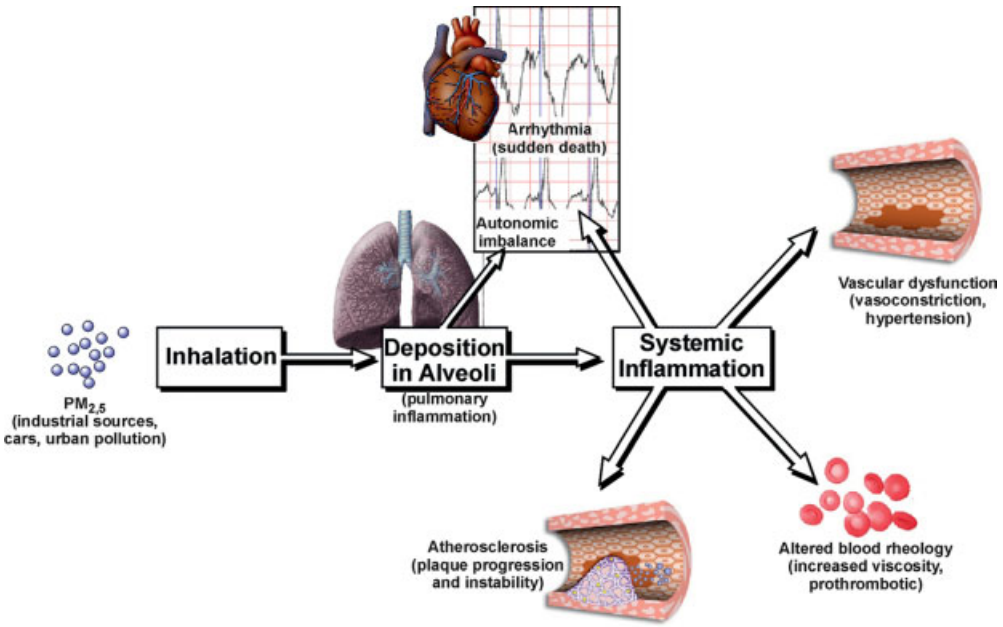


Figure 9.2 General mechanism of cardiovascular disease caused by particulate air pollution exposure. Inhalation of constituents of fine particulate matter ($PM_{2.5}$) can produce pulmonary inflammation. This can directly alter autonomic balance (cardiac rhythm) and lead to a systemic-wide inflammatory response capable of triggering acute and chronic cardiovascular disease. (Reproduced with permission from [49]).

myocardial infarction. Adjusting for the level of exercise on a bicycle or for getting up in the morning changed the estimated effect of exposure to traffic only slightly (OR for myocardial infarction, 2.73; 95% CI: 2.06 to 3.61, $p < 0.001$). The subject's use of a car was the most common source of exposure to traffic; nevertheless, there was also an association between time spent on public transportation and the onset of a myocardial infarction 1 h later [54].

Time-series studies have reported significant reductions in heart rate variability in association with higher ambient air pollution levels in elderly subjects [55, 56] and with occupational exposure concentrations in healthy young men [57]. Decreased heart rate variability reflects a disturbance of cardiac autonomic function and predicts an increased risk for sudden death. Peters [52] reviewed the association between particulate matter and heart disease and concluded that epidemiological studies have demonstrated coherent associations between daily changes in concentrations of ambient particles and cardiovascular disease mortality, hospital admission, disease exacerbation in patients with cardiovascular disease and early physiological responses in healthy individuals consistent with a risk factor profile deterioration.

9.5 Respiratory Disease

The APHEA 2 project investigated short-term health effects of particles in eight European cities and confirmed that particle concentrations were positively associated with increased numbers of hospital admissions for respiratory diseases [58]. Lung diseases attributed to environmental air pollution are (1) deterioration of lung function and respiratory symptoms, exacerbations of chronic obstructive lung disease (COPD) and chronic bronchitis, (2) asthma and allergies and (3) lung cancer.

9.5.1 Deterioration of Lung Function and Respiratory Symptoms

There are few sources of widespread urban air pollution that rival diesel exhaust. The combustion of diesel fuel leads to an emission aerosol that is nanoparticle in primary particle size, but rapidly forms aggregates of 80 nm nanoparticle (accumulation) mode with a solid carbon core. In Europe, exhaust from motor vehicle traffic is considered to contribute to more than 50% of ambient particulate matter (PM_{10}) [59]. For ultrafine particles, the contribution of automobile traffic is even higher. Traffic-related air pollution increases the risk of non-allergic respiratory symptoms and disease. This has been observed in so many epidemiological studies that only one review [60] and one study from The Netherlands are cited here as examples.

Diesel typically is emitted at ground level and ambient diesel levels are highest near highways and busy roads. Brunekreef *et al.* [61] studied children in six areas located near major motorways in The Netherlands and showed that lung function was associated with truck traffic density. The association was stronger in children living closest (<300 m) to the motorways. The results indicated that exposure to

traffic-related air pollution, in particular diesel exhaust particles, leads to reduced lung function in children living near major motorways.

The carbon content of particulate matter, which can be measured as elemental carbon (EC) or organic carbon (OC), served in several epidemiological studies as an exposure surrogate for traffic-related air pollution. It can be shown that alveolar macrophages are loaded with carbonaceous material. Bunn *et al.* [62] collected alveolar macrophages from 22 children aged from 3 months to 16 years with no respiratory symptoms by bronchoalveolar lavage prior to elective surgery. In each child, the size and composition of environmental particles within single sections from 100 separate alveolar macrophages were determined by electron microscopy and microanalysis. Particles consisted of a carbonaceous core and all were ultrafine (<0.1 μm). Other elements such as metals and silicon were not detected. The percentage of particle-containing AM did not change with age, but was increased in children whose parents lived on a main road compared with those living on a quiet residential road (median 10% vs. 3%, $p = 0.014$).

In a recent investigation, Kulkarni *et al.* [63] studied airway macrophages obtained by sputum induction from 64 healthy children in Leicestershire. The authors used the carbon content of airway macrophages as a marker of individual exposure to particulate matter derived from fossil fuel. Each increase in primary PM_{10} of 1.0 μg m^{-3} near each child's home address was associated with an increase of 0.10 μm^2 in the carbon content of airway macrophages and each increase of 1.0 μm^2 in carbon content of airway macrophages was associated with a reduction of 17% in forced expiratory volume in 1 s, of 12.9% in forced vital capacity and of 34.7% in the forced expiratory flow between 25 and 75% of the forced vital capacity. All these reductions were highly statistically significant. The data strengthened the evidence for a causal association between the inhalation of carbonaceous particles and impaired lung function in children.

9.5.2
Asthma and Allergies

Peterson and Saxon [64] reviewed the prevalence of allergic rhinitis and asthma and found an increase in frequency over the past two centuries. They suggested that certain pollutants such as those produced from the burning of fossil fuels, which have been shown to enhance *in vitro* and *in vivo* IgE production, may be partly responsible for the increased prevalence of allergic respiratory disease.

Laboratory studies in humans and animals have shown that particulate toxic pollutants, particularly diesel exhaust particulates, can enhance allergic inflammation and can induce allergic immune responses. Although road traffic pollution from automobile exhausts may be a risk factor for atopic sensitization, the evidence in support of this view remains contradictory [65, 66]. Some investigators have reported a clear association between the prevalence of allergy and road traffic-related air pollution, whereas such a difference was not observed in other studies.

Asthma is characterized by airway obstruction, with air trapping and increases in lung residual volume. Increases in alveolar volume would be expected to enhance diffusional deposition, the primary mechanism of deposition for UFPs, although

impaired alveolar ventilation would counter this increase. A panel study of subjects with asthma [36] found that peak flow varied more closely with the 5-day mean of UFP number than with fine particle mass concentration, suggesting that the UFP component of fine particle pollution contributes to airway effects in asthmatics. Penttinen *et al*. [40] noted that UFP number concentrations tended to be inversely but not significantly associated with measures of lung function. However, some epidemiological studies have not found associations between UFP exposure and health effects [41]. Inhaled UFPs have a high predicted deposition efficiency in the pulmonary region [67]. Thus, the expected number of particles retained in the lung with each breath is greater for UFPs than for larger particles. A study on 16 subjects with mild to moderate asthma demonstrated an efficient respiratory deposition of ultrafine particles especially in subjects with asthma [68]. Deposition was measured during spontaneous breathing at rest and exercise. The deposition fraction increased during exercise by particle number and mass concentration and reached a maximum for the smallest particles. When both the increased deposition fraction and minute ventilation were considered, the total number of particles retained in the lung was 74% greater in subjects with asthma than in healthy subjects. Thus, people with asthma have a higher total respiratory dose of UFPs for a given exposure, which may contribute to their increased susceptibility to the health effects of air pollution.

The association between particulate air pollution and asthma medication use and symptoms was assessed in a panel study of 53 adult asthmatics in Erfurt, Germany, in winter 1996–97. The results suggest that reported asthma medication use and symptoms increase in association with particulate air pollution (0.01–0.1 μm in diameter) and gaseous pollutants such as nitrogen dioxide [38].

9.5.3 Lung Cancer

Large cohort studies in the USA and Europe suggest that air pollution may increase lung cancer risk. For example, the Adventist Health Study found an increased risk of newly diagnosed lung cancers in a cohort study of 6338 non-smoking, white Californian adults, followed from 1977 to 1992, associated with elevated long-term ambient concentrations of PM_{10} [69]. In the Harvard Six Cities Study air pollution was positively associated with death from lung cancer [21]. Also, in the Cancer Prevention II study of the American Cancer Society it was quantitatively evaluated that for each $10\,\mu g\,m^{-3}$ elevation in fine particulate air pollution there is an increase of 8% in lung cancer mortality [24].

In Europe, the association between incidence of lung cancer and long-term air pollution exposure was investigated in a cohort of Oslo men followed from 1972–73 to 1998. During the follow-up period, 418 men developed lung cancer. For a $10\,\mu g\,m^{-3}$ increase in average home address exposure to nitrogen oxides NO_x – a traffic-related gas of urban air pollution – between 1974 and 1978, the risk of developing lung cancer increased by 8%, controlling for age, smoking habits and length of education [70]. To estimate the relationship between air pollution and lung cancer, a nested case–control study was set up within EPIC (European Prospective

Investigation on Cancer and Nutrition). There was a non-significant association between lung cancer and residence nearby heavy traffic roads as an indicator of exposure to air pollution [71].

Since epidemiological studies in railroad workers [72, 73] had suggested that diesel exhaust was a human lung carcinogen, public concern was aroused. Inhalation studies in rats exposed to high levels of diesel exhaust had also resulted in lung tumors [74, 75], although the results were not replicable in other species. Later studies in humans in motor exhaust-related occupations such as truck, forklift and other drivers of diesel vehicles, operators of heavy construction equipment, farm workers operating diesel equipment, bus maintenance garage workers and loading dock workers [76–81] supported the evidence that diesel exhaust might be a potential occupational carcinogen. Bhatia *et al.* [82] evaluated the relation between occupational exposure to diesel exhaust and cancer of the lung in a meta-analysis of 29 published cohort and case–control studies. Pooled effect measures weighted by study precision indicated a statistically significant increased relative risk of lung cancer from occupational exposure to diesel exhaust. This meta-analysis supported a causal association between increased risk of lung cancer and exposure to diesel exhaust. The International Agency for Research on Cancer (IARC) classified diesel exhaust as probably carcinogenic to humans (Group 2A) [83].

Numerous experimental studies *in vitro* and *in vivo* have provided unambiguous evidence for genotoxicity of air pollution. Several studies found an association between external measures of exposure to air pollution and increased levels of DNA adducts, with an apparent leveling off of the dose–response relationship. Due to the organic extracts of particulate matter, especially various polycyclic aromatic hydrocarbons (PAHs), particulate air pollution induces oxidative damage to DNA. Lung cancer develops through a series of progressive pathological changes occurring in the respiratory epithelium. Molecular alterations such as loss of heterozygosity, gene mutations and aberrant gene promoter methylation have emerged as potentially promising molecular biomarkers of lung carcinogenesis [84].

9.6 Diseases of the Central Nervous System

Transitional metals such as copper, manganese and iron have been associated with pathological lesions of the brain characteristic of a variety of neurodegenerative diseases such as Parkinson's disease, Alzheimer's disease and amyotrophic lateral sclerosis [85]. Metals are essential in the synthesis of DNA and RNA and are also cofactors of numerous enzymes, particularly those involved in respiration. In addition, several modifications indicative of oxidative stress have been described in association with neurons, neurofibrillary tangles and senile plaques in Alzheimer's disease.

These findings became even more important after inhalation experiments with rats by Oberdörster [14] suggested that ^{13}C-labeled nanoparticles with a size about 35 nm may migrate along the olfactory nerve into the olfactory bulb of the brain after

deposition on the olfactory mucosa in the nasal region. If this observation proves to be a route of entry of nanoparticles into the brain, it would circumvent the tight blood–brain barrier and might play a role in neurodegenerative disease.

In Mexico, neuropathological findings for 32 dogs from Southwest Metropolitan Mexico City, a highly polluted urban region, were compared with those for eight dogs from Tlaxcala, a less polluted, control city [86]. The report describes early and progressive alterations in the nasal respiratory and olfactory mucosa. Early changes included expression of nuclear neuronal NF-kappaB and iNOS in cortical endothelial cells occurring at ages 2 and 4 weeks; subsequent damage included alterations of the blood–brain barrier (BBB), degenerating cortical neurons, apoptotic glial white matter cells, deposition of apolipoprotein E (apoE)-positive lipid droplets in smooth muscle cells and pericytes, non-neuritic plaques and neurofibrillary tangles. The authors concluded that persistent pulmonary inflammation and deteriorating olfactory and respiratory barriers may play a role in the degenerative neuropathology observed in the brains of highly exposed dogs.

The greatest exposure to metals is likely to occur in occupational settings such as mining, alloy production and welding. Welding and laser operations are well known for their potential to produce large numbers of nanosized particles [87] (see Table 9.1), for example manual metal arc welding with covered electrodes releases particles in the size ranges 20–400 and 10–20 nm for gas-shielded metal arc welding.

A review by Tjälve and Henriksson [88] deals with the mechanism of uptake and transport of metals in the olfactory system. Metals discussed are mainly manganese, cadmium, nickel and mercury. Manganese was found to have a unique capacity to be taken up via the olfactory pathways and be passed transneuronally to other parts of the brain. It is considered that the occupational neurotoxicity of inhaled manganese may be related to an uptake of the metal into the brain via the olfactory pathways. Airborne manganese levels during welding practice were measured in a study on 97 welders engaged in electric arc welding in a vehicle manufacturer. Ambient manganese levels in welders' breathing zone were the highest inside the vehicle and the lowest in the center of the workshop. Serum levels of manganese in welders were about three-fold ($p < 0.01$) higher than those of controls [89].

The neurotoxicity of manganese has been known since the nineteenth century. In 1837, Couper described "manganism" characterized by extrapyramidal motor system dysfunction and in particular, Parkinson's disease and dystonia. Manganese is rapidly cleared from the blood by the liver, but elimination from the central nervous system takes a very long time. The neurological signs of manganism have received close attention because they resemble several clinical disorders collectively described as extrapyramidal motor system dysfunction and, in particular, Parkinson's disease and dystonia. Semchuk *et al.* [90], in a population-based case–control study in Calgary, Alberta, reported no significant increase in risk of Parkinson's disease associated with a history of rural exposure to manganese. In contrast, Gorell *et al.* [91], in a population-based case–control study at Henry Ford Health System (HFHS), Detroit, MI, found a significant association of Parkinson's disease with manganese with more than 20 years of occupational contact (OR 10.6, 95% CI: 1.06 to 105.83),

although only three cases and one control subject had such a lengthy exposure to manganese. The small number reporting such an exposure requires that the association be interpreted with caution.

Racette *et al*. performed a case–control study [92] that compared the clinical features of 15 career welders with two control groups with idiopathic Parkinson's disease. One control group was ascertained sequentially to compare the frequency of clinical features and the second control group was sex- and age-matched to compare the frequency of motor fluctuations. Welders were exposed to a mean of 47 144 welding hours. Welders had a younger age at onset (46 years) of Parkinson's disease compared with sequentially ascertained controls (63 years; $p < 0.0001$). There was no difference in frequency of tremor, bradykinesia, rigidity, asymmetric onset, postural instability, family history, clinical depression, dementia or drug-induced psychosis between the welders and the two control groups. Parkinsonism in welders was distinguished clinically only by age at onset, suggesting that welding may be a risk factor for Parkinson's disease.

9.7
Particulate Air Pollution at the Workplace

The inhalation of dust at work has historically always been and still remains one of the most important causes of ill-health related to work. Dust is responsible for serious and disabling diseases such as pneumoconiosis, interstitial lung disease and fibrosis, lung cancer and asthma. Research related to dust exposure at the workplace has historically always focused on the effects on the lung and only recently – probably triggered by environmental epidemiological studies – has started to look for implications regarding the cardiovascular system. For example, in a retrospective cohort study, an increased risk of mortality due to ischemic heart disease (OR 1.32, 95% CI: 1.13 to 1.55) was observed among heavy equipment operators [93] exposed to diesel motor emissions. The paragraphs below summarize the effects of particle inhalation on the lungs, as the impact of dust on the lungs has attracted most attention over the last 50 years in occupational medicine.

Pneumoconiosis, one of the civilization's oldest known occupational respiratory diseases, is caused by the inhalation of dust and is characterized by a reactive reparative process that leads to the formation of nodular fibrotic changes in the lungs. Gradually, the alveoli of the lungs become replaced by fibrotic tissue, causing an irreversible loss of the tissue's ability to transfer oxygen into the bloodstream. The fibrogenic potential of inorganic dusts varies considerably, with silica and asbestos having greater fibrogenic potential than coal dusts, iron and man-made mineral fibers. Silicosis, a condition of fibrosis of the lungs marked by shortness of breath and resulting from prolonged inhalation of crystalline silica dust, is also associated with lung cancer. The hypothesis that diffuse fibrotic disorders of the lung are associated with increased lung cancer risk stems from early observations at autopsy that lung cancer was often associated with fibrosis of the lung. This finding could be substantiated in a meta-analysis of lung cancer and silicosis [94].

The pooled RR estimate for the 23 studies that could be combined was 2.2, with a 95% CI of 2.1 to 2.4. The authors considered the association between silicosis and lung cancer as causal, either due to silicosis itself or due to a direct effect of the underlying exposure to silica. The IARC concluded that there is sufficient evidence for carcinogenicity of crystalline silica in humans [95].

In 2006, the IARC in Lyon, France, reassessed the carcinogenicity of carbon black and titanium dioxide and the results will be published as volume 93 of the IARC Monographs. Both substances are produced in the particulate form. Exposure to carbon black occurs mainly with aggregates with particle size 50–600 nm. The primary particles of titanium dioxide are typically 200–300 nm in diameter, but larger aggregates and agglomerates are readily formed. Ultrafine grades of titanium dioxide (10–50 nm) are used in sunscreens and plastics to block ultraviolet light. For carbon black and titanium dioxide, the Monograph Working Group of the IARC concluded that existing epidemiological studies provided inadequate evidence of carcinogenicity, but overall – taking also into account the sufficient evidence of carcinogenicity from toxicological experiments in laboratory animals – carbon black and titanium dioxide were classified as possibly carcinogenic to human beings (Group 2B).

Asbestos is the name given to a group of minerals that occur naturally as bundles of fibers which can be separated into thin threads. These fibers are not affected by heat or chemicals and do not conduct electricity. For these reasons, asbestos has been widely used in many industries. When asbestos fibers are set free and inhaled, exposed individuals are at risk of developing an asbestos-related disease such as asbestosis, lung cancer, mesothelioma of the pleura or peritoneum and other cancers, such as those of the larynx and oropharynx [96]. Asbestos remains the primary occupational carcinogenic substance affecting workers all over the world. Outside the workplace, asbestos is second only to tobacco as an environmental source of cancer.

Carbon nanotubes have attracted a great deal of attention due to their potential technological applications, but also – with their shape and physical appearance resembling those of asbestos fibers – have aroused considerable concern. Even though carbon nanotubes consist only of carbon, it does not seem adequate to classify them (and also fullerenes in general) as graphite. Varying physical shapes might well be associated with entirely different properties. Toxicity studies on nanomaterial will be extremely complex, as 10, 20 m, 50 and 500 nm titanium dioxide crystals, for example, will all be different. Despite the prominence of carbon nanotubes in nanotechnology, exploration of their interactions with biological materials remains sparse. In part, this reflects the challenge of observing nanotubes in biological environments. Single-walled carbon nanotubes in tissues evade detection by elemental analysis, as they contain only carbon, and often also by electron microscopy. One successful method described recently is near-infrared fluorescence microscopy [97].

Lam *et al.* [98] and Warheit *et al.* [99] have published the results of studies of the toxicity of single-walled carbon nanotubes in mice and rats, respectively. Using an intratracheal route of administration, they compared different means of nanotube production with effects of carbon black and quartz particles. In Lam *et al.*'s study,

the nanotubes were found to produce dose-dependent lung lesions. The effects of carbon black were distinctively different. The study by Warheit *et al.* was more comprehensive. It showed multifocal pulmonary granuloma but without evidence of ongoing pulmonary inflammation or cellular proliferation. These effects were different from those of quartz, carbon black and graphite. The conclusion from these two studies was that carbon nanotubes have different toxicological properties from other forms of carbon [100].

Another example of how the physical shape of nanoparticles will have an impact on cellular function was provided by Zhao *et al.* [101]. C_{60} fullerenes were found to bind to double-stranded DNA, either at the hydrophobic ends or at the minor groove of the nucleotide. They also bound to single-stranded DNA, deforming the nucleotides significantly. When the DNA molecule was damaged (specifically, a gap was created by removing a piece of the nucleotide from one helix), fullerenes could stably occupy the damaged site. The authors speculated that this strong association may negatively impact the self-repairing process of the double stranded DNA.

There have been two reports [102, 103] describing fibrotic lung disease that developed after exposure to indium tin oxide (ITO). ITO is a sintered alloy containing a large proportion of indium oxide and a small proportion of tin oxide and is used in the making of thin-film transistor liquid crystal displays (LCDs) for television screens, portable computer screens, cellphone displays and video monitors. One patient was engaged in wet surface grinding of ITO targets for 3 years and the other was exposed for 4 years to ITO as an aerosol while making transparent conductive films. Both patients came from the same factory and developed pulmonary fibrosis. One died of bilateral pneumothorax. The autopsy demonstrated interstitial pneumonia with numerous fine particles scattered throughout the lungs. Intrapulmonary deposition of indium and tin was shown by X-ray energy spectrometry in the fine particles. The level of serum indium was extremely high. According to Chonan and Taguchi [104], among 115 workers from the same metal plant, 14 revealed interstitial fibrosis on chest CT.

In experimental or occupational settings, exposure to airborne particles, fibers and fumes have long been recognized as causing fibrotic lung disease, with idiopathic pulmonary fibrosis (IPF) being the most distinct entity. IPF is a progressive and devastating lung disorder with a median survival of 2–4 years after diagnosis [105], yet the course of individual patients is highly variable. In various populations, the prevalence estimates for IPF have ranged from 6 to 32 per 100 000 persons. Approximately two-thirds do not have a known cause (idiopathic), whereas one-third result from known causes such as sarcoidosis, connective tissue disease, complication of certain drug exposures or radiation and occupational exposures.

In a meta-analysis of six case–control studies conducted in three countries, several exposures were significantly associated with IPF, including ever smoking (OR 1.58, 95% CI: 1.27 to 1.97), agriculture/farming (OR 1.65, 95% CI: 1.20 to 2.26), livestock (OR 2.17, 95% CI: 1.28 to 3.68), wood dust (OR 1.94, 95% CI: 1.34 to 2.81), metal dust (OR 2.44, 95% CI: 1.74 to 3.40) and stone/sand (OR 1.97, 95% CI: 1.09 to 3.55) [106]. Although multiple exogenous agents can initiate an inflammatory

alveolitis and result in interstitial lung disease, it is likely that the underlying pathogenetic mechanisms that mediate the development and progression of pulmonary fibrosis are similar. The natural history and the pathogenic mechanisms remain unknown; the long-prevailing hypothesis sustains the idea that chronic inflammation plays an essential role. According to this hypothesis, the alveolar epithelial alterations are caused by an unresolved inflammatory process. More recently, however, research emphasis changed from a focus on inflammation to alveolar epithelial injury, fibrogenesis in fibroblastic foci [107].

New cases of occupational asthma in France are collected by a national surveillance program, based on voluntary reporting, named Observatoire National des Asthmes Professionnels (ONAP) [108]. The most frequently incriminated agents were flour (20.3%), isocyanates (14.1%), latex (7.2%), aldehyde (5.9%), persulfate salts (5.8%) and wood dust (3.7%). The highest risks of occupational asthma were found in bakers and pastry makers, car painters, hairdressers and wood workers. Another voluntary surveillance scheme, SHIELD, for occupational asthma is located in the West Midlands, a highly industrialized region of the UK [109]. Spray painters represented the occupation at the highest risk of developing occupational asthma, followed by electroplaters, rubber and plastic workers, bakery workers and molders. Although the percentage of reported cases was low among healthcare workers, there was an increasing trend. Isocyanates still remained the most common causative agents, with 190 (17.3%) out of the total 1097 cases reported to the surveillance scheme in 7 years. There was a decrease in the reported cases due to colophony (from 9.5 to 4.6%) and flour and wheat (from 8.9 to 4.9%). There was an increase of reported cases due to latex (from 0.4 to 4.9%) and glutaraldehyde (from 1.3 to 5.6%).

The best evidence to support the hypothesis that it is the ultrafine fraction of PM_{10} that is responsible for the adverse health effects comes from toxicology [110]. Ultrafine particles have extra toxicity and inflammogenicity compared with fine, respirable particles of the same material when delivered at the same mass dose. This has been shown for a range of different materials of generally low toxicity, such as carbon black and titanium dioxide. Ultrafine particles cause inflammation in the lungs even when composed of relatively low toxicity materials. The mechanism of the induction of inflammation appears to be via oxidative stress and Ca^{2+} and signaling perturbations [111]. Particularly the large surface area of ultrafine particles provides a unique interface for catalytic reaction of surface-located agents with biological targets such as proteins and cells [112]. *In vivo* experiments showed that within hours after the respiratory system is exposed to nanoparticles, they may appear in many compartments of the body, including the liver, heart and nervous system. Inhalation experiments with rats resulted in ultrafine titanium dioxide particles being found on the luminal side of airways and alveoli, in all major lung tissue compartments and cells and within capillaries. Particles within cells were not membrane bound and hence had direct access to intracellular proteins, organelles and DNA, which may greatly enhance their toxic potential [113].

Current aerosol standards at the workplace are expressed in terms of mass concentration of particulate matter conforming to a particle size fraction.

Instruments able to measure particles below 100 nm were first introduced for environmental studies and are not in use for operational supervision due to a lack of regulations. This is surprising, as nanoparticles had been around the workplace for a long time before nanotechnology appeared on the scene. Aerosols in workplace environments may be derived from mechanical processes (e.g. the breaking or fracture of solid or liquid material) and may come from a variety of sources such as mining, chemical manufacture, textiles and agriculture. The size range can be anything from micrometer and submicrometer particles down to 100 nm and below. In a workplace study [114], nanoparticles occurring in different work processes were measured. Typical examples include welding fumes, metal fumes, soldering fumes, plasma cutting fumes, plasma spraying emissions, polymer fumes, vulcanization fumes, powder coating emissions, oil mists, aircraft engine emissions, bakery oven emissions, meat smokery fumes and particulate diesel motor emissions. The particles were for the most part the products of condensation in thermal and chemical reactions, the primary particles created having a size of only a few nanometers. The most frequently occurring particle size was between 160 and 300 nm. The total concentration of all particles in the measurement range 14–673 nm was between 500 000 and 2 500 000 particles per cm^3. A comparison of the occurrence of nanoparticles in different workplace atmospheres is given in Table 9.1 [87].

Most plasma and laser deposition and aerosol processes are performed in evacuated or at least closed reaction chambers. Therefore, exposure to nanoparticles is more likely to happen after the manufacturing process itself, except in those cases of failures during the processing. In processes involving high pressure (e.g. supercritical fluid techniques) or with high-energy mechanical forces, particle release could occur in the case of failure of sealing of the reactor or the mills. Furthermore, many particles, including metallic particles, are highly pyrophoric and there is a considerable risk of dust explosions [116].

Table 9.1 Comparison of nanoparticles in workplace air [87].

Workplace	Total concentration in measurement range 14–673 nm (particles cm^{-3})	Maximum of number concentration (nm range)
Outdoor, office	Up to 10 000	
Silicon melt	100 000	280–520
Metal grinding	Up to 130 000	17–170
Soldering	Up to 400 000	36–64
Plasma cutting	Up to 500 000	120–180
Bakery	Up to 640 000	32–109
Airport field	Up to 700 000	<45
Hard soldering	54 000–3 5000 000	33–126
Welding	100 000–40 000 000	40–600

References

1 G. Oberdörster, J. Ferin, B. E. Lehnert, *Environ. Health Perspect.* 1994, **102**, Suppl 5, 173–179.

2 A. Seaton, W. MacNee, K. Donaldson, D. Godden, *Lancet* 1995, **345**, 176–178.

3 H. E. Wichmann, C. Spix, T. Tuch, G. Woelke, A. Peters, J. Heinrich, W. G. Kreyling, J. Heyder, *Health Effects Institute Research Report 2000*, Health Effects Institute, Boston, 2001.

4 International Commisson on Radiological Protection (ICRP) , Human respiratory tract model for radiological protection. ICRP Publication 66. *Ann. ICRP* 1994, **24**, 1–3.

5 D. L. Swift, N. Montassier, P. K. Hopke, K. Karpen-Hayes, Y. S. Cheng, Y. F. Su, J. C. Strong, *J. Aerosol Sci.* 1992, **23**, 65–72.

6 K. E. Driscoll, J. M. Carter, D. G. Hassenbein, B. Howard, *Environ Health Perspect.* 1997, **105**, Suppl 5, 1159–1164.

7 P. A. Jaques and C. S. Kim, *Inhal. Toxicol.* 2000, **12**, 715–731.

8 C. C. Daigle, D. C. Chalupa, F. R. Gibb, P. E. Morrow, G. Oberdörster, M. J. Utell, M. W. Frampton, *Inhal. Toxicol.* 2003, **15**, 539–552.

9 J. S. Brown, K. L. Zeman, W. D. Bennett, *Am. J. Respir. Crit Care Med.* 2002, **166**, 1240–1247.

10 A. Nemmar, P. H. Hoet, B. Vanquickenborne, D. Dinsdale, M. Thomeer, M. F. Hoylaerts, H. Vanbilloen, L. Mortelmans, B. Nemery, *Circulation* 2002, **105**, 411–414.

11 W. G. Kreyling, M. Semmler, F. Erbe, P. Mayer, S. Takenaka, H. Schulz, G. Oberdörster, A. Ziesenis, *J. Toxicol. Environ Health A* 2002, **65**, 1513–1530.

12 G. Oberdörster, Z. Sharp, V. Atudorei, A. Elder, R. Gelein, A. Lunts, W. Kreyling, C. Cox, *J. Toxicol. Environ. Health A* 2002, **65**, 1531–1543.

13 S. Takenaka, E. Karg, C. Roth, H. Schulz, A. Ziesenis, U. Heinzmann, P. Schramel, J. Heyder, *Environ Health Perspect.* 2001, **109**, Suppl 4, 547–551.

14 G. Oberdörster, *Inhal. Toxicol.* 2004, **16**, 437–445.

15 A. Elder, R. Gelein, V. Silva, T. Feikert, L. Opanashuk, J. Carter, R. Potter, A. Maynard, Y. Ito, J. Finkelstein, G. Oberdörster, *Environ Health Perspect.* 2006, **114**, 1172–1178.

16 W. P. D. Logan, *Lancet* 1953, *i*, 336–338.

17 J. M. Samet, F. Dominici, F. C. Curriero, I. Coursac, S. L. Zeger, *N. Engl. J. Med.* 2000, **343**, 1742–1749.

18 F. Dominici, A. McDermott, M. Daniels, S. L. Zeger and J. M. Samet, (eds.) Revised Analyses of the National Morbidity, Mortality, and Air Pollution Study (NMMAPS), Part II: Mortality Among Residents of 90 Cities. In *Revised Analyses of Time-series Studies of Air Pollution and Health*, Health Effects Institute, Boston, 2003, 9–24.

19 K. Katsouyanni, G. Touloumi, E. Samoli, A. Gryparis, A. Le Tertre, Y. Monopolis, G. Rossi, D. Zmirou, F. Ballester, A. Boumghar, H. R. Anderson, B. Wojtyniak, A. Paldy, R. Braunstein, J. Pekkanen, C. Schindler, J. Schwartz, *Epidemiology* 2001, **12**, 521–531.

20 A. Analitis, K. Katsouyanni, K. Dimakopoulou, E. Samoli, A. K. Nikoloulopoulos, Y. Petasakis, G. Touloumi, J. Schwartz, H. R. Anderson, K. Cambra, F. Forastiere, D. Zmirou, J. M. Vonk, L. Clancy, B. Kriz, J. Bobvos, J. Pekkanen, *Epidemiology* 2006, **17**, 230–233.

21 D. W. Dockery, C. A. Pope, X. Xu, J. D. Spengler, J. H. Ware, M. E. Fay, B. G. Ferris, F. E. Speizer, *N. Engl. J. Med.* 1993, **329**, 1753–1759.

22 F. Laden, J. Schwartz, F. E. Speizer, D. W. Dockery, *Am. J. Respir. Crit. Care Med.* 2006, **173**, 667–672.

23 L. Clancy, P. Goodman, H. Sinclair, D. W. Dockery, *Lancet* 2002, **360**, 1210–1214.

24 C. A. Pope, R. T. Burnett, M. J. Thun, E. E. Calle, D. Krewski, K. Ito, G. D. Thurston, *J. Am. Med. Assoc.* 2002, **287**, 1132–1141.

25 J. Sunyer, J. Schwartz, A. Tobias, D. Macfarlane, J. Garcia, J. M. Anto, *Am. J. Epidemiol.* 2000, **151**, 50–56.

26 M. S. Goldberg, R. T. Burnett, J. C. BailarIII, R. Tamblyn, P. Ernst, K. Flegel, J. Brook, Y. Bonvalot, R. Singh, M. F. Valois, R. Vincent, *Environ. Health Perspect.* 2001, **109**, Suppl 4, 487–494.

27 H. J. Kwon, S. H. Cho, F. Nyberg, G. Pershagen, *Epidemiology* 2001, **12**, 413–419.

28 S. von Klot, G. Wölke, T. Tuch, J. Heinrich, D. W. Dockery, J. Schwartz, W. G. Kreyling, H.-E. Wichmann, A. Peters, *Eur. Respir. J.* 2002, **20**, 691–720.

29 T. F. Bateson and J. Schwartz, *Epidemiology* 2004, **15**, 143–149.

30 C. A. Pope, R. T. Burnett, M. J. Thun, E. E. Calle, D. Krewski, K. Ito, G. D. Thurston, *J. Am. Med. Assoc.* 2002, **287**, 1132–1141.

31 A. Peters, A. Doring, H. E. Wichmann, W. Koenig, *Lancet* 1997, **349**, 1582–1587.

32 J. Pekkanen, E. J. Brunner, H. R. Anderson, P. Tiittanen, R. W. Atkinson, *Occup. Environ. Med.* 2000, **57**, 818–822.

33 A. Ibald-Mulli, J. Stieber, H. E. Wichmann, W. Koenig, A. Peters, *Am. J. Public Health* 2001, **91**, 571–577.

34 A. Ibald-Mulli, K. L. Timonen, A. Peters, J. Heinrich, G. Wolke, T. Lanki, G. Buzorius, W. G. Kreyling, J. de Hartog, G. Hoek, H. M. Ten Brink, J. Pekkanen, *Environ. Health Perspect.* 2004, **112**, 369–377.

35 H. E. Wichmann, C. Spix, T. Tuch, G. Wolke, A. Peters, J. Heinrich, W. G. Kreyling, J. Heyder, *Health Effects Institute Research Report*, Health Effects Institute, Boston, 2000, 5–86

36 A. Peters, H. E. Wichmann, T. Tuch, J. Heinrich, J. Heyder, *Am. J. Respir. Crit. Care Med.* 1997, **155**, 1376–1383.

37 J. Pekkanen, K. L. Timonen, J. Ruuskanen, A. Reponen, A. Mirme, *Environ. Res.* 1997, **74**, 24–33.

38 S. von Klot, G. Wolke, T. Tuch, J. Heinrich, D. W. Dockery, J. Schwartz, W. G. Kreyling, H. E. Wichmann, A. Peters, *Eur. Respir. J.* 2002, **20**, 691–702.

39 P. Penttinen, K. L. Timonen, P. Tiittanen, A. Mirme, J. Ruuskanen, J. Pekkanen, *Environ. Health Perspect.* 2001, **109**, 319–323.

40 P. Penttinen, K. L. Timonen, P. Tiittanen, A. Mirme, J. Ruuskanen, J. Pekkanen, *Eur. Respir. J.* 2001, **17**, 428–435.

41 J. J. De Hartog, G. Hoek, A. Peters, K. L. Timonen, A. Ibald-Mulli, B. Brunekreef, J. Heinrich, P. Tiittanen, J. H. van Wijnen, W. Kreyling, M. Kulmala, J. Pekkanen, *Am. J. Epidemiol.* 2003, **157**, 613–623.

42 J. Ruuskanen, T. Tuch, H. Ten Brink, A. Peters, A. Khlystov, A. Mirme, G. P. Kos, B. Brunekreef, H. E. Wichmann, G. Buzorius, M. Vallius, J. Pekkanen, *Atmos. Environ.* 2001, **35**, 3729–3738.

43 J. Pekkanen, A. Peters, G. Hoek, P. Tiittanen, B. Brunekreef, J. de Hartog, J. Heinrich, A. Ibald-Mulli, W. G. Kreyling, T. Lanki, K. L. Timonen, E. Vanninen, *Circulation* 2002, **106**, 933–938.

44 A. Henneberger, W. Zareba, A. Ibald-Mulli, R. Rückerl, J. Cyrys, J. P. Couderc, B. Mykins, G. Woelke, H. E. Wichmann, A. Peters, *Environ. Health Perspect.* 2005, **113**, 440–446.

45 D. W. Dockery, *Environ. Health Perspect.* 2001, **109**, Suppl 4, 483–486.

46 M. W. Frampton, *Environ. Health Perspect.* 2001, **109**, Suppl 4, 529–532.

47 C. A. Pope III, R. T. Burnett, G. D. Thurston, M. J. Thun, E. E. Calle, D. Krewski, J. J. Godleski, *Circulation* 2004, **109**, 71–77.

48 WHO, *Air Quality Guidelines for Europe*, 2nd edn., WHO Regional Office for Europe, Copenhagen, 2000, http://www.euro.who.int/air/Activities/20020620_1.

49 R. D. Brook, J. R. Brook, S. Rajagopalan, *Curr. Hypertens. Rep.* 2003, **5**, 32–39.

50 R. Rückerl, A. Ibald-Mulli, W. Koenig, A. Schneider, G. Woelke, J. Cyrys, J. Heinrich, V. Marder, M. Frampton,

H. E. Wichmann, A. Peters, *Am. J. Respir. Crit. Care Med.* 2006, **173**, 432–441.

51 A. Peters, M. Frohlich, A. Doring, T. Immervoll, H. E. Wichmann, W. L. Hutchinson, M. B. Pepys, W. Koenig, *Eur. Heart J.* 2001, **22**, 1198–1204.

52 A. Peters, *Toxicol. Appl. Pharmacol.* 2005, **207**, 2 Suppl., 477–482.

53 A. Peters, E. Liu, R. L. Verrier, J. Schwartz, D. R. Gold, M. Mittleman, J. Baliff, J. A. Oh, G. Allen, K. Monahan, D. W. Dockery, *Epidemiology.* 2000, **11**, 11–17.

54 A. Peters, S. von Klot, M. Heier, I. Trentinaglia, A. Hormann, H. E. Wichmann, H. Lowel, *N. Engl. J. Med.* 2004, **351**, 1721–1730.

55 D. R. Gold, A. Litonjua, J. Schwartz, E. Lovett, A. Larson, B. Nearing, G. Allen, M. Verrier, R. Cherry, R. Verrier, *Circulation* 2000, **101**, 1267–1273.

56 C. A. Pope, III, R. L. Verrier, E. G. Lovett, A. C. Larson, M. E. Raizenne, R. E. Kanner, J. Schwartz, G. M. Villegas, D. R. Gold, D. W. Dockery, *Am. Heart J.* 1999, **138**, 890–899.

57 S. R. Magari, R. Hauser, J. Schwartz, P. L. Williams, T. J. Smith, D. C. Christiani, *Circulation* 2001, **104**, 986–991.

58 R. W. Atkinson, H. R. Anderson, J. Sunyer, J. Ayres, M. Baccini, J. M. Vonk, A. Boumghar, F. Forastiere, B. Forsberg, G. Touloumi, J. Schwartz, K. Katsouyanni, *Am. J. Respir. Crit. Care Med.* 2001, **164**, 1860–1866.

59 N. Kunzli, R. Kaiser, S. Medina, M. Studnicka, O. Chanel, P. Filliger, M. Herry, F. Horak, Jr., V. Puybonnieux-Texier, P. Quenel, J. Schneider, R. Seethaler, J. C. Vergnaud, H. Sommer, *Lancet* 2000, **356**, 795–801.

60 D. J. Ward and J. G. Ayres, *Occup. Environ. Med.* 2004, **61**, e13ff.

61 B. Brunekreef, N. A. H. Janssen, J. d. Hartog, H. Harssema, M. Knape, P. v. Vliet, *Epidemiology* 1997, **8**, 298–303.

62 H. J. Bunn, D. Dinsdale, T. Smith, J. Grigg, *Thorax* 2001, **56**, 932–934.

63 N. Kulkarni, N. Pierse, L. Rushton, J. Grigg, *N. Engl. J. Med.* 2006, **355**, 21–30.

64 B. Peterson and A. Saxon, *Ann. Allergy Asthma Immunol.* 1996, **77**, 263–268.

65 R. Polosa, S. Salvi, G. U. Di Maria, *Arch. Environ. Health* 2002, **57**, 188–193.

66 J. Heinrich and H. E. Wichmann, *Curr. Opin. Allergy Clin. Immunol.* 2004, **4**, 341–348.

67 International Commisson on Radiological Protection (ICRP), ICRP Publication 66, *Ann ICRP* 1994, 24, 1–3.

68 D. C. Chalupa, P. E. Morrow, G. Oberdörster, M. J. Utell, M. W. Frampton, *Environ. Health Perspect.* 2004, **112**, 879–882.

69 W. L. Beeson, D. E. Abbey, S. F. Knutsen, *Environ. Health Perspect.* 1998, **106**, 813–823.

70 P. Nafstad, L. L. Haheim, T. Wisloff, F. Gram, B. Oftedal, I. Holme, I. Hjermann, P. Leren, *Environ. Health Perspect.* 2004, **112**, 610–615.

71 P. Vineis, G. Hoek, M. Krzyzanowski, F. Vigna-Taglianti, F. Veglia, L. Airoldi, H. Autrup, A. Dunning, S. Garte, P. Hainaut, C. Malaveille, G. Matullo, K. Overvad, O. Raaschou-Nielsen, F. Clavel-Chapelon, J. Linseisen, H. Boeing, A. Trichopoulou, D. Palli, M. Peluso, V. Krogh, R. Tumino, S. Panico, H. B. Bueno-De-Mesquita, P. H. Peeters, E. E. Lund, C. A. Gonzalez, C. Martinez, M. Dorronsoro, A. Barricarte, L. Cirera, J. R. Quiros, G. Berglund, B. Forsberg, N. E. Day, T. J. Key, R. Saracci, R. Kaaks, E. Riboli, Air pollution and risk of lung cancer in a prospective study in Europe. *Int. J. Cancer* 2006.

72 E. Garshick, M. B. Schenker, A. Munoz, M. Segal, T. J. Smith, S. R. Woskie, S. K. Hammond, F. E. Speizer, *Am. Rev. Respir. Dis.* 1987, **135**, 1242–1248.

73 E. Garshick, M. B. Schenker, A. Munoz, M. Segal, T. J. Smith, S. R. Woskie, S. K. Hammond, F. E. Speizer, *Am. Rev. Respir. Dis.* 1988, **137**, 820–825.

74 U. Heinrich, H. Muhle, S. Takenaka, H. Ernst, R. Fuhst, U. Mohr, F. Pott, W. Stober, *J. Appl. Toxicol.* 1986, **6**, 383–395.

75 J. L. Mauderly, *Environ. Health Perspect.* 1994, **102**, Suppl 4, 165–171.

76 R. B. Hayes, T. Thomas, D. T. Silverman, P. Vineis, W. J. Blot, T. J. Mason, L. W. Pickle, P. Correa, E. T. Fontham, J. B. Schoenberg, *Am. J. Ind. Med.* 1989, **16**, 685–695.

77 P. Gustavsson, N. Plato, E. B. Lidstrom, C. Hogstedt, *Scand. J. Work Environ. Health* 1990, **16**, 348–354.

78 K. Steenland, D. Silverman, D. Zaebst, *Am. J. Ind. Med.* 1992, **21**, 887–890.

79 A. Emmelin, L. Nystrom, S. Wall, *Epidemiology* 1993, **4**, 237–244.

80 I. Brüske-Hohlfeld, M. Möhner, W. Ahrens, H. Pohlabeln, J. Heinrich, M. Kreuzer, K. H. Jöckel, H. E. Wichmann, *Am. J. Ind. Med* 1999, **36**, 405–414.

81 I. Brüske-Hohlfeld, *Environ. Health Perspect.* 1999, **107**, Suppl 2,253–258.

82 R. Bhatia, P. Lopipero, A. H. Smith, *Epidemiology* 1998, **9**, 84–91.

83 International Agency for Research on Cancer, *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans*, Vol. 46, IARC, Lyon, 1989.

84 P. Vineis, K. Husgafvel-Pursiainen, *Carcinogenesis* 2005, **26**, 1846–1855.

85 A. Campbell, M. A. Smith, L. M. Sayre, S. C. Bondy, G. Perry, *Brain Res. Bull.* 2001, **55**, 125–132.

86 L. Calderon-Garciduenas, B. Azzarelli, H. Acuna, R. Garcia, T. M. Gambling, N. Osnaya, S. Monroy, M. R. DELTizapantzi, J. L. Carson, A. Villarreal-Calderon, B. Rewcastle, *Toxicol. Pathol.* 2002, **30**, 373–389.

87 C. Möhlmann, *Gefahrstoffe-Reinhaltung Luft* 2005, **65**, 469–471.

88 H. Tjalve and J. Henriksson, *Neurotoxicology* 1999, **20**, 181–195.

89 L. Lu, L. L. Zhang, G. J. Li, W. Guo, W. Liang, W. Zheng, *Neurotoxicology* 2005, **26**, 257–265.

90 K. M. Semchuk, E. J. Love, R. G. Lee, *Can. J. Neurol. Sci.* 1991, **18**, 279–286.

91 J. M. Gorell, C. C. Johnson, B. A. Rybicki, E. L. Peterson, G. X. Kortsha, G. G. Brown, R. J. Richardson, *Neurotoxicology* 1999, **20**, 239–247.

92 B. A. Racette, L. McGee-Minnich, S. M. Moerlein, J. W. Mink, T. O. Videen, J. S. Perlmutter, *Neurology* 2001, **56**, 8–13.

93 M. M. Finkelstein, D. K. Verma, D. Sahai, E. Stefov, *Am. J. Ind. Med.* 2004, **46**, 16–22.

94 A. H. Smith, P. A. Lopipero, V. R. Barroga, *Epidemiology* 1995, **6**, 617–624.

95 International Agency for Research on Cancer, *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, Vol. 68, Silica, Some Silicates, Coal Dust and para-Aramid Fibrils*, IARC, Lyon, 1997.

96 International Agency for Research on Cancer, *IARC Monographs on the Evaluation of the Carcinogenic Risk of Chemicals to Humans, Vol 14, Asbestos*, IARC, Lyon, 1977.

97 P. Cherukuri, S. M. Bachilo, S. H. Litovsky, R. B. Weisman, *J. Am. Chem. Soc.* 2004, **126**, 15638–15639.

98 C. W. Lam, J. T. James, R. McCluskey, S. Arepalli, R. L. Hunter, *Crit. Rev. Toxicol.* 2006, **36**, 189–217.

99 D. B. Warheit, B. R. Laurence, K. L. Reed, D. H. Roach, G. A. Reynolds, T. R. Webb, *Toxicol. Sci.* 2004, **77**, 117–125.

100 K. L. Dreher, *Toxicol. Sci.* 2004, **77**, 3–5.

101 X. Zhao, A. Striolo, P. T. Cummings, *Biophys. J.* 2005, **89**, 3856–3862.

102 S. Homma, A. Miyamoto, S. Sakamoto, K. Kishi, N. Motoi, K. Yoshimura, *Eur. Respir. J.* 2005, **25**, 200–204.

103 T. Homma, T. Ueno, K. Sekizawa, A. Tanaka, M. Hirata, *J. Occup. Health* 2003, **45**, 137–139.

104 T. Chonan, O. Taguchi, *Nihon Kokyuki Gakkai Zasshi* 2004, **42**, 185.

105 D. S. Kim, H. R. Collard, T. E. KingJr., *Proc. Am. Thorac. Soc.* 2006, **3**, 285–292.

106 V. S. Taskar, D. B. Coultas, *Proc. Am. Thorac. Soc.* 2006, **3**, 293–298.

107 V. J. Thannickal, J. C. Horowitz, *Proc. Am Thorac. Soc.* 2006, **3**, 350–356.

108 J. Ameille, G. Pauli, A. Calastreng-Crinquand, D. Vervloet, Y. Iwatsubo, E. Popin, M. C. Bayeux-Dunglas, M. C. Kopferschmitt-Kubler, *Occup. Environ. Med.* 2003, **60**, 136–141.

109 F. Di Stefano, S. Siriruttanapruk, J. McCoach, M. Di Gioacchino, P. S. Burge, *Allerg. Immunol. (Paris)* 2004, **36**, 56–62.

110 K. Donaldson, V. Stone, *Ann. Ist. Super. Sanita* 2003, **39**, 405–410.

111 V. Stone, D. M. Brown, N. Watt, M. Wilson, K. Donaldson, *Inhal. Toxicol.* 2000, **12**, 345–351.

112 A. Andersen, E. Bjelke, F. Langmark, *Br. J. Cancer* 1989, **60**, 112–115.

113 M. Geiser, B. Rothen-Rutishauser, N. Kapp, S. Schurch, W. Kreyling, H. Schulz, M. Semmler, H. Im, V, J. Heyder, P. Gehr, *Environ. Health Perspect.* 2005, **113**, 1555–1560.

114 C. Möhlmann, *Gefahrstoffe-Reinhaltung Luft* 2005, **65**, 469–471.

115 R. D. Brook, B. Franklin, W. Cascio, Y. Hong, G. Howard, M. Lipsett, R. Luepker, M. Mittleman, J. Samet, S. C. SmithJr., I. Tager, *Circulation* 2004, **109**, 2655–2671.

116 W. Luther, *Technological Analysis Industrial Application of Nanomaterials – Chances and Risks*, http://www.nano.uts.edu.au/nanohouse/nanomaterials%20risks.pdf 2004.

10
Impact of Nanotechnological Developments on the Environment

Harald F. Krug and Petra Klug

10.1
Problem

Since Feynman's legendary statement, 'There's plenty of room at the bottom' [1], natural scientists in physics, chemistry, electronics and other fields have been occupied with newly combining the smallest units of matter. Using the resources of modern analysis, especially atomic force microscopy, not only can the properties of matter and atoms be examined, but even single atoms can be manipulated. Aside from the known results, this also gave rise to speculation that manipulation of matter at the single atom level was interpreted in such a way that the possibility exists of generating engines and machines that are able to replicate themselves and therefore possibly get out of control. Without wanting to evoke once again the discussion that has been going on for a long time between Eric Drexler, representative of the hazard hypothesism and Richard Smalley, representative of the safety hypothesis [2], the reactions to this nevertheless show that here (i) sensible communication is necessary in order to point out the real hazards and (ii) accompanying safety-relevant research is needed in order to identify the real hazards and to face them. What we have to count on in all probability in the near future is increased production of nanomaterials and nanoparticles and associated with that its possible release into the air, water and soil.

Bayer, as an example, has produced carbon nanotubes for over 2 years and the production capacity will increase over the next 5 years from

- 2005 3 tons

to

- 2006 30 tons
- 2007 60 tons
- 2009 200 tons
- 2012/13 3000 tons.

Within this book, several examples have been shown where nanomaterials can be used within the environment. The opportunities for applications are nearly endless;

Nanotechnology. Volume 2: Environmental Aspects. Edited by Harald Krug

ISBN: 978-3-527-31735-6

nevertheless, along with their use, numerous threats may arise when they are released into the environment and become distributed everywhere. Nanoparticles and nanomaterials are used in various applications from which they reach the water and the soil. Titanium dioxide and zinc oxide from sunscreens and surface coatings from textiles, glass or other surfaces may be washed off and contaminate natural water [3]. We are responsible for these products, their use and their disposal, hence we must be careful in distributing all these new materials before we know their exact behavior and fate within the environment. It is obvious that nanomaterials will reach the environment and exposure is therefore probable. If there is a biological effect within the organisms that are exposed to these products, then we can postulate a possible risk that has to be addressed:

$$\text{risk} = f(\text{exposure}, \text{hazard})$$

10.2 Risk Management

The first step in risk management is the identification of potential risks and their cause. A reasonable risk identification must include all areas of a technology, both the internal and the external factors (Figure 10.1). In order to achieve this, intensive research is necessary concerning both health-relevant and environment-relevant questions [4–6]:

- particle absorption by living organisms
- accumulation of nanoparticles in certain organs (e.g. lung, liver, spleen, brain, fetus)
- specific effects of nanoparticles in the respiratory tract (e.g. inflammation)
- fate and behavior of substances in the environment (e.g. mobilization of heavy metals, binding to and of toxic substances)

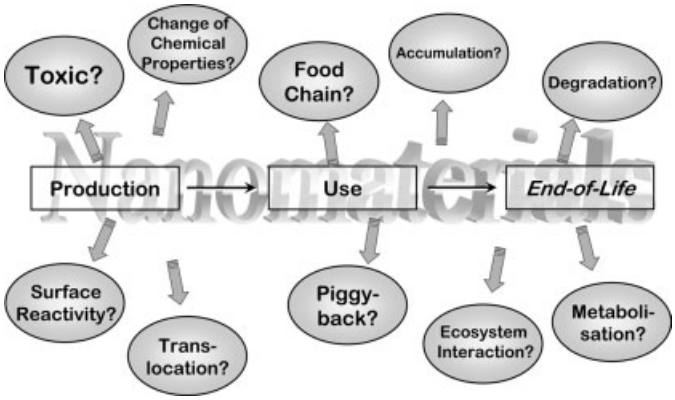


Figure 10.1 Questions of hazard identification of nanoparticles that can occur in the environment during the entire life cycle (Adapted from [5]).

- possible accumulation via the food chain
- desorption/adsorption
- unexpected effects.

A very important aspect in the judgment of possible risks from nanotechnological products lies in the differentiation of free and fixed nanoparticles, because there is obviously a large difference in their mobility. Furthermore, one must differentiate between particles and materials that are being manufactured as technical products and those that are created accidentally in technical processes and are released into the environment (e.g. diesel exhaust, fly ash, catalytic dust, candle soot). Humans are and have been exposed to such ultrafine particles (UFPs), which mainly result from combustion processes, since the beginning of their biological development. Whereas in former times forest fires, volcanoes and sand- and other storms occurred, since the industrial revolution and since the increase in motor traffic, a dramatic rise in UFPs in the air has taken place over the last century.

With the presumed and very fast development in the area of nanotechnological applications, it must now be taken into account that a further source of such minute particles will emerge that will reach humans via the environment: either from the environment to the people or vice versa.

The exposure that goes along with that to humans via the respiratory system, nutrition and skin and also the direct injection of nanoparticles in the medical sector could lead to adverse effects [7–10]. With the knowledge that newly synthesized nanomaterials possess completely new properties in view of chemical, physical and electronic applications, entirely new effects on living organisms can be postulated. For these reasons, the behavior of nanoparticles in the environment and in living organisms cannot simply be extrapolated; a significant prediction of the toxicity of nanoparticles on the basis of the knowledge concerning conventional materials cannot be made. The situation is, in addition to the above-mentioned new effects, not assessable, also for the following reasons:

- the large number of different materials
- the large number of different structures, surfaces, shapes and sizes.

Hence information about the safety and the possible hazards from nanomaterials is urgently needed. Toxicologists can, by all means, benefit from the previously performed studies on the effects of ultrafine particles on the environment, since this is where a multitude of findings already exist. Since the Middle Ages and earlier there have been well-documented cases on workplace-related exposure with effects on health. Especially workers in mines are subject to exposure to inhaled dust of any size for long periods of their lives, which can lead to pneumoconiosis and fibrosis of the lung. It has been shown that especially the fractions of ultrafine particles in the air lead to the greatest effects on health [11–17].

Present scientific knowledge about substances and devices produced using nanotechnologies precludes going further than identifying hazards – the first step in risk assessment – and providing some elements of hazard characterization – the

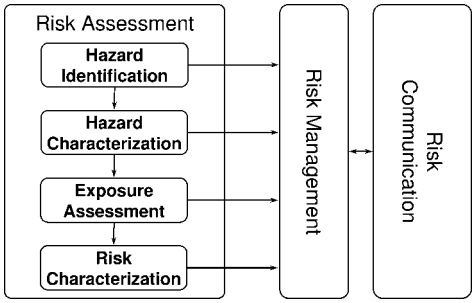


Figure 10.2 From risk assessment to risk communication.

second step in risk assessment. Research on the behavior of nanoparticles in different compartments of the environment and also research on the impacts of nanoparticles on animals and humans, depending on different ways of assimilation of nanoparticles (via the lung, via digestion or via the skin) is on the agenda. Together with assessments of the exposure of humans to nanoparticles at different locations (exposure at workplaces in industry or for consumers of nanoparticle-based products) – which would require prospective analyses of the production and distribution of nanoparticle-based products – the risk management chain (Figure 10.2) could be completed by scientific knowledge. Moreover, if there are some situations where risks could be assumed, we have to communicate these to politicians, the public and consumers. Affected people must have the chance to decide whether they want to accept such risks or not.

10.3 Sources of Nanoparticles: New Products

The results of these and many other studies could lead us to expect a cautionary and understandably biased view towards new sources of nanoparticles. Various products have already been on the market for a long time (Table 10.1).

Some of these products will sooner or later lead to a rise in the amount of particles in the environment, even if not all released particles will be in the ultrafine range under 100 nm. The use of nanotechnological products in all areas of life raises some important questions:

- Is the use immediately followed by an increased exposure towards nanoparticles and which exposure routes are involved?
- Where do the nanoparticles go after being released and where do they stay?
- Do the new materials pose an unacceptable risk?

The manufacture of nanomaterials and nanoparticles at present does not make a noteworthy contribution to the amount of particles in the environment. The public discussion and the voiced opinions in the daily media are currently based purely on speculation. On the other hand, it should also be made clear that the corresponding

Table 10.1 Use of nanoscale metal oxide and carbon modifications in various products of commercial interest (examples).

Type	Products (examples)
Metal oxides	
Silicon dioxide (SiO_2)	Additives in polymer composites
Titanium dioxide (TiO_2)	UV-A protection/photocatalysis
Aluminum oxide (Al_2O_3)	Solar cells
Iron oxide (Fe_3O_4, Fe_2O_3)	Pharmacology/medicine/catalysis
Zirconium oxide (ZrO_2)	Additives to scratch-resistant surfaces
Zinc oxide (ZnO)	UV-A protection/photonic
Carbon modifications	
Carbon black	Car tires, printers, copy machines
Fullerenes	
Buckminsterfullerenes (C_{60})	Mechanical and tri-biological applications/additives in lubrications/greases Additives to cosmetics
Carbon nanotubes	
Single-wall carbon nanotubes	Additives in polymer composites
Multi-wall carbon nanotubes	Electronic field emission Batteries Fuel cells
Carbon nanowires	
Various conformations	Mechanical and tri-biological applications Carrier material for catalysts Additives in polymer composites Elastic spumes

data for assessing an exposure and the hazards are still lacking and that the gaps in knowledge must definitely be closed. For this, a new research field has been established: nanotoxicology [18]. This specializes in analyzing the biological safety of technical nanostructures and nanomaterials. The same rules apply for this part of toxicological research as for the hazard assessment of other chemical impacts on the environment (Figure 10.3).

Within the assessment of relevant environmental problems of xenobiotics or chemicals, the predicted environmental concentration (PEC) and the predicted 'no-effect' concentration (PNEC) are of great importance. If the PEC value is low and the PNEC value is high (PEC/PNEC < 1), a low risk can be calculated and no measures for reduction are necessary, whereas in the opposite case, a high PEC value and a low PNEC value (PEC/PNEC > 1), further measures are indicated. On the one hand this can mean that a limited use will be the result, but it can be corrected by extended test procedures if it results in the relation between the two values being <1 again.

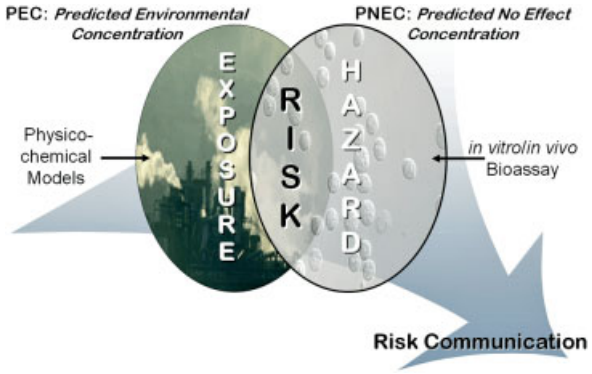


Figure 10.3 Conventional approach to ecotoxicological risk assessment.

10.4 Production and Use of Nanomaterials

As mentioned above, the production and use of nanomaterials will increase dramatically and all areas of daily life will be affected. Exposure to nanomaterials will become more probable, whereas an important differentiation will be whether these materials are coated or uncoated. Coatings change the properties of the material, which has a direct impact on the particle and its behavior in biological systems. Also of importance in view of possible far-reaching effects is the stability and along with that the survival time of the particles in the environment. Both the environmental-relevant and the health-relevant effects are largely dependent on whether the nanoparticles are persistent or not. As far as stable particles are concerned, basically the same is true as for long-lived chemicals: they could, if absorbed by living organisms, accumulate, concentrate in certain target organs and eventually develop a critical effect if the dosage reached is sufficiently high. Therefore, precautionary safety measures must be taken in good time, in order to recognize those stable materials with critical consequences and to prevent their release. It has been demanded by several researchers that the following points must be strictly observed:

- potential routes for human exposure
- possible industrial sources of occupational exposure
- level of exposure
- means of and effectiveness of control measures
- potential numbers exposed
- trends in the use of nanotechnological products
- timely recognition of effects that are caused by the change from the laboratory scale to the industrial scale.

These issues can be transferred in this or a similar fashion to the conditions in the environment; for example, nanoparticles in personal care products reach wastewater

and therefore also the environment, in small but continuous quantities, through body cleansing [3].

10.5 Workplace and the Environment: Effects and Aspects of Nanomaterials

The technical possibilities of nanotechnology to develop new types of miniature sensors, pollutant filters and fuel-cell catalysts can make a considerable contribution to the improvement of the environment. However, these developments are only at the beginning stage and the insecurity that is connected with the production of these new materials currently outweighs their possible advantages. To convey these advantages and to measure them against their possible disadvantages are a huge challenge for the developers of this technology. However, there are already a number of ideas for better 'end-of-pipe' technologies that could make a contribution to air pollution control, wastewater, soil and waste purification and energy production and storage. Some examples of uses in the environment that have been described in detail in the other chapters of this book are:

- Syntheses or fabrication processes can take place at room temperature and under normal pressure in order to save energy.
- The use of non-toxic catalysts leads to the minimal formation of pollutants and reduces material usage and emission.
- Water-based reactions can help save on solvents and reduce contaminants and an adapted 'just-in-time' production can reduce ecological environmental pollution and over-production.
- Nanoscale information technology will improve the tracking of products and product routes in order to control recycling, further use and 'end-of-life' disposal in an environmentally safe way.
- Nanoscale iron could be applied very efficiently for groundwater treatment along with further possibilities for improvement by using additional metals, such as palladium.
- Various nanomaterials can be used as semiconductor films for producing sensors or photocatalysts and these sensors could detect organic pollutants and degrade them by photocatalytic reactions.
- Single-molecule detection could help recognize pollutants early so that precautionary measures are possible.
- Nano-building blocks could analyze chemicals by specific reactions; biogenous toxins could be detected and food could be monitored more efficiently.

Already lead-off studies are being published where carbon nanotubes are used as sensors for gases [19], metal oxides show a sensitive reaction to moisture or hydrogen [20], silver polymer nanoparticles are used for the detection of aromatic hydrocarbons [21] and nanoparticles are used as sorbents of environmental contaminants [22].

Aside from these positive aspects of nanotechnological uses in the environment, we must not lose sight of the above-mentioned possible negative effects on the environment and on living organisms. In the meantime, recent studies have already been conducted which directly examine an exposure to nanoparticles and their effects on mortality. However, so far no significant increases in the standardized mortality rate for specific causes of death in rats that were subjected to high concentrations of nanoscale TiO_2 were found, or in manufacturing workers in a cohort study [23]. The existing data are anything but sufficient in order to serve as a basis for a scientific discussion. Even if the pure nanomaterials did not have a negative effect on living systems, according to the opinions of toxicologists and occupational physicians, they could still be able to bind other contaminants on their surface, enabling more facile transport into the air or water and subsequently leading to an increased burden of the organisms in the environment. Most nanomaterials and nanoproducts are currently only used in laboratories, so that environmental pollution so far low or negligible. However, with the increasing number of possible applications, the possible impact on the environment and the exposure of living organisms will increase [4, 24].

10.6 Distribution of Nanoparticles in Ambient Air

Depending on their production and use, nanoparticles can reach the water or air, from where they will eventually reach the groundwater or the soil. Furthermore, their use in disposable articles necessitates increased caution where recycling and waste management are concerned. Since nanoparticles in the air act more like gas molecules and hardly sediment at all, they could cover great distances. Currently, the long-term effects cannot be calculated by any means based on the poor data available. However, the air is one of the best-examined environmental compartments and its pollution has been recorded for many decades. The organisms that are exposed in ambient air are being examined intensively and many studies are examining the adverse effects of particular mass in the air (see Chapter 9 for detailed information). Nevertheless, the unreasonably hazardous effects especially of ultrafine particles in the air have been discovered during the past 10 years [16, 17, 25–27]. It could be assumed that the number of ultrafine particles in the air will increase in the coming years, first at the workplace and then in the environment. Especially due to the surface treatment of engineered nanoparticles, release of nanoscale particles could occur, while oxidic metal particles and carbon nanomaterials normally start to aggregate soon after their synthesis and form agglomerates, which frequently sediment faster than the primary particles. Immediately after their formation, the primary particles act more like a gas or a vapor phase where diffusion processes prevail (for details, see Chapter 7). The primary particles show high diffusion coefficients and blend in well in aerosol systems. This is an important aspect in the control of such minute particles in the air. In closed synthesis systems that show a leak, it is easier for nanoparticles to escape unnoticed, due to their much higher

Table 10.2 Coagulation half-life period of nano particles.

Particle diameter (nm)	Half-life at a concentration of			
	1 g m^{-3}	1 mg m^{-3}	1 μg m^{-3}	1 ng m^{-3}
1	2.20 μs	2.20 ms	2.20 s	36.67 min
2	12.00 μs	12.00 ms	12.00 s	3.34 h
5	0.12 ms	0.12 s	2.00 min	33.34 h
10	0.70 ms	0.70 s	11.67 min	8.10 d
20	3.80 ms	3.80 s	63.34 min	43.98 d

(Original values from [47]).

mobility, than it is for their larger counterparts. Due to the faster distribution, the measurable concentration directly at the leakage is rather low; in return, a larger number of individuals/workers could be exposed since the particles are distributed more extensively over a larger area. For this reason, gas detection systems must be used and not particle detectors, ultra-sensitive systems in any case. On the other hand, the higher speed of the smaller particles also leads to a larger number of collisions, which support the tendency for aggregation and agglomeration and facilitate particle growth. This growth process is directly dependent on the number of particles in the given volume and their mobility. Although this process runs very fast (Table 10.2), it must be considered that a newly formed nanoparticle resulting from these collisions, which is obviously larger than the primary particles, will nevertheless have at least one of its dimensions smaller than 100 nm.

One difficulty is, and also will be in future studies, that airborne particles do not display a uniform population, but rather result from very different sources. Hence, possibly emitted engineered particles can mix already at the workplace with, for example, diesel exhaust fumes during packing and loading of nanomaterials [28], so that assessment and distinct source definition are often difficult. Nevertheless, there are indications that during the processing of nanomaterials, the same ones are released into the air [29]. However, these are so far the only current studies that have dealt with the release of produced nanomaterials. Two further reports deal with the general aspects of the distribution of ultrafine particles in the atmosphere and at the workplace, but largely particles that are created unintentionally and not those produced technically [30, 31].

10.7 Distribution of Nanoparticles in Water

In all probability, and viewed over a longer period of time, distribution and exposure to the environment and to humans in the water and the soil will take place. If the growth in production in fact moves as fast as is predicted, increased concentrations of nanomaterials in the groundwater or in the soil could be an essential exposure route

that should be taken into consideration when assessing the risk for the environment [32]. Such products are already in use as titanium dioxide nanoparticles in sunscreens and paints, as carbon nanotubes in composite materials (car tires) or as aluminum particles in shampoos [3].

As mentioned above already, different metallic and polymeric nanoparticles can be used for groundwater remediation (for details, see Chapter 4) and are therefore a possible source of pollution [33–35].

In this context, two issues are of particular importance: on the one hand, the direct effects of nanomaterials in the environment must be examined, and on the other, the routes of the nanomaterials into the environment (and through it, e.g. through the porous soil matrix) must be clarified. Some studies have shown a large difference between the mobility of nanoparticles in porous media, depending on whether they are oxidic particles or carbon particles such as fullerenes [34, 36]. During their transport in the water or through the soil matrix, even harmless particles could react with other chemicals or adsorb them and therefore contribute to a possible hazard themselves [37–39]. Many substances have the capability of sticking to the surface of nanoparticles [21, 40]; these adsorbed chemicals could then cause a biological effect, the nanoparticles itself especially increase the uptake by organisms or directly lead to adverse effects (Figure 10.4). It is conceivable that the particles themselves or the bound pollutants after they have been taken up by living organisms cause, for instance, lysosomal damage, thereby increasing autophagy and in effect cell damage, as suggested by Moore [41].

Using the fullerenes as an example, it could be shown that the availability and the mobility of nanomaterials in the water and in the soil depend very strongly on the physico-chemical properties of the surface of these particles. At different ionic strengths and pH values, three fullerene preparations and four different oxidic materials behaved very differently with regard to their transport properties in the aquifer [34, 36]. Whereas the transport of mineral nanoparticles could be described by established models for the transport of particles in porous media, the same did not hold true for the fullerenes since they behaved in a totally unexpected way and showed unusual properties. Especially that particular form of fullerenes that has recently caused concern in biological tests [10] is the least mobile one in soil and water tests, so

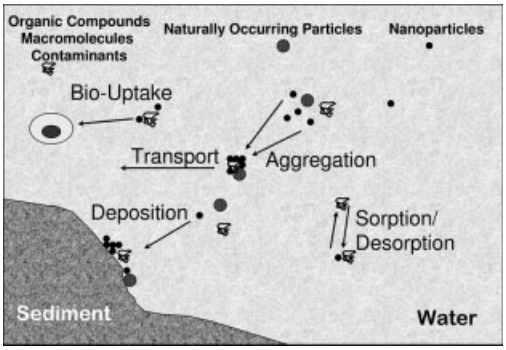


Figure 10.4 Distribution and possible reactions of nanoparticles in the aquatic system.

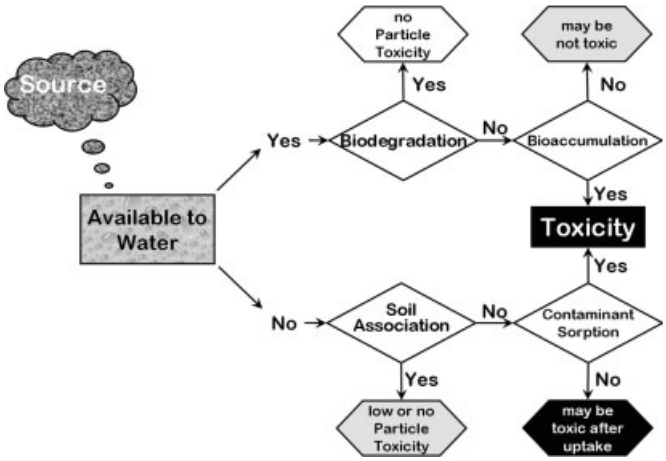


Figure 10.5 Decision scheme for an assessment of nanomaterials in environmental compartments.

that we can rather expect a reduced risk here. This example is intended to demonstrate that our knowledge about the behavior of nanomaterials in the environment is still totally insufficient and that the resulting debate about possible risks has not yet been professionally established. Hence a decision should be made for every single material, based on the knowledge about the availability and transport in water and/or soil (Figure 10.5).

The above-mentioned work by Oberdörster [10] showed, for the first time, a direct effect of fullerenes on fish. After exposure to 0.5 ppm of water-soluble buckyballs (C_{60}) for 48 h, the fish showed significantly increased levels of lipid peroxidation in the brain. Furthermore, the gene expression was checked, for example in the liver, where it was striking that there were genes that were turned on and others that were turned off. Also, there was evidence of a systemic effect of the fullerenes, despite the fact that the author herself judged the concentration to be very high and allocated the fullerenes only 'moderate toxicity'. Unanswered are the questions of how much of the material was really taken up by the fish and whether this material can really accumulate in the food chain or in specific organs, such as the brain. Such an accumulation would then suggest a hazard for other organisms. Therefore, it is necessary to increase the efforts in order to acquire more information about the bioaccumulation of nanomaterials, since many of them are not likely to be biodegradable and could therefore, under certain conditions, behave as persistent pollutants. Some follow-up studies by different authors resulted in controversial data as the high capacity for toxic effects of fullerenes could not documented by the same group when fullerenes were resuspended directly in water [42]. A solution of fullerenes in tetrahydrofuran (THF) has been shown to be highly toxic ([10]; [43]) and we could demonstrate that this is dependent on the peroxides formed in THF spontaneously. Hence it is not always the nanomaterials that are the toxic component band the solvents or contaminants often have a more dramatic effect on living organisms, as has been published elsewhere [44, 45].

10.8 Conclusions

On the basis of the displayed level of knowledge so far, five fundamental considerations can be taken into account about the eco-toxicological risk management of nanomaterials:

1. Research on the toxicology of nanomaterials has so far been more of a description of the symptoms, hence it is essential to find out more about the biological mechanisms on the cellular and molecular level.
2. The further development of models and model systems is necessary in order to understand cellular and physiological processes better and to be able to include the communication between the cells within the investigations.
3. It should be possible to establish a relationship between molecular, cellular and pathophysiological end-points with ecological consequences.
4. A more precise and preventive assessment of hazardous effects and risks of new developments in the sector of nanotechnology should be possible by strong improvement of the available data.
5. Of essential relevance is an increasing knowledge of the life-cycle of nanotechnological products, from production, through their use and until their disposal. This includes possible exposure scenarios in addition to the systemic effects in living organisms, to close knowledge gaps, where screening studies may be preferred instead of detailed analysis.

Only on the basis of improved knowledge about the potential dangers in the entire life cycle of the products will a risk assessment be possible and the corresponding measures can then be implemented in order to reduce a possible hazard.

This was also made clear at a workshop of the National Science Foundation (NSF) and the Environmental Protection Agency (EPA) in the USA, whose results were summarized by Dreher [46] as follows:

- valid exposure assessment for engineered nanoparticles.
- toxicity of nanoparticles
- extrapolation of their toxicity using existing data about particles and ambient fibers
- environmental and biological fate, transport, persistence and the possible transformation of nanoparticles
- recyclability and sustainability of nanotechnological products.

Hence the question remains unanswered of whether all nanomaterials are also simultaneously nanonoxes. The entire subject matter of environment and health is displayed very clearly in a recent review that deals with the possible routes of nanoparticles in the environment and the exposure routes by organisms [18]. At this point it is important to state that the currently available data are not sufficient for a realistic assessment of exposure, hazards and the associated risks. Moreover, nobody takes into account that there is a natural background exposure with several particle types

Table 10.3 Most frequent elements in the Earth's crust (italics: typically as oxides comparable to engineered nanoparticles).

Element	Concentration (%)
Oxygen	47
Silicon	*28*
Aluminum	*8*
Iron	*4.5*
Calcium	3.5
Potassium	2.5
Sodium	2.5
Magnesium	2
Titanium	*0.5*
Hydrogen	0.2
Carbon	0.2
Others	<1

used in technical applications as well. Some elements produced as oxides at the nanoscale in very large amounts are equally distributed all over the Earth. Therefore, we cannot discriminate between synthetic materials and naturally occurring materials when the engineered nanoparticles have reached the environment (Table 10.3).

Therefore, specific regulatory measures are unnecessary at this point since it is completely unclear what they should actually be aimed at. Nevertheless, contact with nanomaterials should be given increased attention during development in the research laboratory and also during large-scale technical production since the unique properties of the new materials are able to show not only technical but also biological effects, and that the behavior of the nanoparticles towards bulk materials will certainly be changed. In the meantime, this has also been recognized by the funding organizations and both European and German governmental funding have initiated relevant projects. In these projects it is also taken into account that the field of nanotoxicology can only be approached in a multidisciplinary way, that is, in addition to industry and the agencies, also chemists, physicists, materials scientists, engineers, medical professionals, biologists, toxicologists, ecologists, statisticians and additional branches of study, who have to deal with all aspects of nanotechnology, including the ethical questions and the sustainability of this technology, are also challenged.

References

1 Feynman, R.P. (1960) There's plenty of room at the bottom. *Engineering and Science*, **23**, 22–36.

2 Baum, R. (2003) Nanotechnology – Drexler and Smalley make the case for and against 'molecular assemblers'. *Chemical and Engineering News*, **81**, 37–42.

3 Daughton, C.G. and Ternes, T.A. (1999) Pharmaceuticals and personal care products in the environment: agents of

subtle change? *Environmental Health Perspectives*, **107** (Suppl 6), 907–938.

4 Aitken, R.J., Creely, K.S. and Tran, C.L. (2004) Nanoparticles: an Occupational Hygiene Review. Norwich, Crown Copyright.

5 Helland, A. (2004) *Nanoparticles: a Closer Look at the Risks to Human Health and the Environment*. IIIEE, Lund University, Lund.

6 Royal Society (2004) *Nanoscience and Nanotechnologies: Opportunities and Uncertainties*. Royal Society, London.

7 Krug, H.F. (2003) Nanopartikel: Gesundheitsrisiko, Therapiechance? *Nachrichten aus der Chemie*, **51**, 1241–1246.

8 Krug, H.F. and Diabaté, S. (2003) Ultrafeine Partikel: Gesundheitsrisiko versus Therapiechance!? *Umwelt Medizin Gesellschaft*, **16**, 250–255.

9 Krug, H.F., Kern, K. and Diabaté, S. (2004) Toxikologische Aspekte der Nanotechnologie. Versuch einer Abwägung. *Technikfolgenabschätzung: Theorie und Praxis*, **13**, 58–64.

10 Oberdörster, E. (2004) Manufactured nanomaterials (fullerenes, C60) induce oxidative stress in the brain of juvenile largemouth bass. *Environmental Health Perspectives*, **112**, 1058–1062.

11 Oberdörster, G. (2000) Toxicology of ultrafine particles: *in vivo* studies. *Philosophical Transactions of the Royal Society of London, Series A*, **358**, 2719–2739.

12 Eikmann, T. and Seitz, H. (2002) Klein, aber oho! Von der zunehmenden Bedeutung der Feinstäube. *Umweltmed Forsch Prax*, **7**, 63–64.

13 Heinrich, J., Grote, V., Peters, A. and Wichmann, H.E. (2002) Gesundheitliche Wirkungen von Feinstaub: Epidemiologie der Langzeiteffekte. *Umweltmed Forsch Prax*, **7**, 91–99.

14 Pekkanen, J., Peters, A., Hoek, G., Tiittanen, P., Brunekreef, B., de Hartog, J., Heinrich, J., Ibald-Mulli, A., Kreyling, W.G., Lanki, T., Timonen, K.L. and Vanninen, E. (2002) Particulate air pollution and risk of ST-segment depression during repeated submaximal exercise tests among subjects with coronary heart disease: the Exposure and Risk Assessment for Fine and Ultrafine Particles in Ambient Air (ULTRA) study. *Circulation*, **106**, 933–938.

15 Peters, A., Heinrich, J. and Wichmann, H.E. (2002) Gesundheitliche Wirkungen von Feinstaub - Epidemiologie der Kurzzeiteffekte. *Umweltmed Forsch Prax*, **7**, 101–115.

16 de Hartog, J.J., Hoek, G., Peters, A., Timonen, K.L., Ibald-Mulli, A., Brunekreef, B., Heinrich, J., Tiittanen, P., van Wijnen, J.H., Kreyling, W.G., Kulmala, M. and Pekkanen, J. (2003) Effects of fine and ultrafine particles on cardiorespiratory symptoms in elderly subjects with coronary heart disease: the ULTRA study. *American Journal of Epidemiology*, **157**, 613–623.

17 Kappos, A.D., Bruckmann, P., Eikmann, T., Englert, N., Heinrich, U., Hoppe, P., Koch, E., Krause, G.H., Kreyling, W.G., Rauchfuss, K., Rombout, P., Schulz-Klemp, V., Thiel, W.R. and Wichmann, H.E. (2004) Health effects of particles in ambient air. *International Journal of Hygiene and Environmental Health*, **207**, 399–407.

18 Oberdörster, G., Oberdörster, E. and Oberdörster, J. (2005) Nanotoxicology: an emerging discipline evolving from studies of ultrafine particles. *Environmental Health Perspectives*, **113**, 823–839.

19 Modi, A., Koratkar, N., Lass, E., Wei, B. and Ajayan, P.M. (2003) Miniaturized gas ionization sensors using carbon nanotubes. *Nature*, **424**, 171–174.

20 Varghese, O.K. and Grimes, C.A. (2003) Metal oxide nanoarchitectures for environmental sensing. *Journal of Nanoscience and Nanotechnology*, **3**, 277–293.

21 De Jesus, M.A., Giesfeldt, K.S. and Sepaniak, M.J. (2004) Factors affecting the sorption of model environmental pollutants onto silver polydimethylsiloxane nanocomposite

Raman substrates. *Applied Spectroscopy*, **58**, 1157–1164.

22 Yuan, G. (2004) Natural and modified nanomaterials as sorbents of environmental contaminants. *Journal of Environmental Science and Health, Part A*, **39**, 2661–2670.

23 Fryzek, J.P., Chadda, B., Marano, D., White, K., Schweitzer, S., McLaughlin, J.K. and Blot, WJ. (2003) A cohort mortality study among titanium dioxide manufacturing workers in the United States. *Journal of Occupational and Environmental Medicine*, **45**, 400–409.

24 Luther, W. (2004) *Industrial Application of Nanomaterials – Chances and Risks. Technology Analysis*, Future Technologies Division of VDI, Technologiezentrum, Düsseldorf.

25 Wichmann, H.E., Spix, C., Tuch, T., Wolke, G., Peters, A., Heinrich, J., Kreyling, W.G. and Heyder, J. (2000) Daily mortality and fine and ultrafine particles in Erfurt, Germany. Part I: role of particle number and particle mass. *Research Report/Health Effects Institute*, **98**, 5–86.

26 Ibald-Mulli, A., Wichmann, H.E., Kreyling, W.G. and Peters, A. (2002) Epidemiological evidence on health effects of ultrafine particles. *Journal of Aerosol Medicine*, **15**, 189–201.

27 Schulz, H., Harder, V., Ibald-Mulli, A., Khandoga, A., Koenig, W., Krombach, F., Radykewicz, R., Stampfl, A., Thorand, B. and Peters, A. (2005) Cardiovascular effects of fine and ultrafine particles. *Journal of Aerosol Medicine*, **18**, 1–22.

28 Kuhlbusch, T.A., Neumann, S. and Fissan, H. (2004) Number size distribution, mass concentration and particle composition of PM1, PM2.5 and PM10 in bag filling areas of carbon black production. *Journal of Occupational and Environmental Hygiene*, **1**, 660–671.

29 Maynard, A.D., Baron, P.A., Foley, M., Shvedova, A.A., Kisin, E.R. and Castranova, V. (2004) Exposure to carbon nanotube material: aerosol release during the handling of unrefined single-walled carbon nanotube material. *Journal of Toxicology and Environmental Health, Part A*, **67**, 87–107.

30 Brown, L.M., Collings, N., Harrison, R.M., Maynard, A.D. and Maynard, R.L. (2003) *Ultrafine Particles in the Atmosphere*, Imperial College Press, London.

31 Möhlmann, C. (2003) *Ultrafine Aerosols at Workplaces*. HVBG Berufsgenossenschaftliches Institut für Arbeitsschutz, Sankt Augustin.

32 Colvin, V.L. (2003) The potential environmental impact of engineered nanomaterials. *Nature Biotechnology*, **21**, 1166–1170.

33 Wang, C.B. and Zhang, W.X. (1997) Synthesizing nanoscale iron particles for rapid and complete dechlorination of TCE and PCBs. *Environmental Science and Technology*, **31**, 2154–2156.

34 Lecoanet, H.F. and Wiesner, M.R. (2004) Velocity effects on fullerenes and oxide nanoparticle deposition in porous media. *Environmental Science and Technology*, **38**, 4377–4382.

35 Tungittiplakorn, W., Lion, L.W., Cohen, C. and Kim, J.Y. (2004) Engineered polymeric nanoparticles for soil remediation. *Environmental Science and Technology*, **38**, 1605–1610.

36 Lecoanet, H.F., Bottero, J.Y. and Wiesner, M.R. (2004) Laboratory assessment of the mobility of nanomaterials in porous media. *Environmental Science and Technology*, **38**, 5164–5169.

37 Moore, M.N. and Willows, R.I. (1998) A model for cellular uptake and intracellular behavior of particulate-bound micro-pollutants. *Marine Environmental Research*, **46**, 509–514.

38 Gerde, P., Muggenburg, B.A., Lundborg, M., Tesfaigzi, Y. and Dahl, AR. (2001) Respiratory epithelial penetration and clearance of particle-borne benzo[*a*] pyrene. *Research Report/Health Effects Institute*, 5–25.

39 Xia, T., Korge, P., Weiss, J.N., Li, N., Venkatesen, M.I., Sioutas, C. and Nel, A.

(2004) Quinones and aromatic chemical compounds in particulate matter induce mitochondrial dysfunction: implications for ultrafine particle toxicity. *Environmental Health Perspectives*, **112**, 1347–1358.

40 Mudroch, A., Kaiser, K.L.E., Comba, M.E. and Neilson, M. (1994) Particle-associated PCBs in Lake Ontario. *Science of the Total Environment*, **158**, 113–125.

41 Moore, MN. (2002) Biocomplexity: the post-genome challenge in ecotoxicology. *Aquatic Toxicology*, **59**, 1–15.

42 Zhu, S., Oberdörster, E. and Haasch, M.L. (2006) Toxicity of an engineered nanoparticle (fullerene, C60) in two aquatic species, *Daphnia* and fathead minnow. *Marine Environmental Research*, **62** (Suppl), S5–S9.

43 Lovern, S.B., Strickler, J.R. and Klaper, R. (2007) Behavioral and physiological changes in *Daphnia magna* when exposed to nanoparticle suspensions (titanium dioxide, nano-C_{60} and $C_{60}H_xC_{70}H_x$). *Environmental Science and Technology*, **41**, 4465–4470.

44 Wörle-Knirsch, J.M., Pulskamp, K. and Krug, H.F. (2006) Oops they did it again! Carbon nanotubes hoax scientists in viability assays. *Nano Letters*, **6**, 1261–1268.

45 Pulskamp, K., Diabate, S. and Krug, H.F. (2007) Carbon nanotubes show no sign of acute toxicity but induce intracellular reactive oxygen species in dependence on contaminants. *Toxicology Letters*, **168**, 58–74.

46 Dreher, K.L. (2004) Health and environmental impact of nanotechnology: toxicological assessment of manufactured nanoparticles. *Toxicological Sciences*, **77**, 3–5. (see also National Science Foundation and US Environmental Protection Agency, Nanotechnology Grand Challenge in the Environment Research Planning Workshop, Session E: Nanotechnology Implications in Health and the Environment, May 2003, http://es.epa.gov/ncer/publications/nano/nanotechnology4-20-04.pdf).

47 Preining, O. (1998) The physical nature of very, very small particles and its impact on their behavior. *Journal of Aerosol Science*, **29**, 481–495.

48 Kern, K., Wörle-Knirsch, J.M. and Krug, H.F. (2004) Nanonoxen: nanoparticle uptake, transport and toxicity. *Signal Transduction*, **3–4**, 149.

I
Basic Principles and Theory

Nanotechnology. Volume 3: Information Technology I. Edited by Rainer Waser

ISBN: 978-3-527-31738-7

1
Phase-Coherent Transport

Thomas Schäpers

1.1
Introduction

From elementary quantum mechanics it is known that electrons possess wave properties in addition to their appearance as a particle. Often, these wave properties are difficult to observe directly, the main reason being that in many cases the electron wavelength is quite small – that is, in metals the wavelength of the electrons at the Fermi energy is only of the order of a few nanometers. Therefore, one possible approach to observing the phenomena related to the wave properties of the electrons is to reduce the sample size to dimensions close to the electron wavelength, as performed in a quantum point contact. Nevertheless, the wave nature of the electrons is sometimes revealed under much more relaxed conditions. An essential perquisite here is that the coherent wave propagation is maintained over sufficiently long distances, so that interference effects can occur. In most cases this condition is only fulfilled at low temperatures in the Kelvin range, where inelastic scattering is suppressed to a large extent.

In diffusive conductors, one possible way to achieve electron interference is if the diffusive motion allows electrons to propagate coherently in closed loops. This so-called "weak localization effect" can even be observed in macroscopic structures. The electron interference can be significantly modified if spin precession (i.e., due to spin-orbit coupling) comes into play. Well-controlled electron interference can be achieved if the wave propagation is guided by the shape of the conductor, and an excellent example in this respect is the Aharonov–Bohm effect, which is observed in ring-shaped conductors.

This discussion of phase-coherent transport in nanostructures begins by introducing the relevant length scales and the different transport regimes in Section 1.2. Subsequently, in Section 1.3 the Landauer–Büttiker formalism and ballistic transport through a split-gate point contact are discussed. Section 1.4 provides an explanation for the weak localization effect, which leads to an enhanced resistance, whilst in Section 1.5 it is shown that spin precession can result in the reversal of the weak localization effect. Phase-coherent transport in ring-shaped structures is discussed in Sections 1.4 and 1.5, while in Section 1.6 it is shown that the finite number of

Nanotechnology. Volume 3: Information Technology I. Edited by Rainer Waser

ISBN: 978-3-527-31738-7

scattering centers in very small structures can result in pronounced fluctuations in conductance. Although, within this chapter, transport phenomena in two- and one-dimensional structures are outlined, zero-dimensional structures – namely quantum dots – are discussed in detail in Chapter 2.

1.2 Characteristic Length Scales

Transport in nanoelectronic systems can be classified by relating its size to some specific characteristic length scales [1, 2] which determine how the carriers propagate through the sample. In the following sections, the elastic and inelastic mean free path are introduced, which quantify the degree of elastic and inelastic scattering occurring in the structure, respectively. A length scale, which provides information about loss of the phase memory is termed the phase-coherence length.

1.2.1 Elastic Mean Free Path

The elastic mean free path l_e is a measure of the distance between subsequent elastic scattering events. Such events occur due to the fact that the conductor is not ideal but rather contains irregularities in the lattice, such as impurities or dislocations. The scattering can be considered as *elastic*, which means that the electron energy is conserved. A typical example is the scattering of an electron at a charged impurity. If we assume a stationary scattering center, then effectively no energy is transferred during the scattering event, whereas the direction of the electron momentum can change greatly.

In order to determine the elastic mean free path l_e within the Drude model, one must first calculate the average time between elastic scattering events, τ_e. Its value can be extracted from the electron mobility μ_e, given by

$$\mu_e = \frac{e\tau_e}{m^*} \tag{1.1}$$

The quantities m^* and e are the effective electron mass and the elementary charge, respectively. The electron mobility is a measure of the increase of the drift velocity v_{drift} in a conductor with increasing electric field E: $v_{\text{drift}} = -\mu_e E$. In practice, the electron mobility is determined from the electron concentration n_e and the Drude conductivity σ_0 by

$$\mu_e = \frac{\sigma_0}{e n_e} \tag{1.2}$$

Experimentally, the electron concentration n_e is obtained from Hall measurements, while the conductivity σ_0 is deduced from resistance measurements at zero magnetic field.

Effectively, only electrons at the Fermi energy E_F contribute to the electron transport. Therefore, the elastic mean free path l_e is given by the length an electron

with the Fermi velocity v_F propagates until it is elastically scattered after the elastic scattering time τ_e:

$$l_e = \tau_e v_F \tag{1.3}$$

As an example, for a typical two-dimensional (2-D) electron gas in an AlGaAs/GaAs heterostructure (see Section 1.3), low-temperature mobilities of around $10^6\,\mathrm{cm}^2(\mathrm{Vs})^{-1}$ at $n_e = 3\times10^{11}\,\mathrm{cm}^{-2}$ are achieved. For a 2-D system the Fermi velocity is given by $v_F = \hbar\sqrt{2\pi n_e}/m^*$. With $m^* = 0.067m_e$ and using Equation 1.3, the length of the elastic mean free path is 9 µm.

1.2.2
Inelastic Mean Free Path

In addition to the elastic scattering discussed above, electron scattering can also be connected to an energy transfer. A typical example is the effect of lattice vibrations on electron transport. An electron moving within a crystal will be scattered by these lattice vibrations and either lose or gain energy, depending on whether it excites the lattice vibrations or is excited by them. As an energy transfer occurs, these scattering processes are considered to be *inelastic*. Similar to the previous discussion, one can define an inelastic scattering length l_{in} as a measure for the length between inelastic scattering events. Besides electron–phonon scattering, electron–electron scattering is another possible process, where a considerable amount of energy can be exchanged between both scattering partners [3].

1.2.3
Phase-Coherence Length

The phase-coherence length l_φ is the relevant length scale, which determines if phase-coherent transport can be observed in nanolectronic systems [2]. It is a measure of the distance that the electron propagates phase coherently before its phase is randomized. At low temperatures, the phase-coherence length can be larger than the elastic mean free path l_e. Thus, a number of elastic scattering events occur before the phase information is finally lost. During an elastic scattering event (i.e., at an impurity), the phase of an electron is not randomized; it is only shifted by well-defined amount. If the electron propagates along the identical path a second time, the phase accumulation will be exactly the same. This is in strong contrast to inelastic scattering events (e.g., electron–phonon scattering), where the scattering target changes with time. Consequently, the phase shift that the electron would acquire is different each time. However, care must be taken to identify l_φ right away with the inelastic mean free path l_{in}, as they are not identical in all cases; that is, spin-flip scattering is considered to be phase-breaking and thus contributing to l_φ whilst it may be elastic at the same time. In addition, small-energy-transfer electron–electron scattering, which is due to the fluctuation of the electric field produced by the electrons (Nyquist contribution), can contribute to a large extent to l_φ [4]. As mentioned above, at low temperatures a number of elastic scattering events occur

Table 1.1 Comparison of the different transport regimes.

Diffusive	Classical	$\lambda_F,\ l_e \ll L,\ l_\varphi < l_e$
	Quantum	$\lambda_F,\ l_e \ll L,\ l_\varphi > l_e$
Ballistic	Classical	$\lambda_F \ll L < l_e,\ l_\varphi$
	Quantum	$\lambda_F \approx L < l_e,\ l_\varphi$

until the phase is broken, implying that the characteristic phase-breaking time τ_φ is larger than the elastic scattering time τ_e. Owing to the diffusive motion during the time τ_φ, the phase-coherence length l_φ must be expressed by

$$l_\varphi = \sqrt{D\tau_\varphi} \tag{1.4}$$

Here, D is the diffusion constant defined as

$$D = \frac{1}{d} v_F^2 \tau_e \tag{1.5}$$

with d the dimensionality of the system. Typical values for the phase-coherence length of an AlGaAs/GaAs 2-D electron gas below 1 K are of the order of several micrometers [5].

1.2.4 Transport Regimes

By comparing l_e and l_φ with the dimension L of the sample and the Fermi wavelength λ_F, different transport regimes can be classified, and these are summarized in Table 1.1. For the case where the elastic mean free path l_e is smaller than the dimensions of the sample, many elastic scattering events occur while the electrons propagate through the structure. The carriers are traveling randomly (*diffusive*) through the crystal, as illustrated in Figure 1.1a. If the phase-coherence

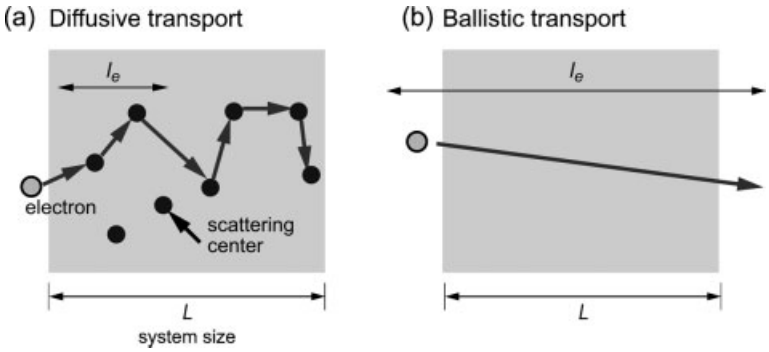


Figure 1.1 Illustration of (a) a diffusive conductor, and (b) a ballistic conductor. In the diffusive transport regime many elastic scattering events occur, while the electron crosses the sample. In the ballistic regime, the electron crosses the sample without any elastic scattering event.

length l_φ is shorter than the elastic mean free path l_e, the transport is considered as classical. In contrast, if $l_\varphi > l_e$, then quantum effects owing to the wave nature of the electrons can be expected. This diffusive regime is thus called the *quantum* regime. As illustrated in Figure 1.1b, in the case that l_e is larger than the dimensions of the sample, the electrons can transverse the system without any scattering; this regime is called *ballistic*. Depending on the magnitude of the Fermi wavelength λ_F in comparison to the dimension of the sample, the transport can either be regarded as classical ballistic or quantum ballistic. In the following section, ballistic transport will first be discussed, and later the transport phenomena in the diffusive regime.

1.3 Ballistic Transport

In this section transport in the ballistic transport regime will be discussed; that is, where the elastic mean free path exceeds the dimensions of the sample. First, the Landauer–Büttiker formalism is explained, where the resistance of a sample is described in terms of transmission and reflection probabilities, which is a very convenient scheme to analyze the transport in the ballistic regime. Subsequently, the quantized conductance of a split-gate point contact will be discussed, making use of the Landauer–Büttiker formalism.

1.3.1 Landauer–Büttiker Formalism

In order to analyze the electronic transport properties of a sample, usually a current is allowed to flow between two contacts while the response of the system is measured by two voltage probes. The latter are not necessarily different from the current contacts. The ratio between the voltage drop U and the current I can be defined as a *macroscopic resistance*. (The expression *macroscopic* is used here as only the global properties of the sample are measured.)

A very intuitive interpretation of the macroscopic resistance, R, of a sample can be obtained if the so-called Landauer–Büttiker formalism is used [6–9]. In this model, the resistance

$$\mathcal{R}_{mn,kl} = \frac{U_{kl}}{I_{mn}} \tag{1.6}$$

is defined by the voltage measured between contacts k and l and the current flowing between contacts n and m.

In order to keep things simple, the discussion is restricted to a conductor connected via ideal one-dimensional (1-D) ballistic leads to four corresponding reservoirs. The geometry of the sample is depicted in Figure 1.2. The ballistic wires should consist of only a single 1-D. The reservoirs with the corresponding chemical potentials μ_i ($i = 1, \ldots, 4$) serve as source and drain for carriers flowing in and out of

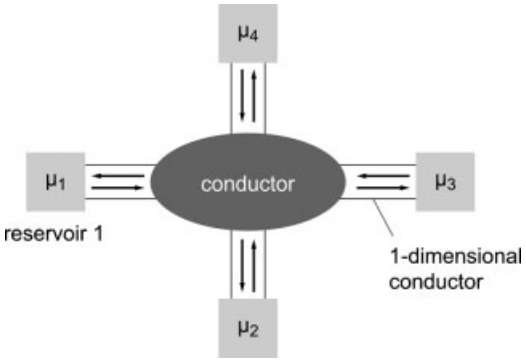


Figure 1.2 Schematic illustration of a four-terminal resistance measurement set-up. The conductor is connected by ideal one-dimensional leads to four corresponding reservoirs.

the conductor. At zero temperature, the i-th reservoir can supply electrons to the conductor up to a maximum energy of μ_i. Each carrier from the lead, which reaches the reservoir is absorbed by the reservoir, irrespective of the phase and energy of the carriers. As discussed above, inelastic scattering is forbidden within the leads, so that electrons once injected into the conductor maintain their energy until they reach one of the reservoirs.

As an example, we will study the current contributions in the 1-D lead 1, which results in the net current I_1. The current injected from reservoir 1 is given by:

$$I_{inj} = e \int_0^{\mu_1} D_{1D}(E) v(E) dE \tag{1.7}$$

where $v(E)$ is the velocity of the electrons. As the wire is 1-D, the density of states of a 1-D system must be inserted, which is given by

$$D_{1D}(E) = \frac{2}{hv(E)} \tag{1.8}$$

So far, only the states propagating from reservoir 1 are considered, and the density of states used here is half of the commonly known value because there is only one direction of propagation [1]. It can be seen directly that the product of the 1-D density of states $D_{1D}(E)$ and the velocity $v(E)$ is constant, and therefore the current leaving reservoir 1 has the following simple form:

$$I_{inj} = \frac{2e}{h} \mu_1 \tag{1.9}$$

Part of the current supplied by reservoir 1 will be reflected back into the conductor. If R_{ii} is defined as the reflection probability for a reflection of carriers from lead i back into lead i, then the current reflected into lead 1 can be written as

$$I_R = -\frac{2e}{h} R_{ii} \mu_i \tag{1.10}$$

In addition, electrons are transmitted from the other three leads into lead 1. By defining the transmission probability from lead j into lead i ($i \leftarrow j$) as T_{ij}, we arrive at the following expression for the current transmitted into lead 1:

$$I_T = -\frac{2e}{h}\sum_{j=2}^{4} T_{1j}\mu_j \tag{1.11}$$

By summing all of these contributions it can be seen that the net current flowing in lead 1 is finally given by:

$$\begin{aligned} I_1 &= I_{inj} + I_R + I_T \\ &= \frac{2e}{h}\left[(1-R_{11})\mu_1 - \sum_{j=2}^{4} T_{1j}\mu_j\right], \end{aligned} \tag{1.12}$$

or, more generally, the current in lead i is given by

$$I_i = \frac{2e}{h}\left[(1-R_{ii})\mu_i - \sum_{j\neq i} T_{ij}\mu_j\right] \tag{1.13}$$

By using Equation 1.13, the above-defined resistance $\mathcal{R}_{mn,kl}$ can be determined for given reflection and transmission probabilities of the sample. According to the initial definition of $\mathcal{R}_{mn,kl}$, as given in Equation 1.6, the net current I_{mn} flows between contacts n and m. The leads k and l do not carry a net current in case of an ideal voltage measurement. The voltage drop U_{kl} is given by the difference of the electrochemical potentials divided by e: $(\mu_k - \mu_l)/e$. In the following section, Equation 1.13 will serve as a basis to describe the transport properties of a split-gate quantum point contact.

1.3.2
Split-Gate Point Contact

In split-gate quantum point contacts the transport is limited to only one dimension. This is obtained by first restricting the propagation of the electrons to a plane. In these so-called "two-dimensional electron gases" (2DEGs), the carriers are confined at an interface of two different semiconductor layers. A typical example of a 2DEG realized in an AlGaAs/GaAs layer system is depicted in Figure 1.3. Here, the carriers are located at the AlGaAs/GaAs interface and, owing to the conduction band offset between AlGaAs and GaAs, a triangular quantum well is formed at the interface. The electrons in the quantum well are supplied by an n-type δ-doped (very thin) layer. In order to prevent ionized impurity scattering, the electrons in the quantum well are separated from the δ-doped layer by an undoped AlGaAs spacer layer. Using this scheme, very large electron mobilities and thus very long elastic mean free paths of the order of several micrometers can be achieved.

A further restriction of the electron propagation to only one dimension can be realized by using split-gate point contacts [10, 11]. As illustrated in Figure 1.4, two opposite gate fingers are separated by a distance of a few hundreds of nanometers. Split-gate electrodes are usually prepared by using electron beam lithography. Since

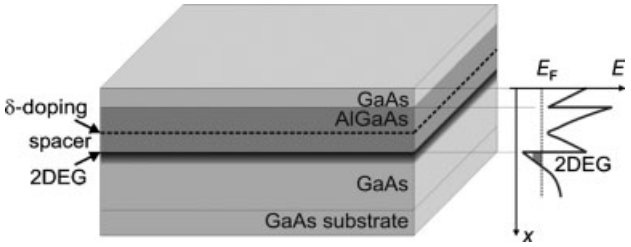


Figure 1.3 Layer sequence of an AlGaAs/GaAs heterostructure containing a two-dimensional electron gas at the AlGaAs/GaAs interface. A schematic illustration of the conduction band profile is shown on the right-hand side.

the Fermi wavelength λ_F of a 2-D electron gas is typically a few tenths of a nanometer, the separation of the split-gates is comparable with λ_F. The length of the channel formed by the gate electrodes is usually smaller than 1 μm, and thus smaller than the elastic mean free path l_e. According to the classification introduced in Section 1.2, the transport can be considered as ballistic.

By applying a sufficiently large negative voltage to the gate fingers, the underlying 2-D electron gas is depleted underneath the gate fingers (see Figure 1.4a). Only a small opening between the gate fingers remains for the electrons to propagate from one side to the opposite side; however, by varying the gate voltage it is possible to control the effective width of the opening. An increase of the negative bias voltage enlarges the depletion area and thus reduces the opening width. At sufficiently large negative bias voltages the opening can even be closed completely (pinch-off).

Owing to the depletion area underneath the split-gate electrodes, it can be assumed that the electrons in the 2DEG are confined in a potential well along the y-axis, while the free propagation takes place along the x-axis. If the potential profile in the plane of the 2DEG induced by the split-gate electrodes is expressed by $V(x, y)$, the Hamilton

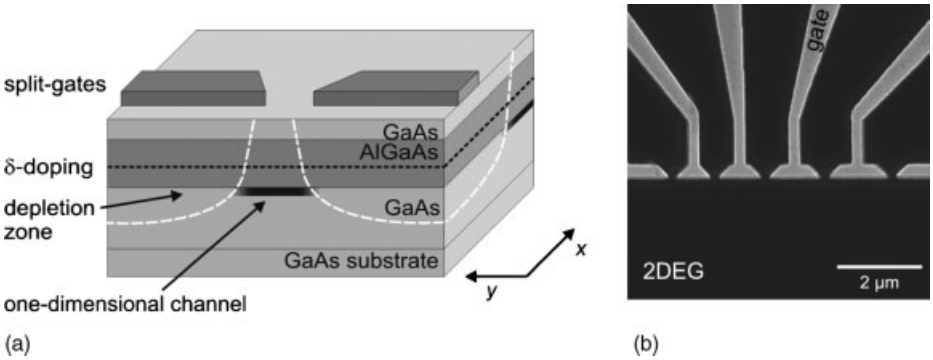


Figure 1.4 (a) Schematic illustration of a split-gate point contact on an AlGaAs/GaAs heterostructures. By applying a negative gate voltage to the split-gate electrodes, the electron gas underneath is depleted. The electrons can only pass the small opening. (b) An electron beam micrograph of split-gate point contacts.

operator has the following form:

$$H = \frac{\hbar^2}{2m^*}\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right) + V(x, y) \tag{1.14}$$

In order to determine the precise shape of the potential $V(x, y)$ as a function of the gate voltage, elaborated self-consistent simulations are required [12]. However, for most applications it is sufficient to assume an approximated potential profile. For low gate voltages an appropriate approximation is a rectangular potential profile, while for higher negative gate voltages the potential well can be approximated by a parabolic potential. As an example, we will consider here the latter potential shape. Due to the short length of the channel formed by the split-gates, the 2-D potential profile will be saddle-shaped. However, if the potential shaped along the constriction is smooth (adiabatic limit), it is sufficient to consider only the narrowest point of the channel, which can be expressed by

$$V(y) = \frac{1}{2} m^* \omega_0^2 y^2 + V_0 \tag{1.15}$$

Here, ω_0 is the characteristic frequency of the parabolic potential, while V_0 represents the height of the inflection point of the saddle-shaped potential. For the energy dispersion of the 1-D subbands in the point contact, we obtain

$$E_n(k_x) = E_n^0 + V_0 + \frac{\hbar^2 k_x^2}{2m^*}, \quad n = 1, 2, 3, \ldots, \tag{1.16}$$

with

$$E_n^0 = (n - 1/2)\hbar\omega_0, \tag{1.17}$$

the energy eigenvalues of the harmonic oscillator. By changing the gate voltage at the split-gate electrodes, the effective width of the opening can be adjusted. In the parabolic approximation ω_0 is increased if a more negative gate voltage is applied, and this leads to an increased separation of the energy eigenvalues. As a consequence, lesser levels are occupied up to the Fermi energy (see Figure 1.5a and b).

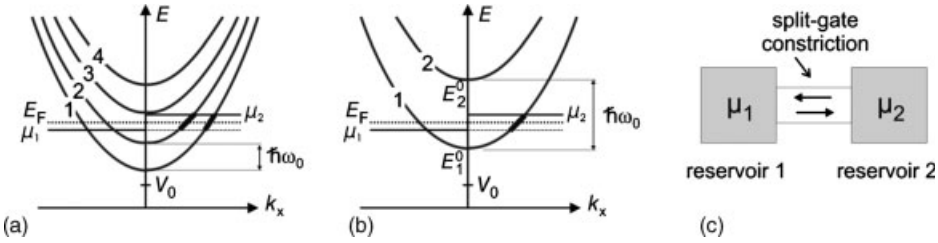


Figure 1.5 (a) Energy dispersion of a one-dimensional channel with the two lowest levels lying below the Fermi energy E_F. (b) Corresponding situation with only one subband occupied. The energy separation between the levels given by $\hbar\omega_0$ is larger compared to the situation shown in (b). (c) A one-dimensional conductor; that is, the channel formed by the split-gate electrodes, connected by two reservoirs with the electrochemical potential μ_1 and μ_2, respectively.

Before examining the experimental outcome of measurement of the split-gate point contact resistance, the conduction of a 1-D conductor by using the Landauer–Büttiker formalism will be briefly discussed. It must first be assumed that the conductor is connected on both terminals to reservoirs with the electrochemical potentials μ_1 and μ_2, respectively (i.e., the 2DEG on both sides of the split-gates), as shown in Figure 1.5c.

For a set-up with only two reservoirs, and where only the lowest subband is occupied, the following expression is obtained according to the Landauer–Büttiker formalism [cf. Equation 1.13]:

$$(h/2e)I = (1 - R_{11})\mu_1 - T_{12}\mu_2 \tag{1.18}$$

$$-(h/2e)I = (1 - R_{22})\mu_2 - T_{21}\mu_1 \tag{1.19}$$

At zero magnetic field ($B = 0$), the transport is time-inversion invariant so that the following relationships hold:

$$T_{12} = T_{21} = T = 1 - R_{11} = 1 - R_{22} \tag{1.20}$$

Thus, finally we arrive at the expression for the conductance of the constriction:

$$G = \frac{I}{U} = \frac{Ie}{\mu_1 - \mu_2} = \frac{2e^2}{h}T \tag{1.21}$$

As illustrated in Figure 1.5b, only carriers with energy between μ_1 and μ_2 contribute to the conductance. If backscattering is neglected ($T = 1$), the conduction through a constriction is simply given by:

$$G = \frac{2e^2}{h}. \tag{1.22}$$

It should be stressed that the constant conductance is a result of the cancellation of the energy dependence of the density of states and the velocity for the 1-D case [cf. Equation 1.7], which is not the case for 2-D or three-dimensional (3-D) systems. In analogy, the conductance can be calculated if N subbands are occupied. The occupied subbands taking part in the transport are usually called channels; the situation for two channels ($N = 2$) is illustrated in Figure 1.5a. If N one-dimensional channels are assumed, then the total transmission probability from reservoir j to reservoir i ($i \leftarrow j$) can be expressed as

$$T_{ij} = \sum_{mn}^{N} T_{ij,mn} \tag{1.23}$$

where $T_{ij,mn}$ denotes the transmission probability from the n-th subband of lead j into the m-th subband of lead i. If ideal transmission and no intersubband scattering is assumed, then the total transmission probability of a 1-D channel with N subbands is given by $T = N$. Thus, each subband contributes with $2e^2/h$ to the conductance so that the total conductance of a constriction with N subbands occupied is given by

$$G = \frac{2e^2}{h}N \tag{1.24}$$

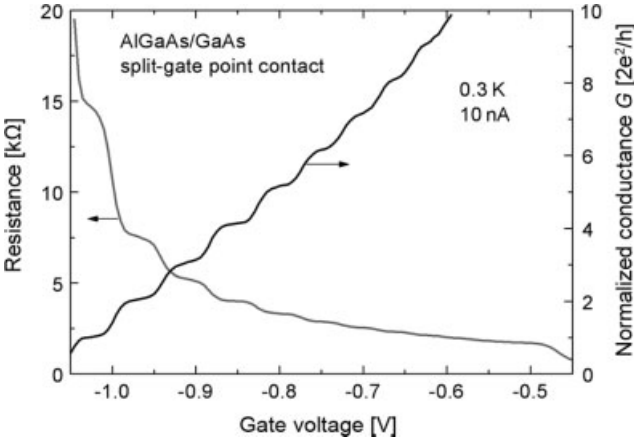


Figure 1.6 Resistance and conductance of an AlGaAs/GaAs split-gate point contact as a function of the gate voltage. The conductance is plotted in units of $2e^2/h$.

This remarkable result indicates that the conductance of a 1-D constriction changes in steps equal to $2e^2/h$, if the number of channels is altered by adjusting the widths of the constriction. The latter can be achieved by applying an appropriate voltage to the split-gate electrodes.

An experimental result of the resistance and conductance of quantum point contact based on a 2DEG in an AlGaAs/GaAs heterostructure is shown in Figure 1.6. With a more negative gate voltage, the resistance of the point contact increases, as the width of the constriction becomes increasingly narrower. As can be seen in Figure 1.6, if the conductance G is plotted, it can clearly be seen that G decreases stepwise by multiples of $2e^2/h$ with increasing negative gate voltage.

The experimentally observed curves can deviate in many aspects from the ideal curves. The calculations given above were restricted to zero temperature, but at finite temperatures the broadening of the Fermi distribution function results in a broadening of the steps owing to the partial occupation and emptying of the 1-D channels. The geometrical shape of the point contact opening also affects the transmission through the point contact. For example, sharp edges of the point contact opening can result in reflections of the incoming and transmitted electrons waves at the inlet and outlet of the 1-D channel. As a result, oscillations are expected in the plateaus of the steps [13, 14].

1.4 Weak Localization

Interference effects of electron waves due to phase coherent transport can be seen even in large samples, where the phase coherence length is much smaller than the dimensions of the sample. This effect, called weak localization, results in an increased resistance compared to the classically expected value [15, 16]. Weak localization is observed if the temperature is sufficiently low so that the phase coherence length l_φ is

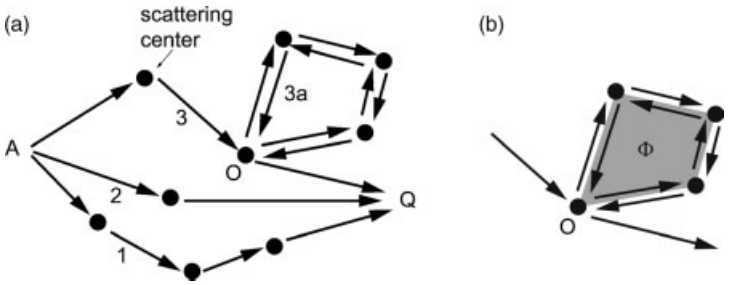


Figure 1.7 (a) Possible trajectories of electrons propagating from point *A* to *Q*. The trajectory 3a represents a closed loop. (b) Detail of a closed loop with a magnetic flux Φ penetrating this loop.

larger than the elastic scattering length l_e. As we will see below, the effect of weak localization depends strongly on the dimensionality of the system. The lower the dimension of the system is, the stronger the effect of weak localization is, that is in quasi one-dimensional wire structures weak localization is most pronounced. In order to illustrate the general mechanisms leading to weak localization, we will first introduce a simple model. Later on more quantitative expressions for the conductivity corrections will be given.

1.4.1 Basic Principles

Let us consider a diffusive conductor, in which an electrons starting at point *A* propagate to point *Q*. Some typical trajectories of an electron are sketched in Figure 1.7, illustrating that there are many possibilities for an electron to propagate from *A* to *Q*.

It is assumed that the elastic mean free path l_e is smaller than the distance between *A* and *Q*. Thus, an electron undergoes many elastic scattering events on its way. However, during elastic scattering the electron does not lose its phase memory. If it is assumed that the phase coherence length is longer than the distance between *A* and *Q*, the phase information is not lost. By following Feynman, each path *j* can be described from the initial state *A* to the final state *Q* by a complex probability amplitude C_j given by [17, 18]:

$$C_j = c_j \exp(i\varphi_j) \tag{1.25}$$

Here, φ_j is the phase shift that the electron acquires on its way from *A* to *Q* while propagating along path *j*. Often, there are many possible paths for an electron to propagate between *A* and *Q*. For example, for free electron propagation the phase accumulation along the path *j* can be calculated from the action S_j by

$$\varphi_j = \frac{S_j}{\hbar} \tag{1.26}$$

The non-relativistic action S_j is defined by

$$S_j = \int_{t_A}^{t_Q} dt \mathrm{L}(\boldsymbol{r}, \dot{\boldsymbol{r}}, t) \tag{1.27}$$

with

$$\mathrm{L}(\boldsymbol{r}, \dot{\boldsymbol{r}}, t) = \frac{m}{2}\dot{r}^2 \tag{1.28}$$

the Lagrangian function of a free propagating electron. Here, t_A is the time when the electron starts at A, and t_Q the time when it arrives at Q. The quantities $\boldsymbol{r}$ and $\dot{\boldsymbol{r}}$ are the position and velocity of the particle, respectively. However, the electron acquires not only a phase shift during free propagation but also well-defined phase shifts by the elastic scattering events, so that the total phase accumulated along the path is the sum of both contributions. The total amplitude for the propagation from A to Q is given by the sum of the amplitudes C_j of all undistinguished paths. Finally, the total probability P_{AQ} for an electron to be transported from A to Q is determined by the square of the total amplitude

$$P_{AQ} = \left| \sum_j c_j e^{i\varphi_j} \right|^2 \tag{1.29}$$

In systems with a large number of possible paths, the phases φ_j are usually randomly distributed, and therefore the wave nature should have no effect on the electron transport due to averaging. Nevertheless, the fact that an increase of the resistance is observed, compared to the classical transport, is a result of closed loops (see Figure 1.7a, trajectory 3a). Along these loops, an electron can propagate in two opposite orientations with the corresponding complex amplitudes $C_{1,2} = c_{1,2}\exp(i\varphi_{1,2})$. The current contribution of the current returning to the starting point of the loop (O) is given by

$$P_{OO} = |C_1 + C_2|^2 = |C_1|^2 + |C_1|^2 + 2\mathrm{Re}(C_1^* C_2) \tag{1.30}$$

Since, for time-reversed paths $c_1 = c_2$ and $\varphi_1 = \varphi_2$, we obtain

$$|C_1 + C_2|^2 = 4|C_1|^2 \tag{1.31}$$

For classical non-phase-coherent transport, the probability would simply be $|C_1|^2 + |C_2|^2$, which is a factor of 2 smaller than for the phase-coherent case. A larger probability to return to the origin implies that the net current through the sample is reduced. Hence, the carriers are *localized* within the loop. Such localization does not depend on the size of the loop as long as its length is smaller than the phase-coherence length. It is important to note here that constructive interference occurs for *all* possible closed loops in the conductor, and is therefore not averaged out. As a result, the total resistance is increased compared to the classical case.

1.4.2
Weak Localization in One and Two Dimensions

In the following section, it is briefly sketched how a value for the correction of the conductance due to weak localization can be obtained quantitatively [18]. For the weak localization effect we are interested only in those processes where the electrons return to their starting points. The discussion will first be restricted to a 2-D system, for example a 2-D electron gas in an AlGaAs/GaAs heterostructure. A larger number

of scattering centers increase the probability for backscattering of the electrons. The larger the number of scattering centers is, the smaller is the diffusion constant; as a consequence one obtains for the return probability due to diffusive motion: $1/(4\pi Dt)$. For the total return probability, it must be ensured that the phase of the electrons is preserved up to time τ_φ, which provides a pre-factor $\exp(-t/\tau_\varphi)$. Furthermore, it is required that the electron is at least once elastically scattered; thus, a pre-factor $[1-\exp(-t/\tau_e)]$ must be included. In total, the correction to the conductance can be expressed as [19, 20]:

$$\begin{aligned}\Delta\sigma_{2D} &= -\frac{2\hbar}{m^*}\sigma_0\int_0^\infty dt\frac{1}{4\pi Dt}(1-e^{-t/\tau_e})e^{-t/\tau_\varphi} \\ &= -\frac{e^2}{2\pi^2\hbar}\ln\left(1+\frac{\tau_\varphi}{\tau_e}\right)\end{aligned} \tag{1.32}$$

Here, σ_0 is the classical Drude conductivity of a 2-D system. The localization vanishes, if the phase-breaking time τ_ϕ is much smaller than τ_e, since then the logarithmic factor tends towards zero. The ratio of the correction due to weak localization to the Drude conductivity $\Delta\sigma_{2D}/\sigma_0$ is usually small and of the order of $1/k_F l_e$. Here, $k_F = m^* V_F/\hbar$ is the Fermi wavenumber. For a typical 2-D electron gas with $\mu_e = 10^6\,\mathrm{cm^2\,V\,s^{-1}}$ at $n_e = 3\times10^{11}\,\mathrm{cm^{-2}}$, a correction of less than 0.1% would be expected.

For a quasi 1-D structure of width W with $l_\varphi \gg W$, the diffusion is effectively reduced to one dimension, so that the return probability can now be expressed by $W^{-1}(4\pi Dt)^{-1/2}$. The conductivity correction in this case is given by [20]:

$$\Delta\sigma_{1D} = -\frac{e^2}{\pi\hbar}\frac{l_\varphi}{W}\left[1-\left(1+\frac{\tau_\varphi}{\tau_e}\right)^{-1/2}\right] \tag{1.33}$$

A comparison of the 1-D and 2-D cases reveals that the weak localization correction to the conductivity is much larger for the 1-D case. In the latter case, the ratio $\Delta\sigma_{1D}/\sigma_0$ is of the order $(l_\varphi/W)(1/k_F l_e)$. If a phase-breaking time of $\tau_\phi = 10^{-10}$ s and a width of $W = 200$ nm are assumed, the result is a ratio $\Delta\sigma_{1D}/\sigma_0$ of 6%, for a wire based on the 2-D electron gas as specified above. Clearly, this value is much larger than the corresponding value for a 2-D system.

1.4.3 Weak Localization in a Magnetic Field

If the sample is penetrated by a magnetic field $\boldsymbol{B}$, the phase accumulation along a certain trajectory is modified, since the Lagrangian function L [cf. Equation 1.28] of an electron with charge $-e$ contains an additional term

$$\mathrm{L}(\boldsymbol{r},\dot{\boldsymbol{r}},t) = \frac{m}{2}\dot{\boldsymbol{r}}^2 - e[\dot{\boldsymbol{r}}\boldsymbol{A}(r,t)] \tag{1.34}$$

Here, $\boldsymbol{A}$ is the vector potential defined by $\boldsymbol{B} = rot\ \boldsymbol{A}$. In the presence of a vector potential, the probability amplitude C_1 of a closed loop propagated in clockwise

orientation acquires an additional phase factor

$$C_1 \rightarrow C_1 \exp\left(-i\frac{e}{\hbar}\oint \boldsymbol{A}d\boldsymbol{l}\right) = C_1 \exp\left(i\frac{2\pi\Phi}{\Phi_0}\right) \tag{1.35}$$

Here, $\Phi = BS$ is the magnetic flux penetrating the enclosed area S of the loop, with $\Phi_0 = h/e$ the magnetic flux quantum. For the propagation in the opposite orientation one obtains

$$C_2 \rightarrow C_2 \exp\left(-i\frac{2\pi\Phi}{\Phi_0}\right) \tag{1.36}$$

The phase difference accumulated between both time-reversed paths is therefore

$$\Delta\varphi = 4\pi\frac{\Phi}{\Phi_0} \tag{1.37}$$

Thus, if a magnetic field is applied, the property that constructive interference occurs for *all* loops in case of $B = 0$ is lost. Generally, many loops enclosing different areas are found in a diffusive conductor and, depending on the size of the loops, different phase shifts $\Delta\varphi$ develop. Thus, for a particular magnetic field the localization is lifted to a different extent depending on the loop size. If the magnetic field is increased starting from zero, the constructive interference is destroyed first for the largest loops. Finally, if the magnetic field is sufficiently large, the phase difference will be randomly distributed between the ensemble of loops. On average, the degree of localization decreases with increasing magnetic field, resulting in a continuous decrease of the resistance.

For a quantitative approach one must take into account that, in addition to the usual phase breaking at zero magnetic field, the phase is also broken effectively by a magnetic field. Similar to l_φ a length l_m is defined, which is characterized by the condition that the area l_m^2 corresponds to the case that the penetrating flux is equal to Φ_0. Thus, l_m is defined by $\sqrt{\hbar/eB}$. As outlined above, for a flux Φ_0 the phase difference between time-reversed paths is already significant. The characteristic magnetic relaxation time τ_B related to l_m can be estimated from the relationship $l_m \sim \sqrt{D\tau_B}$, in analogy to Equation 1.4 defining l_φ. The expression that quantitatively describes the increase of the conductivity with increasing magnetic field is given by [21, 22]:

$$\Delta\sigma_{2D}(B) - \Delta\sigma_{2D}(0) = \frac{e^2}{2\pi^2\hbar}\left[\Psi\left(\frac{1}{2}+\frac{\tau_B}{2\tau_\varphi}\right) - \Psi\left(\frac{1}{2}+\frac{\tau_B}{2\tau_e}\right) + \ln\left(\frac{\tau_\varphi}{\tau_e}\right)\right] \tag{1.38}$$

where $\Psi(x)$ is the digamma function. The exact expression for τ_B, which must be inserted into Equation 1.38, is given by $\tau_B = l_m^2/2D$. At zero magnetic fields the relevant maximum size of the loops at which the phase coherence is broken is given by l_φ^2. In a finite magnetic field, weak localization is suppressed if a noticeable phase shift between time-reversed loops is accumulated. This is the case for loops with the area of about l_m^2. By comparing both relationships, it is clear that the magnetic field has a significant effect on the conductance for $l_\varphi^2 \approx l_m^2$. This relationship defines a critical

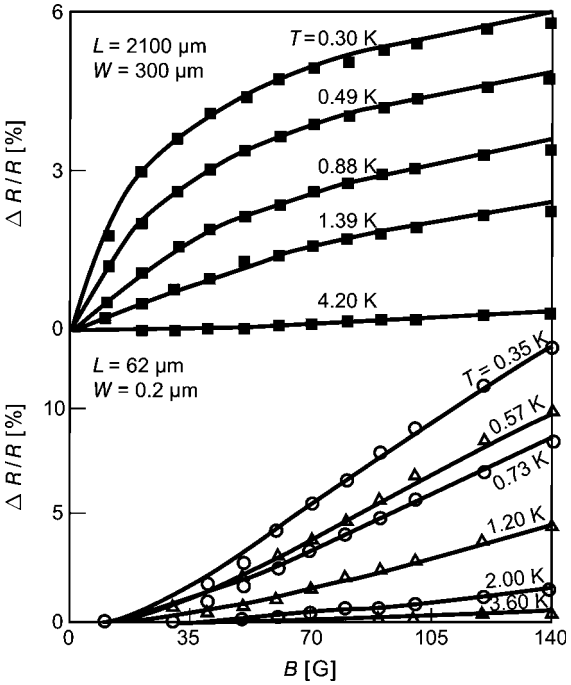


Figure 1.8 Comparison of the weak localization effect in a two-dimensional (upper graph) and a one-dimensional electron gas (lower graph) in AlGaAs/GaAs. For the one-dimensional structures a much higher magnetic field is required to suppress the weak localization effect. (Reprinted with permission from [23]. Copyright (1987) by the American Physical Society.)

magnetic field B_c, which is given by

$$B_c = \frac{\hbar}{2el_\varphi^2} \tag{1.39}$$

Thus, at the characteristic field of B_c one expects a suppression of weak localization. For semiconductor structures, l_φ may be of the order of 1 μm, and result in a critical field of about 1 mT. In the case of a 2-D electron gas, weak localization is suppressed at relatively low magnetic fields (see Figure 1.8).

In 1-D systems in the dirty metal limit, defined as $l_e \ll W \ll l_\varphi$, the closed trajectories contributing to weak localization are quenched in one direction, with a typical enclosed area of the loop given by $W\sqrt{D\tau_B}$ (see Figure 1.9a). For a unit phase shift this area corresponds to l_m^2, resulting in a magnetic relaxation time of $\tau_B \sim l_m^4/DW^2$ and a critical field of $B_c \sim \hbar/eWl_\varphi$. The full expression for the weak localization correction of one-dimensional systems in the dirty limit is given by [24]

$$\Delta\sigma_{1D}(B) = \frac{e^2}{\pi\hbar}\frac{\sqrt{D}}{W}\left[\left(\frac{1}{\tau_\varphi}+\frac{1}{\tau_B}\right)^{-1/2} - \left(\frac{1}{\tau_\varphi}+\frac{1}{\tau_e}+\frac{1}{\tau_B}\right)^{-1/2}\right] \tag{1.40}$$

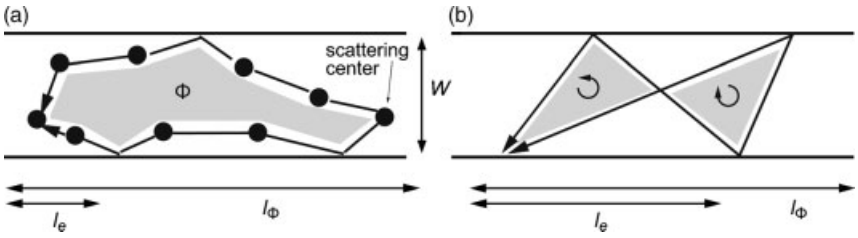


Figure 1.9 (a) Typical closed trajectory in a dirty metal one-dimensional conductor ($l_e \ll W \ll l_\phi$). (b) Typical closed trajectory in a narrow one-dimensional structure with $W \ll l_e$. Here, diffusive boundary scattering results in loops which self-interact. The net flux is cancelled in this configuration.

with magnetic relaxation time in this case given by $\tau_B = 3l_m^4/WD$. It should be noted that, at zero magnetic fields, Equation 1.32 is recovered. Furthermore, a closer inspection of B_c reveals, that if the width of the wire is reduced, the critical field is increased, ensuring that the weak localization effect is preserved up to much higher magnetic fields compared to the 2-D case. This is confirmed by the measurements shown in Figure 1.8, where the magnetoresistance peak is wider in the 1-D case. In wire structures based on high-mobility, 2-D electron gases, the elastic mean free path l_e may be larger than the width of the wire: $W \ll l_e$. In this ballistic regime, the electrons propagate without any scattering between the wire boundaries. As illustrated in Figure 1.9b, owing to diffusive boundary scattering the typical closed loops will self-interact. As both parts of the loop area are traversed in opposite orientation, the net flux is basically cancelled [20]. Clearly, the flux cancellation results in a further increase of the critical field.

1.5 Spin-Effects: Weak Antilocalization

So far, the effect of spin on the electron interference has been neglected, and this approach is valid as long as the spin orientation is conserved. However, in many materials the spin changes its orientation while the electron propagates along the closed loops, resulting in the weak localization effect.

It can be assumed that $|s\rangle$ is the initial spin state, this generally being a superposition of the spin up $|\uparrow\rangle$ and spin down $|\downarrow\rangle$ states. In principle, there are two possibilities of how the spin orientation can be changed:

- The Elliot–Yafet mechanism. Here, the potential profile of the scattering centers can lead to spin-orbit coupling; this results in a spin rotation, while the electron is scattered at the impurities (see Figure 1.10a).
- The so-called D'yakonov–Perel mechanism, where the spin precesses while the electron propagates *between* the scattering centers (see Figure 1.10b). The origin of the spin precession may either be a lack of inversion symmetry (i.e., in zinc blende

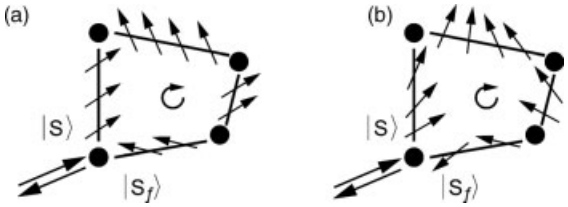


Figure 1.10 (a) Typical closed trajectory in forward direction with spin scattering at the impurities. The initial spin state $|s\rangle$ is transformed to the final spin state $|s_f\rangle$. The spin orientation is preserved while propagation between the scattering centers. (b) The situation where a spin precession occurs while the electron propagates between the scattering centers.

crystals; the Dresselhaus effect [25]), or an asymmetric potential shape of the quantum well forming a 2-D electron gas (the Rashba effect) [26].

Further details on spin precession are provided in Chapter 3 of this volume and Chapter 5 of volume 4 of this series (Bandyopadhyay, S., Monolithic and Hybrid Spintronics. In: Schmid, G. (ed), Nanotechnology, Vol 4, Chapter 5).

Regardless of the underlying mechanism, if an electron propagates along a closed loop, its spin orientation is changed. The modification of the spin orientation can be expressed by a rotation matrix $\mathbf{U}$ [27]. For the propagation along the loop in forward (f) direction the final state $|s_f\rangle$ can be expressed by

$$|s_f\rangle = \mathbf{U}|s\rangle \tag{1.41}$$

where $\mathbf{U}$ is the corresponding rotation matrix. For propagation along the loop in a backwards directions (b), the final spin state is given by

$$|s_b\rangle = \mathbf{U}^{-1}|s\rangle \tag{1.42}$$

Here, use is made of the fact that the rotation matrix of the counter-clockwise propagation is simply the inverse of $\mathbf{U}$. For interference between the clockwise and counter-clockwise electron waves, not only the spatial component is relevant but also the interference of the spin component:

$$\begin{aligned}\langle s_b|s_f\rangle &= \langle \mathbf{U}^{-1}s|\mathbf{U}s\rangle \\ &= \langle s|\mathbf{U}^{\dagger}\mathbf{U}|s\rangle \\ &= \langle s|\mathbf{U}^2|s\rangle.\end{aligned} \tag{1.43}$$

The final expression was obtained by making use of the fact that $\mathbf{U}$ is a unitary matrix: $\mathbf{U}^{-1} = \mathbf{U}^{\dagger}$, with $\mathbf{U}^{\dagger}$ the adjoint (complex conjugated and transposed) matrix of $\mathbf{U}$. Weak localization – and thus constructive interference – is recovered if the spin orientation is conserved in the case that $\mathbf{U}$ is the unit matrix $\mathbf{1}$.

However, if the spin is rotated during electron propagation along a loop, in general no constructive interference can be expected. Moreover, for each loop a different interference will be expected. Interestingly, averaging over all possible trajectories even leads to a reversal of the weak localization effect such that, instead of an increase

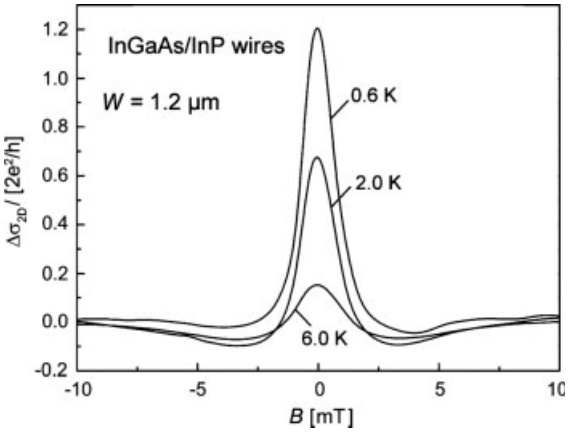


Figure 1.11 Magneto-conductivity measured on a set of 160 InGaAs/InP wires at various temperatures [29]. The Rashba spin-orbit coupling, present in this type of quantum well, results in weak antilocalization, an enhanced conductivity at zero magnetic field. The wires had a geometrical width of 1.2 µm.

in the resistance, a decrease occurs [22, 27, 28]. As the sign of the quantum mechanical correction to the conductivity is reversed, this effect is referred to as "weak antilocalization".

The weak antilocalization measurements of a set of InGaAs/InP wires are shown in Figure 1.11. In contrast to the weak localization effect, an enhanced conductivity is found at $B = 0$. However, if a magnetic field is applied then the weak antilocalization effect is gradually suppressed. Notably, important parameters characterizing the spin scattering and spin precession can be extracted from weak antilocalization measurements. In fact, detailed information of the Rashba and Dresselhaus contributions in a particular material can be obtained by fitting the experimental curves to the appropriate theoretical model. It should be noted that both contributions are important for the spin field effect transistor, as introduced in Chapter 3 of this volume and Chapter 5 of volume 4 of this series (Bandyopadhyay, S., Monolithic and Hybrid Spintronics. In: Schmid, G. (ed), Nanotechnology, Vol 4, Chapter 5).

1.6 Al'tshuler–Aronov–Spivak Oscillations

The fact that, in a metallic conductor, closed loops are responsible for the reduction of the resistance if a magnetic field is applied raises the question: Is it possible to observe resistance oscillations due to interference if the shape of the closed loops are restricted by a fixed, well-defined geometry? In the following sections, it will be shown that indeed these oscillations – the so-called Al'tshuler–Aronov–Spivak oscillations – can be observed in ring-shaped conductors, if the rings are penetrated by a magnetic flux. A series of interconnected ring-shaped conductors is shown schematically in

Figure 1.12 A series of interconnected ring-shaped conductors. Each ring is penetrated by a magnetic flux Φ. The interference of the time-reversed trajectories leads to the Al'tshuler–Aronov–Spivak oscillations as a function of a magnetic field.

Figure 1.12, where the enclosed magnetic flux Φ is the same for each ring. Thus, the phase shift $\Delta\varphi$ between time-reversed paths, as given by Equation 1.37, is approximately the same in all rings. By using Equation 1.30, the total amplitude in a loop is given by [30, 31]:

$$\begin{aligned} P &= \left| C_1 \exp\left(i2\pi \frac{\Phi}{\Phi_0} \right) + C_2 \exp\left(-i2\pi \frac{\Phi}{\Phi_0} \right) \right|^2 \\ &= 2|C_1|^2 [1 + \cos(4\pi\Phi/\Phi_0)]. \end{aligned} \tag{1.44}$$

From the equation given above it can be concluded that the resistance in this type of structure should oscillate with a period of $\Phi_0/2$.

The first demonstration of this type of weak localization resistance oscillations was provided by Sharvin and Sharvin [31], who evaporated a thin Mg film onto the surface of a quartz filament. The magnetic field was applied in axial orientation with respect to the filament while the current was flowing through the Mg film along the filament. A comparison with the cross-section of the filament confirmed, that the resistance oscillations indeed had a period of $\Phi_0/2$.

Beside cylindrical samples, Al'tshuler–Aronov–Spivak oscillations can also be observed in planar quantum wire networks, similar to the structure shown in Figure 1.12 [33]. The closed trajectories are realized by squares connected to a chain or to a mesh, as shown in Figure 1.13 (inset).

The relative resistance difference $\Delta R/R_0$ as a function of a perpendicular magnetic field is shown in Figure 1.13. Pronounced oscillations are found in the chain as well as in the mesh structure. For smaller square elements, the oscillation period is larger as a larger magnetic field is required to generate a magnetic flux of $\Phi_0/2$. In order to observe Al'tshuler–Aronov–Spivak oscillations, the phase-coherence length must be larger than the circumference of the squares. For the chain structure, the total resistance is given by adding the contribution of each single square. Depending on the type of material, each ring produces either a maximum or minimum at $B = 0$, depending on the absence or presence of spin scattering. This ensures that, after summation of the contribution of each element of the chain, the Al'tshuler–Aronov–Spivak oscillations are not averaged out.

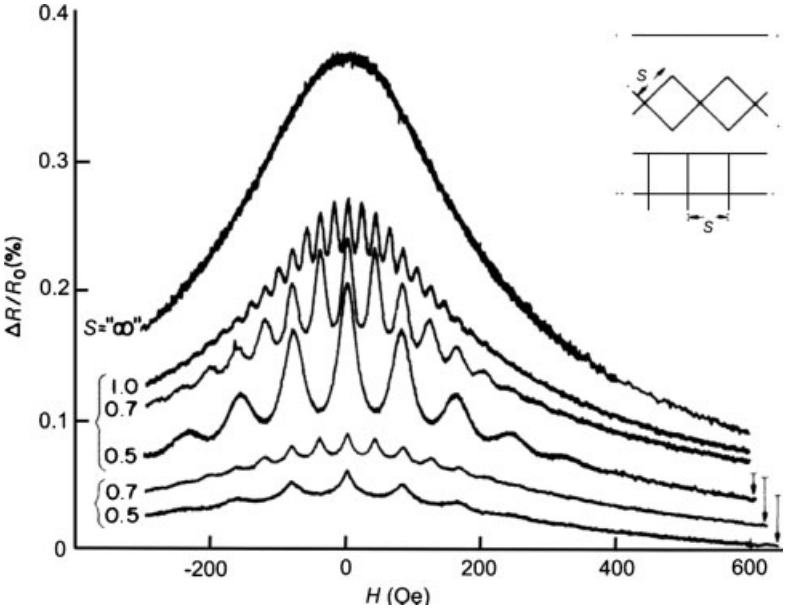


Figure 1.13 Magnetoresistance of 21 nm-thick and 55 nm-wide lithium wires of different geometry measured at 0.13 K. The upper curve shows the measurement of a single wire. Here, a resistance maximum at $B = 0$ due to weak localization is found. The following three curves show the resistance for a chain of squares. The size of the elements decreases for lower curves, respectively. The lowest two curves show the measurement on a mesh structures. The size S of the unit cell (in micrometers) is indicated next to each curve. (Reprinted with permission from [33]. Copyright (1986) by the American Physical Society.)

1.7 The Aharonov–Bohm Effect

In contrast to the Al'tshuler–Aronov–Spivak oscillations, which originate from the interference of electrons propagating along time-reversed paths interfering at the point of departure, the Aharonov–Bohm effect is based on electron waves propagating along two different branches of a ring structure and interfering at the opposite side of the ring. This situation is illustrated schematically in Figure 1.14a.

The Aharonov–Bohm effect was predicted, from a theoretical point of view, in 1959 [34]. The essence of this effect is that the vector potential $\boldsymbol{A}$ affects the interference of the electron waves, even in the case when the magnetic field $\boldsymbol{B}$ in the conductor is zero. In order to clarify this point, the experimental set-up shown in Figure 1.14a will be discussed. Here, the magnetic field $\boldsymbol{B}$ is restricted to an area within the ring structure (the gray-shaded area in Figure 1.14a), and is zero in the ring-shaped conductor. Classically no effect is expected since, at the location of the electrons, no magnetic field is present. However, as seen in Section 1.6, the vector potential $\boldsymbol{A}$, which is non-zero in the conductor, will induce a phase shift of the electron wave and thus affect the electron transport.

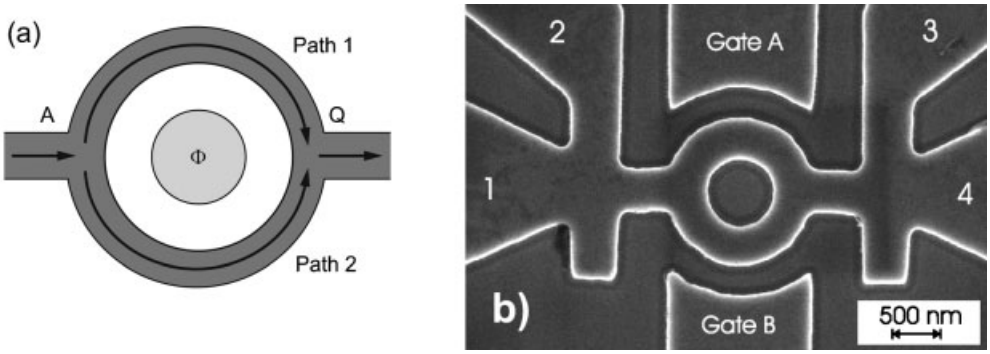


Figure 1.14 (a) Electron trajectories in a ring-shaped conductor. For the Aharonov–Bohm effect the ring is penetrated by a magnetic field within the inner diameter of the conductor. No magnetic field is applied within the conductor. The magnetic flux within the ring is Φ. (b) Electron beam micrograph of an AlGaAs/GaAs ring structure with two in-plane gates (A and B).

The phase difference $\Delta\varphi$ of two electron waves propagating along the upper and the lower branches of the ring (paths 1 and 2 in Figure 1.14a) and interfering at the end point Q of the ring is given by

$$\begin{aligned}\Delta\varphi &= \chi_1 - \chi_2 - \frac{e}{\hbar}\int\limits_{path1} \boldsymbol{A}d\boldsymbol{l} + \frac{e}{\hbar}\int\limits_{path2} \boldsymbol{A}d\boldsymbol{l} \\ &= \Delta\chi + \frac{e}{\hbar}\oint \boldsymbol{A}d\boldsymbol{l}.\end{aligned} \tag{1.45}$$

Here, χ_1 and χ_2 are the phases that the electron waves acquire during their propagation along path 1 and path 2 at zero magnetic field in the interior of the ring. In contrast to weak localization, the paths are different and therefore not time-reversed. Since the impurity configurations on both branches usually differ, the accumulated phases are different in both branches.

We will now return to the Aharonov–Bohm effect itself. By making use of $rot\,\boldsymbol{A} = \boldsymbol{B}$, Equation 1.45 results in

$$\begin{aligned}\Delta\varphi &= \Delta\chi + \frac{e}{\hbar}\int \boldsymbol{B}d\boldsymbol{f} \\ &= \Delta\chi + 2\pi\frac{\Phi}{\Phi_0}.\end{aligned} \tag{1.46}$$

The surface integral over $\boldsymbol{B}$ corresponds to the magnetic flux Φ penetrating the ring. As illustrated in Figure 1.14a, the area penetrated by the magnetic field does not need to be as large as the opening of the ring. As can be inferred from Equation 1.46, a phase shift of 2π is acquired if the magnetic flux is changed by a magnetic flux quantum Φ_0. Thus, the period is twice as large as the period of the Al'tshuler–Aronov–Spivak oscillations discussed above.

For the first experiments demonstrating the Aharonov–Bohm effect, a set-up was used where the electrons were not exposed to a magnetic field [36, 37]. These experiments were performed with an electron beam in a vacuum and a shielded magnet coil. In solid-state the Aharonov–Bohm effect was first demonstrated in Au rings, with the diameter of the ring structure being less than 1 μm and a wire width of a few tens of nanometers [38]. In metallic ring structures, the magnetic field cannot usually be prevented from penetrating the wire itself. Nevertheless, the vector potential $\boldsymbol{A}$ is still responsible for the effect on the electron interference pattern. A typical ring structure defined in an AlGaAs/GaAs semiconductor heterostructures is shown in Figure 1.14b [35]. One important difference between the Aharonov–Bohm experiments on nanoscaled rings and experiments using electron beams in vacuum, is that in the former case the electrons are usually scattered many times within the conductor before reaching the opposite side of ring. Thus, the elastic mean free path is most often smaller than the ring size. In addition, in metallic or semiconducting ring structures the phase-coherence length is in the order of the ring diameter at low temperatures, and consequently many electrons lose their phase memory while propagating through the ring. This is the reason why the oscillation amplitude is considerably smaller than the total resistance of the structure. As can be seen in Figure 1.15, pronounced Aharonov–Bohm oscillations were observed in ring structures based on 2-D electron gases in an $In_{0.77}Ga_{0.23}As/InP$ heterostructure [39]. A comparison of the enclosed area of the ring confirmed that the oscillation period corresponded to a magnetic flux quantum Φ_0. Owing to the low effective electron mass and to the high mobility in these heterostructures, the phase-coherence length can exceed 1 μm at temperatures below 1 K, and consequently large oscillation amplitudes are achieved. Previously, resistance modulations of up to 12% have been observed in this type of structure.

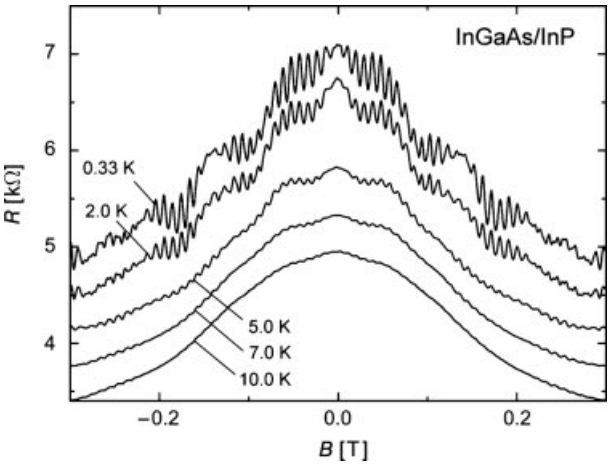


Figure 1.15 Magnetoresistance of an $In_{0.77}Ga_{0.23}As/InP$ ring structure measured at 0.3 K. The ring had a diameter of 820 nm, with a width of the wires forming the ring of about 85 nm. (Reprinted with permission from [39]. Copyright (1995) by the American Physical Society.)

When the Al'tshuler–Aronov–Spivak oscillations were discussed, it was found that at $B = 0$ a maximum is observed in the resistance, owing to the constructive interference between time-reversed loops (see Figure 1.13). For the Aharonov–Bohm effect, the interference at the branching point is determined by the two different paths 1 and 2 along the two branches of the ring, as illustrated in Figure 1.14a. The phase difference $\Delta\chi$ at $B = 0$ between the two paths depends to a large extent on the distribution of scattering centers in the ring; that is, for different rings – regardless of whether they have the same geometry – a different phase shift $\Delta\chi$ is accumulated. As a result, no clear maximum or minimum, as for the Al'tshuler–Aronov–Spivak oscillations, is expected at $B = 0$. In fact, it could be shown that the amplitude of the Aharonov–Bohm oscillations is decreased if the signal of many rings is averaged [40]. This is due to the fact that the contributions of interferences of the different rings with the same oscillation period but with statistically distributed phase shifts $\Delta\chi$ are averaged out.

Besides the magnetic control of the interference pattern of an Aharonov–Bohm ring structure, the oscillation pattern can also be changed electrostatically by means of a gate electrode. A typical AlGaAs/GaAs sample with two in-plane gates is shown in Figure 1.14b. By applying a voltage to one of the gates, the electron concentration in the corresponding branch of the ring is altered. A change of the carrier concentration goes along with a change of the Fermi wavelength and, as a result, the phase accumulated in this branch of the ring is changed. Clearly this will immediately affect the interference pattern, as can be seen in Figure 1.16.

It is interesting to observe that the oscillation pattern of the sample is symmetric with respect to the magnetic field. This can be seen in Figure 1.16, where the resistance oscillations are shown as a grayscale plot as a function of magnetic field

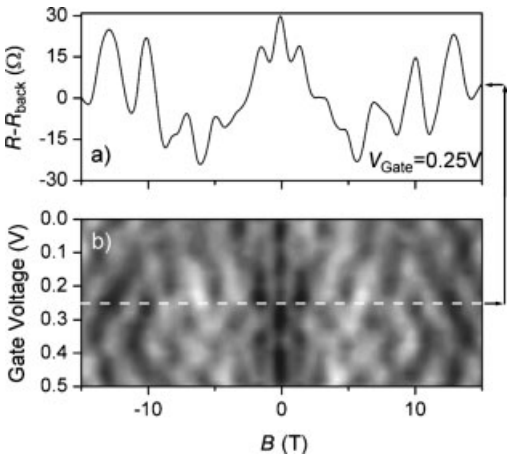


Figure 1.16 (a) Magnetoresistance of the AlGaAs/GaAs ring structure shown in Figure 1.14b for 0.25 V applied to gate B. The voltage at gate A was set to zero. The background resistance R_{back} was subtracted from the total resistance. (b) Grayscale plot of the magnetoresistance $R\text{-}R_{back}$ for different voltages applied to gate A and B. The dark and light regions correspond to large and low resistance values, respectively. The dashed line indicated the measurement shown in (a).

and gate voltage. The symmetric pattern is due to the fact that, although four terminals are used during the measurement, the measurement is effectively a two-terminal measurement. For such a measurement it can be shown in general that the resistance is symmetric under reversal of the magnetic field: $R(-B) = R(B)$ [9]. The reason for the two-terminal nature of the measurement is the coupling of the two contacts on each side, although these are not independent, as would be required for a pure four-terminal measurement.

1.8 Universal Conductance Fluctuations

A closer inspection of the Aharonov–Bohm effect measurements reveals that irregular fluctuations are often superimposed on the regular oscillations. This can be seen clearly in the measurements shown in Figure 1.15 where a long-wavelength underground is observed superimposed on the oscillations. The fluctuations are reproducible if the measurements are repeated for the same sample [42, 43]; however, if different samples with the same geometry and fabricated using the same material are compared, it is found that a different fluctuation pattern belongs to each sample. This is the reason, why the individual fluctuation pattern of a sample is sometimes referred to as a *fingerprint*. Conductance fluctuations are observed in semiconducting structures as well as in metallic samples [44, 45]. However, as the size of semiconductor devices becomes smaller, the statistical distribution of the remaining few dopant atoms will result in a spread of the device characteristics. Thus, resistance fluctuations are not only important in the phase-coherent regime but are also becoming much more of an issue in device applications.

1.8.1 Basic Principles

The reason, why each sample shows a different conductance fluctuation pattern becomes clear when it is realized that very few scattering centers (i.e., impurities) are present in very small structures. In fact, if the sample possesses very few impurities then it is the impurity configuration that governs the transport properties. Moreover, if only a finite number of scattering centers is present, an ensemble average cannot be applied for the theoretical description, as this does not take into consideration the particular spatial distribution of the scattering centers. Of course fluctuations may be observed not only in ring structures but also in a single quantum wire, as they originate from the spatial configuration of the scattering centers in the wire. The conductance fluctuations of a single Au wire are shown in Figure 1.17.

The fluctuations shown in Figure 1.17 cover an interval of approximately $\pm e^2/h$, the fluctuation amplitude of which, as proven by a detailed theoretical analysis, is universal [47, 48]. For a qualitative explanation of the physical origin of the conductance fluctuations, the reader is referred to the above-mentioned Aharonov–Bohm effect. As illustrated in Figure 1.18a, the electron is able to propagate along a certain number of paths in order to cross the wire.

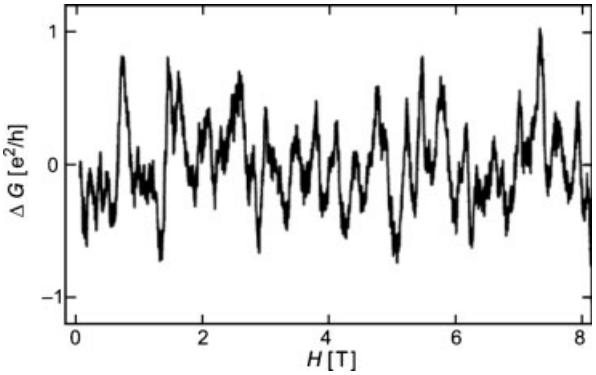


Figure 1.17 Conductance fluctuations as a function of magnetic field of a 310 nm-long and 25 nm-wide Au wire measured at a temperature of 10 mK [46]. (© Taylor &Francis Ltd.).

The total transmission probability results from the squared amplitude of all possible trajectories. Among these trajectories, a limited number of paths may be found which meet again after a certain distance. If a magnetic field is applied, the paths become penetrated by a magnetic flux Φ; subsequently, if the magnetic field is varied, superpositioning of the electron waves of two paths (which cross twice) leads to a variation in the transmission probability due to the Aharonov–Bohm effect. Then, in contrast to a well-defined ring structure, the encircled areas differ among the various locations of the wire, such that a different Aharonov–Bohm period is developed for each area. Superposition of the different quasi Aharonov–Bohm rings then produces an irregular conductance pattern [43]. It is important that only a limited number of trajectories exists, so that an effective averaging out of the

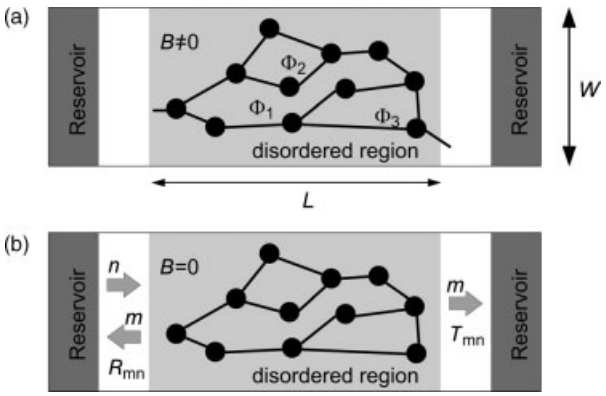


Figure 1.18 (a) Electron trajectories in a quantum wire. If a magnetic field is applied, loops are penetrated by a magnetic flux Φ_1, Φ_2, Φ_3, ... (b) Sample configuration considered for the calculation of the conductance fluctuations using the Landauer–Büttiker formalism. The wire consists of a disordered region connected by ballistic areas to the phase-randomizing reservoirs. Here, *n* denotes the incoming channel, while *m* is the outgoing channel (transmitted or reflected).

oscillations is prevented. In addition to varying the magnetic field, it is also possible to observe conductance fluctuations by increasing the applied voltage, but this will lead to a change in the Fermi wavelength of the electrons.

1.8.2 Detailed Analysis

A detailed theoretical description of conductance fluctuations is based on the particular scattering center configuration [47, 48]. Hence, by using this theoretical approach it was possible to calculate the average oscillation amplitude of sample-specific conductance fluctuations. One important point of the theoretical model is that a variation in the magnetic field or the Fermi energy induces the same type of fluctuation as would an ensemble average (quasi-ergodic hypothesis). This allows the measurement of only a single sample rather than measuring numerous different wires at zero magnetic field and constant Fermi energy.

In the following section it will first be shown that the magnitude of the conductance fluctuations at zero temperature is of the order e^2/h, independent of the sample size. The only requirement is that the transport takes place within the diffusive regime. For the derivation, an ensemble of different wires with different impurity configurations at zero magnetic field is considered. An explanation of the universality of conductance fluctuations follows the approach of Lee [49], although for the following considerations use will be made of the Landauer–Büttiker formalism (see Section 1.3.1), where conductance is expressed by the transmission probabilities of the different quantum channels. A scheme of the conductor is shown in Figure 1.18b, where the current is flowing from the left to the right reservoir. The disordered region of the wire is connected by ballistic areas to the phase-randomizing reservoirs.

The first step of the process is to express the Drude conductance of a diffusive conductor by the Landauer–Büttiker formula. Hence, the Drude conductance for a single spin direction can be written as

$$G = \frac{e^2}{h}\sum_{m,n}^{N} T_{mn} = \frac{e^2}{h}\left(N - \sum_{m,n}^{N} R_{mn}\right) \tag{1.47}$$

where $N = k_F W/\pi$ is the number of quantum channels of a 1-D conductor of width W, and k_F is the Fermi wavenumber. The quantities T_{mn} and R_{mn} denote the transmission and reflection probabilities from channel n into channel m, respectively [cf. Equation 1.23 and 1.20, the indices i, j for the channels are omitted, here]. Interest has been shown in the variations of the conductance for different impurity configurations, and the quantity related to this is the variance of the conductance, which is defined by:

$$\mathrm{var}(G) = \langle \Delta G^2 \rangle = \langle (G - \langle G \rangle)^2 \rangle \tag{1.48}$$

Here, $\langle \cdots \rangle$ denotes the average over different impurity configurations. The square root of the variance, $\delta G \equiv \sqrt{\mathrm{var}(G)}$, is a measure of the magnitude of the conductance

fluctuations, the quantity, which is determined experimentally. With the expression for the conductance, given by Equation 1.47, one obtains for the variance:

$$\begin{aligned} \mathrm{var}(G) &= \left(\frac{e^2}{h}\right)^2 \mathrm{var}\left(N - \sum_{m,n}^{N} R_{mn}\right) \\ &= \left(\frac{e^2}{h}\right)^2 \mathrm{var}\left(\sum_{m,n}^{N} R_{mn}\right) \\ &= \left(\frac{e^2}{h}\right)^2 N^2 \mathrm{var}(R_{mn}), \end{aligned} \tag{1.49}$$

and it is assumed that $\mathrm{var}(R_{mn})$ is independent of m and n. The question might be asked as to why the variance is not calculated directly by using the transmission probabilities T_{mn}. Following Lee [49], this causes a problem, since for transmission processes in the diffusive transport regime with many impurity collisions a sequence of scattering events is shared by different channels of the conductor. As a consequence, the different channels are not completely uncorrelated, so that problems are encountered in the proceeding averaging procedure. The situation is different, however, if reflection processes are considered, where it may be assumed that only a few scattering events are responsible for the back-reflection. This is also the reason, why it can be assumed that the reflections in different channels are uncorrelated.

In order to calculate the variance of R_{mn}, use is made of the concept of Feynman paths. By analogy to the probability to propagate between to points A and Q, as expressed by Equation 1.29, the probability for a reflection from channel n into m by the square of the total amplitude of all possible paths j which propagate from the incoming channel n into the outgoing channel m can be expressed as:

$$R_{mn} = \left|\sum_j C_j\right|^2 \tag{1.50}$$

According to Equation 1.49, the variance of R_{mn} must first be calculated in order to obtain the variance of G. The variance of R_{mn} is given by

$$\mathrm{var}(R_{mn}) = \langle R_{mn}^2 \rangle - \langle R_{mn} \rangle^2 \tag{1.51}$$

The last term is the square of the average reflection probability $\langle R_{mn} \rangle$, which can be expressed by

$$\langle R_{mn} \rangle = \sum_{ij} \langle C_i C_j^* \rangle \tag{1.52}$$

If uncorrelated paths are assumed, as discussed above, so that

$$\langle C_i C_j^* \rangle = 0 \tag{1.53}$$

one finally obtains

$$\langle R_{mn} \rangle = \sum_i \langle C_i C_i^* \rangle = \sum_i \langle |C_i|^2 \rangle \tag{1.54}$$

For the first term in Equation 1.51, the following can be written:

$$\begin{aligned}\langle R_{mn}^2\rangle &= \sum_{ijkl}\langle C_i C_j C_k^* C_l^*\rangle \\ &= \sum_{ijkl}\{\langle|C_i|^2\rangle\langle|C_j|^2\rangle\delta_{ik}\delta_{jl} + \langle|C_i|^2\rangle\langle|C_j|^2\rangle\delta_{il}\delta_{jk}\} \\ &= 2\sum_{ij}\langle|C_i|^2\rangle\langle|C_j|^2\rangle = 2\langle R_{mn}\rangle^2.\end{aligned} \tag{1.55}$$

Thus, the variance of R_{mn} is given simply by $\langle R_{mn}\rangle^2$.

For small transmission probabilities $T_{mn} \to 0$, which is the case for a sufficient number of scattering centers in the wire ($l_e \ll L$), the average reflection probability may be approximated by [49]

$$\langle R_{mn}\rangle \approx \frac{1}{N} \tag{1.56}$$

so that finally one obtains

$$\mathrm{var}(G) = \left(\frac{e^2}{h}\right)^2 N^2 \mathrm{var}(R_{mn}) \approx \left(\frac{e^2}{h}\right)^2 \tag{1.57}$$

for the variance of the conductance in the diffusive limit. As can be seen here, the conductance fluctuations are of the order of e^2/h. The universal magnitude of the conductance fluctuations is found for example in the measurement shown in Figure 1.17.

1.8.3 Fluctuations in Long Wires

Until now, it has been assumed that the wire length is smaller than the phase-coherence length, but the question must also be addressed as to what happens if the length L of the wire exceeds the phase-coherence length l_φ. In this situation, the wire may be cut into $N = L/l_\varphi$ phase-coherent pieces connected in series (see Figure 1.19).

Each of these segments produces resistance fluctuations, δR_0, so that the total resistance fluctuations are given by

$$\delta R = \sqrt{N}\delta R_0 \tag{1.58}$$

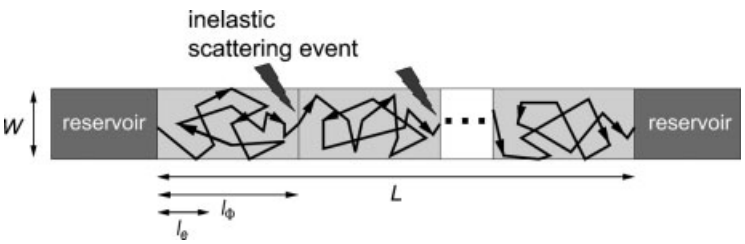


Figure 1.19 The conductance fluctuations are determined by cutting the wire into $N = L/l_\varphi$ coherent pieces.

By using the total resistance $R = NR_0$, where R_0 is the resistance of a single segment, the total conductance fluctuations can be calculated:

$$\delta G = \left| -\frac{\delta R}{R^2} \right| = \frac{e^2}{h}\frac{\sqrt{N}}{N^2} = \frac{e^2}{h} N^{-3/2} \tag{1.59}$$

If N is substituted by the ratio between total length and phase coherence length, the following is finally obtained:

$$\delta G = \frac{e^2}{h}\left(\frac{l_\varphi}{L}\right)^{3/2} \tag{1.60}$$

It is important to note that no exponential decrease of the fluctuations with respect to length is expected. In contrast, only a relatively weak decrease of δG with increasing length is predicted, and this was indeed confirmed experimentally [50].

1.8.4 Energy and Temperature Dependence

In semiconductor structures the electron concentration – and thus the Fermi energy – can be controlled by means of a gate electrode. An impression of how the Fermi energy affects the conductance fluctuations can be gained by comparing the change of E_F with the characteristic correlation energy E_{Th} (Thouless energy).

First, an insight must be obtained into the nature of the correlation energy E_{Th}. For the sake of simplicity, an ideal system can be assumed where any scattering is neglected. Along the length L, the phase develops as

$$\varphi = kL \tag{1.61}$$

If a state is now considered with a slightly different wavevector $k \to k' = k + \Delta k$, the phase difference between both waves is

$$\Delta\varphi = \Delta k L \tag{1.62}$$

The energy difference between both states can be quantified as

$$\Delta E = \frac{dE}{dk}\Delta k = \hbar v_F \frac{\Delta\varphi}{L} = \hbar\Delta\varphi\frac{1}{\tau_{Th}} \tag{1.63}$$

Here, τ_{Th} is the time that the wave requires to propagate along length L. The Thouless energy, E_{Th}, is defined as the energy difference where both states are uncorrelated, at which point if the phase difference $\Delta\varphi$ is equal to 1, the following is obtained:

$$E_{Th} = \frac{\hbar v_F}{L} = \frac{\hbar}{\tau_{Th}} \tag{1.64}$$

Thus, the Thouless energy is connected to the time $\tau_L = L/v_F$ that an electron wave requires to cover the dimension L of the system. Up to now, only a ballistic system has been considered, but in the diffusive regime the corresponding characteristic time is given by $\tau_{Th} = L^2/D$. Only if the Fermi energy is changed by a value comparable to E_{Th} is the next energy level reached and the conductance changed to a value uncorrelated

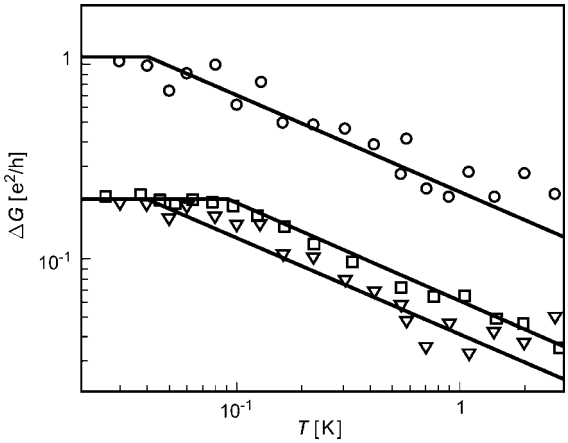


Figure 1.20 Decrease in conductance fluctuations with temperature for a ring structure with a diameter of 820 nm (○). In addition, the temperature dependence of the Aharonov–Bohm oscillations of a 820 nm-wide ring (▽) and for a 325 nm-wide ring (□) are shown [46]. (© Taylor &Francis Ltd.).

to the conductance value of the previous energy value. The correlation energy for small quantum wires has a relative large value, which would not be expected if a large number of uncorrelated trajectories were to be averaged.

If the temperature is increased, then the Fermi distribution becomes smeared out. However, while the temperature remains sufficiently low that the width of the Fermi distribution is smaller than E_{Th}, the maximum fluctuation amplitude is maintained. When the smearing of the Fermi distribution exceeds E_{Th}, a number of approximately $N = (k_B T)/E_{Th}$ segments contribute. As these N segments are uncorrelated, the fluctuation amplitude decreases with $1/\sqrt{N}$, thus $1/\sqrt{T}$. This behavior was confirmed experimentally, as shown in Figure 1.20 [46]. At low temperature the fluctuation amplitude is found initially to be constant, but if the temperature is increased above a critical value a continuous decrease following $1/\sqrt{T}$ is observed. From the starting point of the decrease at $T = 0.1$ K, the correlation energy can be estimated. Typically, a value for E_{Th} of the order of 10 μeV would be obtained for this sample.

1.9 Concluding Remarks

It is clear that many interesting phenomena related to phase-coherence transport can be observed in semiconducting or metallic nanostructures. Although, many of these effects are quite well understood and the theoretical models well established, in other cases open questions remain. For example, does the phase coherence time always saturate if the temperature is sufficiently low [2]? Very recently, the issue of phase

coherence in nanostructures has attracted much attention in connection with solid-state quantum computation, where the maintenance of phase coherence is a critical issue. Furthermore, spin-related phenomena are a subject of current interest, as phase-coherent spin manipulation is regarded as an interesting option for future electronic devices.

Clearly, this chapter can provide only a brief overview of the most important phenomena connected with phase-coherent transport in nanostructures. However, for further information, the reader is referred to various textbooks [1, 51, 52] and reviews [2, 18, 46].

References

1 Datta, S. (1995) *Electron Transport in Mesoscopic Systems*, Cambridge University Press, Cambridge.

2 Lin, J.J. and Bird, J.P. (2002) *Journal of Physics: Condensed Matter*, **14**, R501.

3 Giuliani, G.F. and Quinn, J.J. (1982) *Physical Review B*, **26**, 4421.

4 Al'tshuler, B.L., Aronov, A.G. and, Khmelnitsky, D.E. (1981) *Solid State Communications*, **39**, 619.

5 Choi, K.K., Tsui, D.C. and Alavi, K. (1987) *Physical Review B*, **36**, 7751.

6 Landauer, R. (1957) *IBM Journal of Research and Development*, **21**, 223.

7 Landauer, R. (1987) *Zeitschrift für Physik B*, **68**, 217.

8 Büttiker, M., Imry, Y., Landauer, R. and Pinhas, S. (1985) *Physical Review B*, **31**, 6207.

9 Büttiker, M. (1986) *Physical Review Letters*, **57**, 1764.

10 van Wees, B.J., van Houten, H., Beenakker, C.W.J., Willamson, J.G., Kouwenhoven, L.P., van der Marel, D. and Foxon, C.T. (1988) *Physical Review Letters*, **60**, 848.

11 Wharam, D.A., Thornton, T.J., Newbury, R., Pepper, M., Ahmed, H., Frost, J.E.F., Hasko, D.G., Peacock, D.C., Ritchie, D.A. and Jones, G.A.C. (1988) *Journal of Physics C: Solid State Physics*, **21**, L209.

12 Laux, S.E., Frank, D.J. and Stern, F. (1988) *Surface Science*, **196**, 101.

13 Szafer, A. and Stone, A.D. (1989) *Physical Review Letters*, **62**, 300.

14 Kirczenov, G. (1989) *Physical Review B*, **39**, 10452.

15 Abrahams, E., Anderson, P.W., Licciardello, D.C. and Ramakrishnan, T.V. (1979) *Physical Review Letters*, **42**, 673.

16 (a) Gorkov, L.P., Larkin, A.I. and Khmel'nitskii, D.E. (1979) *Pis'ma Zhurnal Eksperimentalnoi I Teoreticheskoi Fiziki*, **30**, 248; (b) Gorkov, L.P., Larkin, A.I. and Khmel'nitskii, D.E. (1979) *JETP Letters*, **30**, 228.

17 Feynman, R.P. and Hibbs, A.R. (1965) *Quantum Mechanics and Path Intergrals*, McGraw-Hill, New York.

18 Beenakker, C.W.J. and van Houten, H. (1991) (eds. H. Ehrenreich and D. Turnbull), *Semiconductor Heterostructures and Nanostructures in Solid State Physics*, Volume **44**, Academic Press, New York. p. 1.

19 Chakravarty and Schmid, A. (1986) *Physics Reports*, **140**, 193.

20 Beenakker, C.W.J. and van Houten, H. (1988) *Physical Review B*, **38**, 3232.

21 Al'tshuler, B.L., Khmelnitskii, D., Larkin, A.I. and Lee, P.A. (1980) *Physical Review B*, **22**, 5142.

22 Hikami, S., Larkin, A.I. and Nagoka, Y. (1980) *Progress of Theoretical Physics*, **63**, 707.

23 Choi, K.K., Tsui, D.C. and Alavi, K. (1987) *Physical Review B*, **36**, 7751.

24 (a) Al'tshuler, B.L. and Aronov, A.G. (1981) *Pis'ma Zhurnal Eksperimentalnoi I Teoreticheskoi Fiziki*, **33**, 515; (b) Al'tshuler,

B.L. and Aronov, A.G. (1981) *JETP Letters*, **33**, 499.

25 Dresselhaus, G. (1955) *Physical Review*, **100**, 580.

26 Bychkov, Yu.A. and Rashba, E.I. (1984) *Journal of Physical Chemistry (Solid State Physics)*, **17**, 6039.

27 Bergmann, G. (1982) *Solid State Communications*, **42**, 815.

28 Iordanskii, S.V., Lyanda-Geller, Yu.B. and Pikus, G.E. (1994) *JETP Letters*, **60**, 206.

29 Guzenko, V.A., Schäpers, Th., Indlekofer, K.M. and Knobbe, J. (2006) *Physica E*, **32**, 333.

30 (a) Al'tshuler, B.L., Aronov, A.G. and Spivak, B.Z. (1981) *Pis'ma Zhurnal Eksperimentalnoi I Teoreticheskoi Fiziki*, **33**, 101; (b) Al'tshuler, B.L., Aronov, A.G. and Spivak, B.Z. (1981) *JETP Letters*, **33**, 94.

31 Aronov, A.G. and Sharvin, Yu.V. (1987) *Reviews of Modern Physics*, **59**, 755.

32 (a) Sharvin, Yu.D. and Sharvin, Yu.V. (1981) *Pis'ma Zhurnal Eksperimentalnoi I Teoreticheskoi Fiziki*, **34**, 285; (b) Sharvin, D.Yu. and Sharvin, Yu.V. (1981) *JETP Letters*, **34**, 272.

33 Dolan, G.J., Licini, J.C. and Bishop, D.J. (1986) *Physical Review Letters*, **56**, 1493.

34 Aharonov, Y. and Bohm, D. (1959) *Physical Review*, **115**, 485.

35 Krafft, B., Förster, A., van der Hart, A. and Schäpers, Th. (2001) *Physica E*, **9**, 635.

36 Chambers, R.G. (1960) *Physical Review Letters*, **5**, 3.

37 Tonomura, A., Matsuda, T., Suzuki, R., Fukuhara, A., Osakabe, N., Umezaki, H., Endo, J., Shinagawa, K., Sugita, Y. and Fujiwara, H. (1982) *Physical Review Letters*, **48**, 1443.

38 Webb, R.A., Washburn, S., Umbach, C.P. and Laibovitz, R.B. (1985) *Physical Review Letters*, **54**, 2696.

39 Appenzeller, J., Schäpers, Th., Hardtdegen, H., Lengeler, B. and Lüth, H. (1995) *Physical Review B*, **51**, 4336.

40 Murat, M., Gefen, Y. and Imry, Y. (1984) *Physical Review B*, **34**, 659.

41 Büttiker, M. (1986) *Physical Review Letters*, **57**, 1761.

42 Umbach, C.P., Washburn, S., Laibowitz, R.B. and Webb, R.A. (1984) *Physical Review B*, **30**, 4048.

43 Stone, A.D. (1985) *Physical Review Letters*, **54**, 2692.

44 Kaplan, S.B. and Hartstein, A. (1986) *Physical Review Letters*, **56**, 2403.

45 Licini, J.C., Bishop, D.J., Kastner, M.A. and Melngailis, J. (1985) *Physical Review Letters*, **55**, 2987.

46 Washburn, S. and Webb, R.A. (1986) *Advances in Physics*, **35**, 375 (http://www.informaworld.com).

47 (a) Al'tshuler, B.L. (1985) *Pis'ma Zhurnal Eksperimentalnoi I Teoreticheskoi Fiziki*, **41**, 530; (b) Al'tshuler, B.L. (1985) *JETP Letters*, **41**, 648.

48 Lee, P.A. and Stone, A.D. (1985) *Physical Review Letters*, **55**, 1622.

49 Lee, P.A. (1986) *Physica A*, **140A**, 169.

50 Umbach, C.P., van Haesendonk, C., Laibowitz, R.B., Washburn, S. and Webb, R.A. (1986) *Physical Review Letters*, **56**, 386.

51 Ferry, D.K. and Goodnick, S.M. (2005) *Transport in Nanostructures*, Cambridge University Press, Cambridge.

52 Heinzel, T. (2003) *Mesoscopic Electronics in Solid State Nanostructures*, Wiley-VCH, Weinheim.

2
Charge Transport and Single-Electron Effects in Nanoscale Systems

Joseph M. Thijssen and Herre S.J. van der Zant

2.1
Introduction: Three-Terminal Devices and Quantization

In electronics, charges are manipulated by sending them through devices which have a few terminals: a *source* which injects the charge; and a *drain* which removes the charge from the device. Occasionally, a third terminal, called a *gate*, is present, and this is used to manipulate the charge flow through the device. The gate does neither inject charge into, nor removes it from the device. Three-terminal devices are standard elements of electronic circuits, where they act as switches or as amplifying elements. Semiconductor-based three-terminal switches are responsible for the tremendous increase in computer speed achieved over the past few decades.

Feynman, in his famous lecture [1], pointed out that the possible scale reduction from the standards of that period was still enormous, and he also suggested that quantum mechanical behavior may result in a different way of operation of the devices, which may open new horizons for applications. Indeed, as we now know, two aspects become important when the size of the device is reduced. The first aspect is indeed the quantum mechanical behavior, and the second is the quantization of the charges flowing into and out of the devices. It is interesting to analyze how the energy scales at which the two effects become noticeable, depend on the device size.

The charge quantization is subtle in view of quantum mechanics: in principle, the charge carried by an electron is distributed in space. In quantum mechanics, a single charge may be distributed according to $|\psi(r)|^2$, where $\psi(r)$ is the quantum mechanical wave function, and this leaves open the possibility of having a fractional charge inside the device. Therefore, the discrete nature of charge does not seem to play a role in the charge transport. However, if the device were to be uncoupled from its surroundings, we would only find integer charges residing on it. This puzzle is solved by realizing that the expectation value of the electrostatic energy, which must be included into the Hamiltonian governing the electron behavior, is dominated by the charge distribution which occurs most of the time. It can be shown that the charge within a device that is *weakly* coupled to its surroundings, is always very close to an integer. Therefore,

Nanotechnology. Volume 3: Information Technology I. Edited by Rainer Waser

ISBN: 978-3-527-31738-7

in order to observe Coulomb effects resulting from the discreteness of the electron charge, it is necessary to consider devices that are weakly coupled to the surroundings.

For the charge quantization, the energy scale associated with the discreteness of the electron charge is given by

$$E_C = \frac{e^2}{2C},$$

where C is the capacitance of the device. This is the energy needed to add a unit charge to the device – it is called the "charging energy". Taking as an estimate the capacitance of a sphere with radius R, we have

$$E_C = \frac{e^2}{8\pi\varepsilon_0 R} = \frac{1}{2R} E_H, \tag{2.1}$$

where, in the rightmost expression, R is given in atomic units (Bohr radii), as is the energy (E_H is the atomic unit of energy – it is called the Hartree and it is given by 27.212 eV). In Section 2.4, we shall present a more detailed analysis for the case where the device is (weakly) coupled to a source, drain and gate.

The energy scale for quantum effects is given by the distance between the energy levels of an isolated device. As a rough estimate, we consider the particle in the (cubic) box problem with energy levels separated by a level splitting Δ given by

$$\Delta = \text{const} \times \frac{\hbar^2}{mL^2} = \text{const} \times \frac{1}{L^2} E_H, \tag{2.2}$$

where m is the electron mass and L is the box size (which must be given in atomic units in the rightmost expression). The first multiplicative constant is of order 1; it depends on the geometry and on the details of the potential.

In the case of carbon nanotubes, the device is much smaller in the lateral direction than along the tube axis. In such cases it is useful to distinguish between the two sizes. The lateral size leads to a large energy splitting and the longitudinal splitting may become vanishingly small. For a metallic nanotube, the level spacing associated with the tube length L is

$$\Delta = \frac{\hbar v_F}{2L},$$

where v_F is the Fermi velocity $v_F = \hbar k_F/m$ with $\nu_F \approx 8 \times 10^5\,\mathrm{m\,s^{-1}}$.

Equations 2.1 and 2.2 tell us how the typical Coulomb and quantum energies scale with the device size (R or L). In Figure 2.1, several experimental realizations are shown of small gated devices that may be weakly coupled to source, drain. Most of these devices have the layout shown in Figure 2.2. An order of magnitude estimate for the charging energy and level splitting for some typical three-terminal devices is provided in Table 2.1. Semiconducting and nanotube quantum dots have been studied in great detail, and their behavior is fairly well understood; however, at the time of writing, the properties of molecular quantum dots are still much less established mainly because it is difficult to fabricate them in a reliable manner.

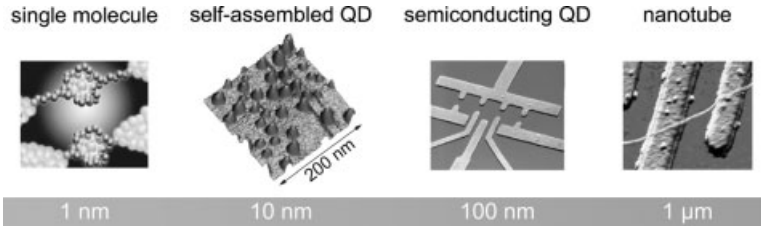


Figure 2.1 Different quantum dot systems.

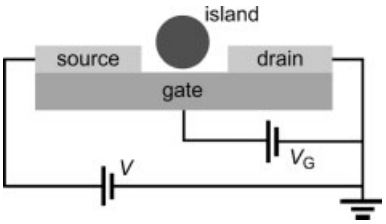


Figure 2.2 A schematic diagram of the three-terminal device layout.

When studying transport through a small island, weakly coupled to a source and a drain, information can be obtained about the quantum level splitting Δ and the charging energy E_C if the energy of the particles flowing through the device can be controlled with precision high enough to resolve these energy splittings. Pauli's principle states that electrons can only flow from an occupied state in the source to an empty state in the drain. The separation between empty and occupied states in the leads is only sharp enough when the temperature is sufficiently low. It can be seen that a low operation temperature is essential for observing the quantum and charge quantization effects. The energy scale associated with the temperature is given by k_BT, so we must have

$$k_BT \leq \Delta, E_C.$$

Note that for molecular devices, with their relatively large values of Δ and E_C, quantum and charge quantization effects should still be observable at room temperature. In a typical metallic island, $\Delta \ll k_BT$, and the Coulomb blockade dominates the level separation. In this case we speak of a *classical dot* (see also Chapter 21 of this volume).

In the present chapter we explain the different aspects of charge transport, with emphasis on those devices in which the level spacing and the charging energy plays an essential role in the transport properties. This is the case in quantum dots and in many molecular devices.

Table 2.1 Typical charging energies and level spacings for various three-terminal devices.

	Ga As quantum dot	Carbon nanotube[a]	Molecular transistor
E_C	0.2–2 meV	3 meV	>0.1 eV
Δ	0.02–0.2 meV	3 meV	>0.1 eV

[a]Metallic nanotube, 500 nm in length.

2.2 Description of Transport

In this section, we present a qualitative discussion of the different transport mechanisms, after which attention will be focused on the weak-coupling case.

The major question here is what picture should be used to describe transport through small devices. In solids, electrons are usually thought of in terms of the independent particle model, in which the wave function of the many-electron system is written in the form of a Slater determinant built from one-electron orbitals. This is an exact solution for a Hamiltonian, which is a sum of one-electron Hamiltonians:

$$H = \sum_i h_i. \tag{2.3}$$

The electrostatic repulsion between the electrons:

$$V_{ES} = \frac{1}{4}\pi\varepsilon_0 \frac{e^2}{|\vec{r_i} - \vec{r_j}|}.$$

does not satisfy this requirement. Also, the electrons couple electrostatically to the motion of the nuclei, which interact among themselves via a similar Coulomb interaction. Several schemes exist for building a Hamiltonian, such as Equation 2.3, in which the interaction between the electrons is somehow moved into a (possibly non-local) average electrostatic potential. The best known such schemes are the Hartree–Fock (HF) and the density functional theory (DFT). The question is now whether the independent electron picture can survive in the study of transport through small devices. The answer is that single-electron orbitals still form a useful basis for understanding this transport, but that the Coulomb and electron–nucleus interaction must be included quite explicitly into the description in order to understand single-electron effects.

2.2.1 Structure of Nanoscale Devices

Although it often cannot be used in the transport itself, the single particle picture is still suitable for the bulk-like systems to which the device is coupled, and for the narrow leads which may be present between the island and the bulk reservoirs. These elements are described in Chapter 1, and their properties will be recalled only briefly here, with emphasis on the issues needed in the context of the present chapter.

2.2.1.1 The Reservoirs

The reservoirs are bulk-like regions where the electrons are in equilibrium. These regions are maintained at a specified temperature, and the number of electrons is variable as they are connected to the voltage source and the leads to the device (see below). The electrons in these reservoirs are therefore distributed according to Fermi functions with a given temperature T and a chemical potential μ:

$$f_{FD}(E) = \frac{1}{\exp[(E-\mu)/k_BT]+1}.$$

This function falls off from 1 at low energy to 0 at high energy. For $(\mu - E_0) \ll k_BT$, where E_0 is the ground state energy, this reduces to a sharp step down from 1 to 0 at $E = \mu$, and μ can in that case be identified with the Fermi energy (the highest occupied single-particle energy level).

In order to have a current running through the device and the leads, the source and drain reservoirs are connected to a voltage source. A bias voltage causes the two leads to have different chemical potentials.

2.2.1.2 The Leads

Sometimes it is useful to consider the leads as a separate part of the system, in particular for convenience of the theoretical analysis. The leads are channels, which may be considered to be homogeneous. They form the connection between the reservoirs and the island (see below). They are quite narrow and relatively long. Electrons in the leads can still be described by single-particle orbitals. If the leads have a discrete or continuous translational symmetry, the states inside them are Bloch waves. By separation of variables, we can write the states as

$$e^{ik_z z} u_T(x, y)$$

with energy

$$E = E_T + \frac{\hbar^2 k_z^2}{2m}.$$

It is seen that the states can be written as a transverse state $u_T(x, y)$ which contributes an amount E_T to the total energy, times a plane wave along z. The quantum numbers of the transverse wave function $u_T(x, y)$ are used to identify a channel.

In this chapter, usually no distinction is made between reservoirs and leads: rather, they are both simply described as baths in equilibrium with a particular temperature and chemical potential (which may be different for the source and drain lead). However, for a theoretical description of transport, it is often convenient to study the scattering of the incoming states into outgoing states – in that case, the simple and well-defined states of the leads facilitate the description.

2.2.1.3 The Island

This is the part of the system which is small in all directions (although in a nanotube, the transverse dimensions are much smaller than the longitudinal); hence, this is the part where the Coulomb interaction plays an important role. To understand the device, it is useful to take as a reference the isolated island. In that case we have a set of quantum states with discrete energies (levels). The density of states of the device consists of a series of delta-functions corresponding to the bound state energies.

Now imagine there is a knob by which we can tune the coupling to the leads. This is given in terms of the rate $\Gamma/\hbar$ at which electrons cross the tunnel barriers separating the island from the leads. The transport through the barriers is a tunneling process which is fast and, in most cases, it can be considered as elastic: the energy is conserved in the tunneling process. Generally speaking, when the island is coupled to the leads (or directly to the reservoirs), the level broadens as a result of the continuous density of states in the leads (or reservoirs), and it may shift due to charge transfer from the leads to the island. Two limits can be considered. For weak coupling, $\Gamma \ll E_C$, Δ, the density of states should be close to that of the isolated device: it consists of a series of peaks, the width of which is proportional to Γ. Sometimes, we wish to distinguish between the coupling to the source and drain lead, and use Γ_S and Γ_D, respectively. For strong coupling, that is, $\Gamma \gg E_C$, Δ, the density of states is strongly influenced by that of the leads, and the structure of the spectrum of the island device is much more difficult to recognize in the density of states of the coupled island.

If we keep the number of electrons within the island fixed, we still have the freedom of distributing the electrons over the energy spectrum. The only constraint is the fact that not more than one electron can occupy a quantum state as a consequence of Pauli's principle. The change in total energy of the device is then mainly determined by the level splitting which is characterized by the energy scale, Δ. If we wish to *add* or *remove* an electron to or from the device, we must pay or we gain in addition a charging energy respectively.

It should be noted that, in principle, Γ may depend on the particular charge state on the island. This is expected to be the case in molecules: the charge distribution usually strongly differs for the different orbitals and this will certainly influence the degree in which that orbital couples to the lead states.

At this stage, an important point should be emphasized. From statistical mechanics, it is known that a particle current is driven by a chemical potential difference. Therefore, the chemical potential of the island is the relevant quantity driving the current to and from the leads. However, in an independent particle picture, a single particle energy is identical to the chemical potential (which is defined as the difference in *total* energy between a system with $N+1$ and N particles). Therefore, if we speak of a single-particle energy of the island, this should often be read as "chemical potential".

2.2.2 Transport

For an extensive discussion of the issues discussed in this paragraph, the reader is referred to the monograph by Datta [2].

As seen above, in the device we can often distinguish discrete states as (Lorentzian) peaks with finite width in the density of states. A convenient representation of transport is then given in Figure 2.3, which shows that the effect of the gate is to shift the levels of the device up and down, while leaving the chemical potentials μ_S and μ_D of the leads unchanged (for small devices, the gate field is inhomogeneous due to the

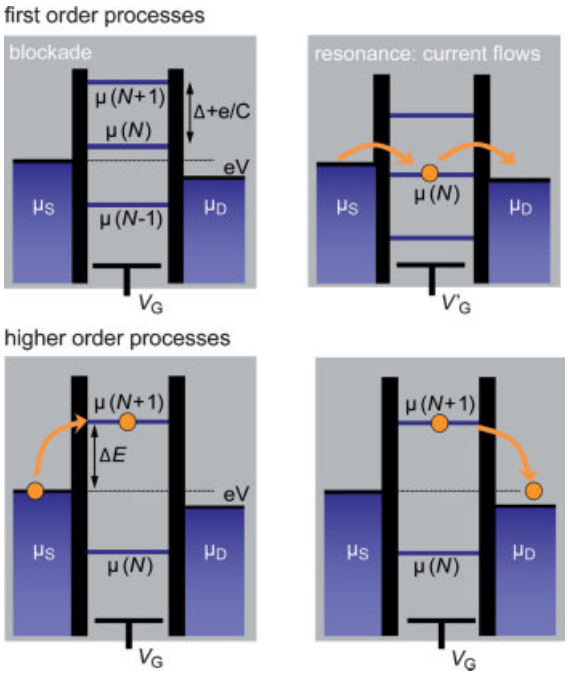


Figure 2.3 Schematic representation of the electrochemical potentials of an island connected to two reservoirs, across which a small (negative) bias voltage V is applied. A voltage on the gate electrode can be used to shift the electrostatic potential of the energy level. Top: Resonant transport becomes possible when the gate voltage pushes one of the levels within the bias window eV. The $\mu(N)$ level is aligned with μ_S and the number of electrons on the dot alternates between N and $N-1$ (sequential tunneling). Bottom: The levels are not aligned. The Coulomb blockade fixes the number of electrons on the dot to N. Transport, however, is possible through a virtual co-tunnel process in which an unoccupied level is briefly occupied. A similar process exists for the occupied level, $\mu(N)$, which may be briefly unoccupied. In contrast to resonant transport, the level is empty (full) most of the time. For all panels it should be noted that, in reality, the levels are not sharp lines but rather have a finite width, Γ. Similarly, the edge between the occupied (blue) and unoccupied states is blurred by temperature via the Fermi–Dirac function.

effect of the leads; moreover, the electrostatic potential in the surface region of the leads will be slightly affected by the gate voltage).

The transport through the device can take place in many different ways. A few classifications will now be provided which may help in understanding the transport characteristics of a particular transport process.

2.2.2.1 Coherent-Incoherent Transport

First, the transport may be either *coherent* or *incoherent*, a notion which pertains to an independent particle description of the electrons where the electrons occupy one-particle orbitals. In the case of coherent transport, the phase of the orbitals evolves deterministically. In the case of incoherent processes, the phase changes in an unpredictable manner due to interactions which are not contained in the indepen-

dent particle Hamiltonian. Such interactions can be either electron–electron or electron–phonon interactions, or between the electrons and an electromagnetic field.

If the electrons spend a long time on the island – which occurs when the couplings to the leads are weak – then the decoherence will be complete. Only for short traversal times the phase will be well preserved.

2.2.2.2 Elastic–Inelastic Transport

Another distinction is that between elastic and inelastic transport. In the latter case, interactions may cause energy loss or gain of the electrons flowing through the device. This energy change may be caused by the same interactions as those causing decoherence (electron–electron, electron–phonon, electron–photon). It should be noted, however, that decoherent transport can still be elastic.

2.2.2.3 Resonant–Off-Resonant Transport

This classification is relevant for elastic tunneling in combination with weak coupling to the leads. In resonant transport, electrons are injected at an energy corresponding to a resonance of the island. Such a resonance corresponds to a discrete energy level of the isolated device. The transport resonance energy corresponds to the center of the shifted peak; this is seen as a peak in the transport current for that energy or, more specifically, an increase of the current as soon as a resonance enters the bias window. The fact that the coupling to the leads is weak causes the time that an electron resides in the device to be rather long. If this time is longer than the time taken for the electron orbital to lose its coherence, we speak of *sequential tunneling*, as the transport process may then be viewed as electrons hopping from the lead to the island where they stay a while before hopping off to the drain. In off-resonant transport, the electrons are injected at energies (far) off the resonance.

2.2.2.4 First-Order versus Higher-Order Processes

The standard technique for calculating the current arising from coherent processes is time-dependent perturbation theory. In this theory, the transition from one particular state to another is calculated in terms of transitions between the initial, intermediate and final states. The first-order process (Figure 2.3, top) corresponds to a direct transition from the initial to the final state and, for this process, the current is proportional to the couplings Γ between device and leads. In off-resonant first-order processes, the current decays rapidly with the energy difference between the closest discrete level on the island and the Fermi energies of the leads (ΔE in Figure 2.3).

Second-order transport processes, often called *co-tunneling*, take place via an intermediate state, as illustrated in the lower panel of Figure 2.3. In these processes, the current is proportional to higher powers of the couplings, but they are less strongly suppressed with increasing distance (in energy) between the states in the leads and on the island. Therefore, they may sometimes compete with – or even supersede – first-order processes, provided that the intermediate state is sufficiently far in energy (chemical potential) from those in the leads. Currents due to second-order processes vary quadratically with the coupling strengths.

Molecules can often be viewed as chains of weakly coupled sites. If the Fermi energy of the source lead (i.e., the injection energy) is at some distance below the on-site energies of the molecule, the dominant transport mechanism is through higher order processes, which in electron transfer theory are known as *superexchange* processes. This term also includes hole transport through levels below the Fermi energy of the leads.

2.2.2.5 **Direct Tunneling**

It should be noted that if the device is very small (e.g., a molecule), there is a possibility of having direct tunneling from the source to the drain, in which the resonant states of the device are not used for the transport.

2.3 Resonant Transport

We start this section by studying resonant transport qualitatively [2]. Suppose we have one or more sharp resonant levels which can be used in the transport process from source to drain. We neglect inelastic processes inside the device during tunneling from the leads to the device, or vice versa. In order to send an electron into the device at the resonant energy, we need occupied states in the source lead. This means that the density of states in that lead must be non-zero for the resonant energy (otherwise there is no lead state at that energy), and that the Fermi–Dirac distribution must allow for that energy level to be occupied. Furthermore, for the electron to end up in the drain, the states in the drain at the resonant energy should be empty according to Pauli's principle. We conclude that for the transport to be possible, the resonance should be *inside the bias-window*. This window is defined as the range of energies between the Fermi energies of the source and the drain.

The process is depicted in Figure 2.3 (top), and from this picture we can infer the behavior of the current as a function of the bias voltage. It can be seen that no current is possible (left panel) for a small bias voltage as a result of a finite difference in energy ΔE between the energy of the resonant state on the island and the nearest of the two chemical potentials of the leads. The current sets off as soon as the bias window encloses the resonance energy (right panel). Any further increase of the bias voltage does not change the current, until another resonance is included. The mechanism described here gives rise to current–voltage characteristics shown in Figure 2.4.

At this point, two remarks are in order. First, the image sketched here supposes weak coupling and a low temperature. Increasing the temperature blurs the sharp edge in the spectrum between the occupied and empty states, and this will cause the sharp steps seen in the I/V curve to become rounded. Second, the differential conductance, dI/dV as a function of the bias voltage V shows a peak at the positions where the current steps up.

In the previous section it was noted that the coupling $\Gamma = \Gamma_S + \Gamma_D$ between leads and device can be given in terms of the rate at which electrons hop from the lead onto the device. From this, an heuristic argument leads via the time–energy uncertainty

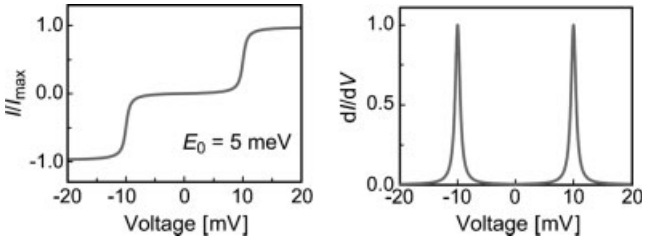


Figure 2.4 Left: Current–voltage characteristic calculated with Equation 2.6 for a level that is located 5 meV from the nearest Fermi energy of one of the electrodes. A symmetric coupling to the leads is assumed with a total broadening of 0.5 meV. Right: Corresponding differential conductance with a peak height equal to the conductance quantum. Note that the peak width is of the order of the total broadening.

relation to the conclusion that Γ gives us the extent to which an energy level[1)] E_0 on the island is broadened. Simple models for leads and device yield a Lorentzian density of states on the device [2]:

$$D(E) = \frac{1}{2\pi}\frac{\Gamma}{(E-E_0)^2 + (\Gamma/2)^2}.$$

Further analysis, which is based on a balance between ingoing and outgoing electrons [2] gives the following expression for the current:

$$I(E) = -\int \frac{e}{\hbar} D(E) \frac{\Gamma_S \Gamma_D}{\Gamma_S + \Gamma_D} \left[f_{FD}(E-\mu_S) - f_{FD}(E-\mu_D) \right] dE. \tag{2.4}$$

It must be remembered that the bias voltage (the potential difference between source and drain) is related to the chemical potentials μ_D and μ_S as

$$-eV = \mu_S - \mu_D;$$

where $e > 0$ is unit charge. A positive bias voltage drives the electrons from right to left, such that the current is then from left to right; this is defined as the positive direction of the current.

If the density of states has a single sharp peak, then current is only possible when this peak lies inside the bias window. Indeed, replacing $D(E)$ by a delta-function centered at E_0 directly gives

$$I = \frac{-e}{\hbar}\frac{\Gamma_S \Gamma_D}{\Gamma_S + \Gamma_D}\left[f_{FD}(E_0-\mu_S) - f_{FD}(E_0-\mu_D) \right].$$

At low temperature, the factor in square brackets is 1 when E_0 lies inside the bias window and 0 otherwise. It can be seen that the maximum value of the current is found as

1) Note that the energy should be identified with the chemical potential of the island; see the comment in the previous section.

$$|I_{\max}| = \frac{e}{\hbar}\frac{\Gamma_S\Gamma_D}{\Gamma_S+\Gamma_D}. \tag{2.5}$$

For low temperature, the Fermi functions in Equation 2.4 become sharp steps, and the integral of the Lorentzian can be carried out analytically, yielding

$$I = \frac{e}{\pi\hbar}\frac{\Gamma_S\Gamma_D}{\Gamma_S+\Gamma_D}\left[\arctan\left(2\frac{\mu_S-E_0}{\Gamma}\right)-\arctan\left(2\frac{\mu_D-E_0}{\Gamma}\right)\right]. \tag{2.6}$$

Equation 2.4 is valid in the limit where we can describe the transport in terms of the independent particle model. It has the form of the Landauer formula:

$$I = \frac{e}{\hbar}\int T(E)[f_{FD}(E-\mu_D)-f_{FD}(E-\mu_S)]dE,$$

which is discussed extensively in Chapter 1 of this volume. In that chapter it is shown that the transmission per channel (which corresponds to the eigenvalues of the matrix $T(E)$) has a maximum value of 1, so that the current assumes for low temperatures a maximum value of

$$I_{\max} = \frac{e^2}{\hbar}nV, \tag{2.7}$$

where n is the number of channels inside the bias window. It should be noted that this maximum occurs only for *reflectionless* contacts, for which a wave incident from the leads onto the device, is completely transmitted. This usually occurs when the device and the leads are made from the same material. The strong-coupling result in Equation 2.7 has been given in order to emphasize that the two Equations 2.5 and 2.7 hold in quite opposite regimes.

Often, in experiments the differential conductance dI/dV is measured, and this can be calculated from the expression in Equation 2.4:

$$\frac{dI}{dV} = -\frac{e^2}{\hbar}\frac{\Gamma_S\Gamma_D}{\Gamma_S+\Gamma_D} \int dE D(E)\{\eta f'_{FD}(E-\bar{\mu}+\eta eV)-(1-\eta)f'_{FD}[E-\bar{\mu}-(1-\eta)eV]\}, \tag{2.8}$$

where f'_{FD} denotes the first derivative of the Fermi–Dirac distribution with respect to its argument, and $\bar{\mu} = (\mu_S+\mu_D)/2$. The parameter η specifies how the bias voltage is distributed over the source and drain contact; for $\eta = 1/2$, this distribution is symmetric. For $T = 0$, the Fermi–Dirac distribution function reduces to a step function, and its derivative is then a delta-function. For low bias ($V \approx 0$), the integral picks up a contribution from both delta functions occurring in the integral in Equation 2.8. The result is

$$\frac{dI}{dV} = 4\frac{e^2}{\hbar}\frac{\Gamma_S\Gamma_D}{\Gamma_S+\Gamma_D}D(\mu),$$

where the energy E in Equation 2.8 is taken at the Fermi energy of either the source or the drain. As the maximum value of $D(E)$ is given as

$$D(E)_{\max} = \frac{2}{\pi}\frac{1}{\Gamma_S+\Gamma_D},$$

it follows that the maximum of the differential conductance occurs when $\Gamma_S = \Gamma_D$ and is then given by e^2/h. Note that this holds even when the current is much smaller than the quantum conductance limit [see Equation 2.7] which follows from the Landauer formula.

At *finite* temperature, for $k_BT \gg \Gamma$ and zero bias, working out the derivative with respect to bias of Equation 2.8 gives:

$$\frac{dI}{dV} = \frac{e^2\Gamma_S\Gamma_D}{4k_BT(\Gamma_S+\Gamma_D)}\left[\cosh\frac{e\alpha(V_G - V_0)}{2k_BT}\right]^{-2}. \tag{2.9}$$

This line shape (see Figure 2.5) is characterized by a maximum value $e^2\Gamma_S\Gamma_D/4k_BT(\Gamma_S+\Gamma_D)$, attained when the gate voltage reaches the resonance $V_0 = E_0/e$. The full-width half maximum (FWHM) of this peak is $3.525k_BT/e\alpha$. The parameter α is the gate coupling parameter: the potential on the island varies linearly with the gate voltage, $\Delta V_I = \alpha\Delta V_G$. These features are often used as a signature for true quantum resonant behavior as opposed to classical dots, where the small value of Δ renders the spectrum of levels accessible to an electron continuous. For a classical dot, the peak height is independent of temperature and the FWHM is predicted to increase by a factor 1.25 [3,4]. It should be noted that, in a quantum dot, Γ sets a lower bound for the temperature dependence of the peak shape: for $\Gamma > k_BT$ the peak height and shape are independent of temperature (not visible in Figure 2.5 due to the small value for Γ chosen there).

Interestingly, the finite width of the density of states, which is given by $\Gamma_S + \Gamma_D$, can in principle be measured experimentally from the resonance line widths at low temperature. It should be noted that the expressions for the current and differential conductance depend only on the combinations $\Gamma_S + \Gamma_D$ and $\Gamma_S\Gamma_D/(\Gamma_S + \Gamma_D)$. If both are extracted from experimental data, the values of Γ_S and Γ_D can be determined (although the symmetry between exchange of source and drain prevents us from identifying which value belongs to the source).

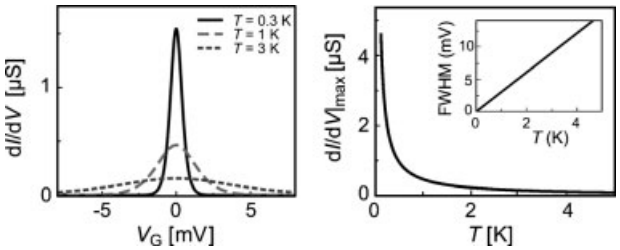


Figure 2.5 Left: Temperature-dependence of the Coulomb peak height [Equation 2.9] in the resonant transport model, showing the characteristic increase as the temperature is lowered. Right: Peak height as a function of temperature. The inset shows the full-width half maximum (FWHM) of the Coulomb peak as a function of temperature (see text). Calculations are performed with $\Gamma = 10^9\,s^{-1}$ and a gate coupling of 0.1 in the regime $\Gamma < k_BT$.

2.4
Constant Interaction Model

In Section 2.1.1 it was seen that in the weak-coupling regime, energy levels can be discrete for two reasons: quantum confinement (the fact that the state must "fit" into a small island), and charge quantization effects. The scale for the second type of splitting is the charging or Coulomb energy, E_C. It is important to realize that this energy will only be noticeable when the coupling to the leads is small in comparison with E_C; this situation is referred to as the *Coulomb blockade* regime. In this situation, a clear distinction should be made a between one or two electrons occupying a level, as their Coulomb interaction contributes significantly to the total energy. The transport process may be analyzed using the so-called *constant interaction model* [3], which is based on the set-up shown in Figure 2.6. Elementary electrostatics provides the following relation between the different potentials and the charge Q on the island:

$$CV_I - C_S V_S - C_D V_D - C_G V_G = Q,$$

where $C = C_S + C_D + C_G$. Note that this equation can be written in the form:

$$V_I = V_{ext} + \frac{Q}{C},$$

with

$$V_{ext} = (C_S V_S + C_D V_D + C_G V_G)/C.$$

It is seen that the potential on the dot is determined by the charge residing on it and by the induced potential V_{ext} of the source, drain and gate.

We take as a reference configuration the one for which all voltages and the charge are zero. For total energies, U rather than E is used in order to avoid confusion with the single-particle energies E_n resulting from solving the single-particle Schrödinger equation. The electrostatic energy $U_{ES}(N)$ with respect to this reference configuration after changing the source, drain and gate potentials and putting N electrons (of charge $-e$) on the island is then identified as the work needed to place this extra charge on the island, and the energy cost involved in changing the external potential when a charge Q is present:

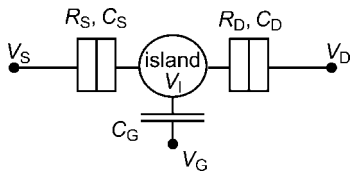


Figure 2.6 The capacitance model. A schematic drawing of an island connected to source and drain electrodes with tunnel junctions; the gate electrode shifts the electrostatic potential of the island.

$$U_{ES}(N) = \int_{Q=0,V_{ext}=0}^{-Ne,V_{ext}} (V_I dQ + Q dV_{ext}) = \frac{(Ne)^2}{2C} - NeV_{ext}.$$

The integral is over a path in Q, V_{ext} space; it is independent of the path – that is, of how the charge and external potential are changed in time.

The result for the total energy, including the "quantum energy" due to the orbital energies is

$$U(N) = \frac{(Ne)^2}{2C} - NeV_{ext} + \sum_{n=1}^{N} E_n.$$

The energy levels E_n correspond to states which can be occupied by the electrons in the device, provided that their total number does not change, as changing this number would change the Coulomb energy, which is accounted for by the first term. This expression for the total energy is essentially the constant interaction model.

From non-equilibrium thermodynamics, it is known that a current is driven by a chemical potential difference; hence, we should compare the chemical potential on the device,

$$\mu(N) = U(N) - U(N-1) = (N - 1/2)\frac{e^2}{C} - eV_{ext} + E_N, \tag{2.10}$$

with that of the source and drain in order to see whether a current is flowing through the device. From the definition of V_{ext} we see that the effective change in the chemical potential due to a change of the gate voltage (while keeping source and drain voltage constant), carries a factor C_G/C; this is precisely the gate coupling, which is called the α factor (this was referred to at the end of Section 2.3).

It is important to be aware of the conditions for which the constant interaction model provides a reliable description of the device. The first condition is weak coupling to the leads; the second condition is that the size of the device should be sufficiently large to make a description with single values for the capacitances possible. Finally, the single-particle levels E_n must be independent of the charge N. The constant-interaction model works well for weakly coupled quantum dots, for which it is very often used. However, for molecular devices the presence of a source and drain both of which are large chunks of conducting material separated by very narrow gaps, reduces the gate field to be barely noticeable close to the leads and far from the gate. This inhomogeneity of the gate field may lead to a dependence of the gate capacitance C_G with N due to the difference in structure of subsequent molecular orbitals, and the chemical potential on the molecule will vary non-linearly with the gate potential.

As will be seen below, the distance between the different chemical potential levels can be inferred from three-terminal measurements of the (differential) conductance. From Equation 2.10, this distance is given by

$$\mu(N+1) - \mu(N) = \frac{e^2}{C} + E_{N+1} - E_N.$$

It should be noted that the difference in energy levels occurring in this expression ($E_{N+1} - \mathrm{E}_N$) is nothing but the splitting Δ mentioned at the very start of this chapter. However, for typical metallic and semiconductor quantum dots, this splitting is usually significantly smaller than the charging energy, so that this quantity determines the distance between the energy levels:

$$\mu(N+1) - \mu(N) = \frac{e^2}{C}.$$

Note that this *addition energy* is twice the energy of a charge on the dot (as the addition energy is the second derivative of the energy with respect to the charge).

Now, the current can be studied as a function of bias and gate voltage. In Section 2.2.2 it was seen that, in the weak coupling regime and at low temperature, the current is suppressed when all chemical potential levels lie outside of the bias window. As the location of these levels can be tuned by using the gate voltage, it is interesting to study the current and differential conductance of the device as a function of the bias *and* of the gate voltage.

We can calculate the line in the V, V_G plane which separates a region of suppressed current from a region with finite current; this line is determined by the condition that the chemical potential of the source (or drain) is aligned with that of a level on the island. Again, it is assumed that the drain is grounded (as in Figure 2.2). From the expression in Equation 2.10 for the chemical potential, and using the definition for V_{ext}, we find the following condition for the chemical potential to be aligned to the source (keeping the dot's charge constant):

$$V = \beta(V_G - V_C),$$

where $\beta = C_G/(C_G + C_D)$ and $V_C = (N - 1/2)\frac{e}{C_G} + \frac{C}{C_G}\frac{E_N}{e}$; that is, the voltage corresponding to the chemical potential on the dot in the absence of an external potential. If the chemical potential is aligned with the drain, we have

$$V = \gamma(V_C - V_G)$$

with $\gamma = C_G/C_S$. The expressions given here are specific for a grounded drain electrode. It is easily verified that, irrespective of the grounding, it holds that

$$\frac{C}{C_G} = \frac{1}{\alpha} = \frac{1}{\beta} + \frac{1}{\gamma}.$$

Each resonance generates two straight lines separating regions of suppressed current from those with finite current. For a sequence of resonances, the arrangement shown in Figure 2.7a is obtained. The diamond-shaped regions are traditionally called "Coulomb diamonds", as they were very often studied in the context of metallic dots, where the chemical potential difference of the levels is mainly made up of the Coulomb energy. The name is also used in molecular transport, although this is – strictly speaking – not justified there as Δ may be of the same order as the Coulomb interaction.

From the Coulomb diamond picture we can infer the values of some important quantities. First, we consider two successive states on the molecule with chemical potentials $\Delta\mu_{(N)}$ and $\Delta\mu_{(N+1)}$. If we suppose that both states have the

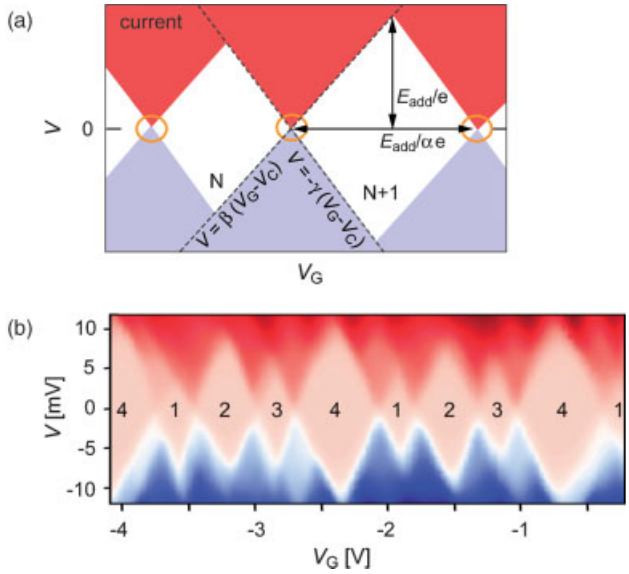


Figure 2.7 Linear transport. (a) Two-dimensional plot of the current as a function of bias and gate voltage (stability diagram). For a small bias, current flows only in the three points corresponding to the situation shown in Figure 2.3 (top right). These points are known as the "degeneracy points". Red = positive currents; blue = negative currents; white = blockade, no current. (b) Measured stability diagram of a metallic, single-walled carbon nanotube, showing the expected fourfold shell filling. The blockade regime is shown in pink. (Data from Ref. [5].).

same gate-coupling parameter α, it can be seen that the upper and lower vertices of the diamond are both at a distance

$$\Delta V = \frac{|\mu(N) - \mu(N+1)|}{e} = \frac{E_{add}}{e}$$

from the zero-bias line. This difference in chemical potentials is the electron addition energy, E_{add}. If the addition energy is dominated by the charging energy, then the total capacitance can be determined. Combining this with the slopes of the sides of the diamond, which provide the relative values of C_G, C_S and C_D, all of these capacitances we can be determined explicitly.

One interesting consequence of the previous analysis is that, if the capacitances do not depend on the particular state being examined, then the height of successive Coulomb diamonds is constant. If, in addition to the Coulomb energy, a level splitting is present, this homogeneity will be destroyed, as can be seen in Figure 2.7b, which shows the diamonds for a carbon nanotube (CNT) [5]. The alternation of a large diamond with three smaller ones can be nicely explained with a model Hamiltonian [6]. In the case of transport through molecules there is no obvious underlying structure in the diamonds.

The electron addition energy is sometimes connected to the so-called HOMO–LUMO gap. [These acronyms represent the Highest Occupied (Lowest Unoccupied)

Molecular Orbital, and denote orbitals within an independent particle scheme.] If the Coulomb interaction is significant, the HOMO–LUMO gap can be related to the excitation energy for an optical absorption process in which an electron is promoted from the ground state to the first excited state, without leaving the system. In that case, the change in Coulomb energy is modest, and the energy difference is mostly made up of the quantum splitting Δ. It should be noted, however, that the HOMO and LUMO are usually calculated using a computational scheme whereby the orbitals are calculated for the ground-state configuration – that is, without explicitly taking into account the fact that all orbitals change when, for example, an electron is excited to a higher level.

The addition energies are partly determined by quantum confinement effects and partly by Coulomb effects. A difficulty here is that these energies will be different for a molecular junction, in which a molecule is either physisorbed or chemisorbed to conducting leads, than for a molecule in the gas phase. There are several effects responsible for this difference. First, if there is a chemical bond present, then the electronic orbitals extend over a larger space, which reduces the confinement splitting. Second, a chemical bond may cause a charge transfer from lead to molecule, which in turn causes the potential on the molecule to change. Third, the charge distribution on the molecule will polarize the surface charge on the leads, which can be represented as an *image charge*. Such charges have the effect of reducing the Coulomb part of the addition energy. In experiments with molecular junctions, much smaller addition energies are often observed than in gas-phase molecules. At the time of writing, there is no quantitative understanding of the addition energy in molecular three-terminal junctions, although the effects mentioned here are commonly held responsible for the observed gaps.

2.5
Charge Transport Measurements as a Spectroscopic Tool

A stability diagram can be used not only for finding addition energies, but also to form a spectroscopic tool for revealing subtle excitations that arise on top of the ground state configurations of an island with a particular number of electrons on it. These excitations appear as lines running parallel to the Coulomb diamond edges. An example taken from Ref. [7] is shown in Figure 2.8a, where the white arrows indicate the excitation lines. At such a line, a new state (electronic or vibrational) enters the bias window, thus creating an additional transport channel. The result is a stepwise increase in the current, and a corresponding peak in the differential conductance. The energy of an excitation can be determined by reading off the bias voltage of the intersection point between the excitation line and the Coulomb diamond edge through the same argument used for finding addition energies. The excitations correspond to the charge state of the Coulomb diamond at which they ultimately end (see Figure 2.8c). The width of the lines in the dI/dV plot (or, equivalently, the voltage range over which the stepwise increase in current occurs) is determined by the larger of the energies k_BT and Γ. In practice, this means that sharp lines – and thus accurate information on spectroscopic features – are obtained at low temperatures and

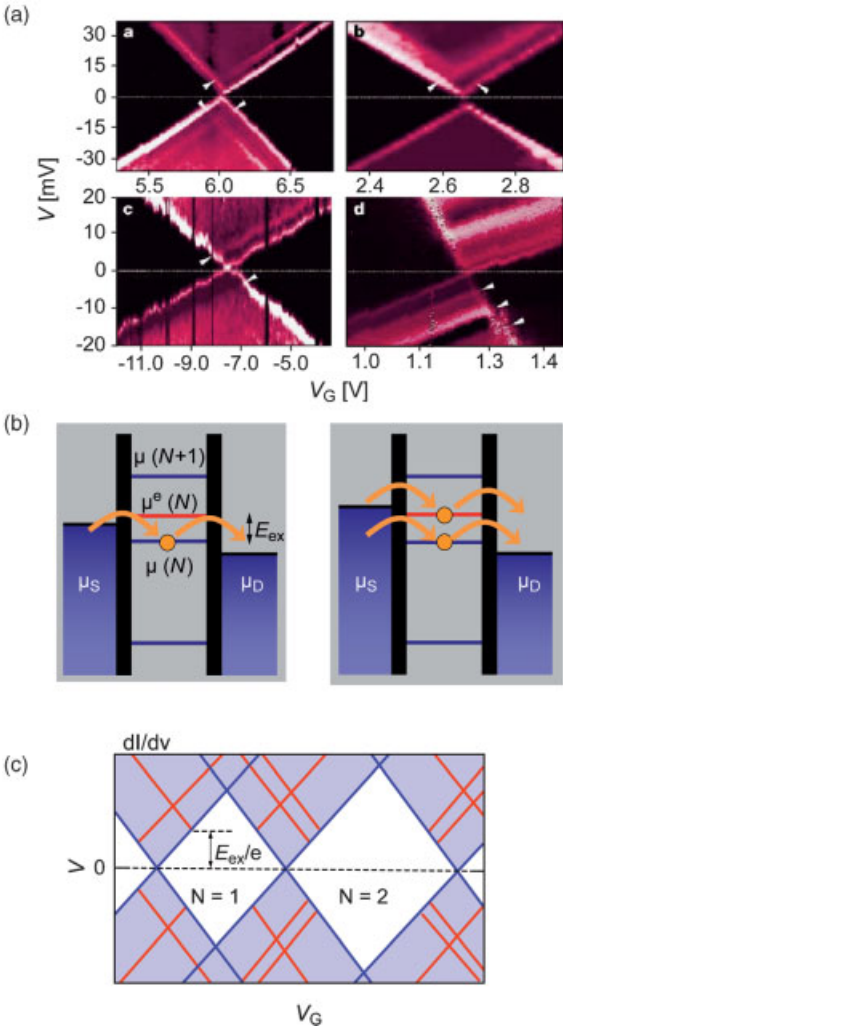


Figure 2.8 Non-linear transport and excited states. (a) Four different conductance maps (stability diagrams) of C_{60} molecules trapped between two electrodes [7]. Excitations lines are indicated by arrowheads. These run parallel to the diamond edges, and are due to vibrational modes of the C_{60} molecule. (b) Electrochemical potential plot of a dot with three electronic energy levels and one excited state (red). Transport through an excited level becomes possible as soon as the red level enters the bias window. (c) Schematic representation of a differential conductance map. The red lines show the positions at which excited states enter the bias window. The associated stepwise increases in current appear as lines running parallel to the edges of the diamond-shaped regions. Blue: dI/dV is zero but the current is not (sequential-tunneling regime). White: current blockade.

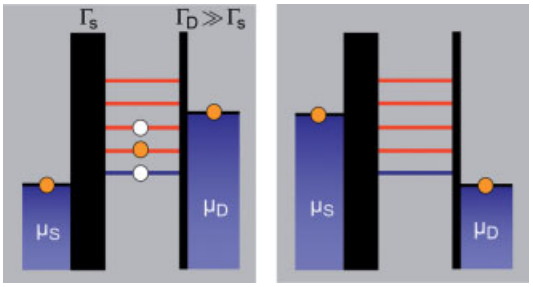


Figure 2.9 Asymmetric coupling to the electrodes leads to an almost full occupancy if tunneling out of the level is limited by the thick barrier (left: low tunnel rate) and almost zero occupancy of the level if tunneling out is determined by the thin barrier (right: high tunnel rate). Note, that of the three levels in the bias window, only one of them can be occupied at the same time. An increase of the bias voltage such that another excited level enters the bias window yields a very small current increase in the case (as depicted on the left-hand side) because the thick barrier remains the limiting factor for the current. In contrast, on the right-hand side a new transport channel becomes available and the current shows a clear stepwise increase.

for weak coupling to the leads. However, it should be noted that the current is proportional to Γ [Equations 2.4 and 2.5], so that Γ should not be too small; in fact, a Γ-value in the order of 0.1–1 meV seems typical in experiments that allow for spectroscopy.

An important experimental issue here is that, for a particular charge state, the lines are often visible on only one side of the Coulomb diamond (see Figure 2.8a, lower right panel). This is due to an asymmetry in the coupling – that is, for $\Gamma_D \gg \Gamma_S$ (or $\Gamma_S \gg \Gamma_D$). The situation at the two "main" diamond edges is illustrated in Figure 2.9. A thick and a thin barrier between the island and the source/drain represent these anti-symmetric couplings. It is clear that if the chemical potential in the lead connected through the thin barrier is the higher one, then the island will have one of its transport channels filled. The limiting step for transport is the thick barrier, and only the occupied orbital will contribute to the current. When an extra transport orbital becomes available, this will have only a minor effect on the total current, but if the chemical potential of the lead beyond the thick barrier is high, then the transport levels on the island will all be empty. The lead electrons which must tunnel through the thick barrier have as many possible channels at their disposal as there are possible empty states: the more orbitals, the more channels there are, and therefore a stepwise increase occurs each time a new excitation becomes available.

2.5.1
Electronic Excitations

In order to study how detailed information on the electronic structure of the island can be obtained from conduction measurements, we consider a system consisting of levels that are separated in energy by the Δ_i (see Figure 2.10). It should be noted that this level splitting does not include a charging energy: the levels can be occupied in charge-neutral excitations. For one extra electron on the island, $N = 1$, the ground

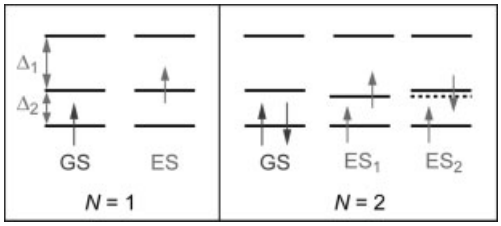


Figure 2.10 Schematic drawing of the ground state (GS) filling and the excited states (ES). Left: The island contains one electron and the first excited state involves a transition to the nearest unoccupied level. (In a zero magnetic field there is an equal probability to find a down spin on the dot.) Right: Two electrons with opposite spin occupy the lowest level. The first excited state involves the promotion of one of the spins to the nearest unoccupied level. A ferromagnetic interaction favors a spin flip. The antiparallel configuration (ES_2) has a higher energy (see text).

state is the one in which it occupies the lowest level. As discussed above, as soon as this level is inside the bias window, the current begins to flow, thereby defining the edges of the Coulomb diamonds. When the bias increases further, transport through the excited level becomes possible. This leads to a step-wise increase of the current as there are now two states available for resonant transport, and this increases the probability for electrons to pass through the island. It should be noted that both levels cannot be occupied at the same time, as this requires a charging energy in addition to the level splitting. The resulting peak in the dI/dV forms a line (red) inside the conducting region (blue), ending up at the "$N=1$" diamond (white), as shown in Figure 2.8c. ($E_{ex}=\Delta_1$ in this case). A second excitation is found at $\Delta_1+\Delta_2$; subsequent excitations intersect the diamond edge at bias voltages $\sum_i \Delta_i$, but they are only visible if $\sum_i \Delta_i < e^2/C$.

Now, we consider the case where two electrons are added to the neutral island ($N=2$). When two electrons occupy the lower orbital, the Fermi principle requires their spin to be opposite. The first excited state is the one in which one of the electrons is transferred to the higher orbital, which costs an energy of Δ_1. A ferromagnetic exchange coupling favors a triplet state with a parallel alignment. If we take only exchange interactions between different orbitals into account, this results in an energy gain of J with respect to the situation with opposite spins. Thus, the first excitation is expected to be at $\Delta_1 - J$, and the second one (corresponding to opposite spins) at Δ_1. The energy difference between the two excitations in Figure 2.8c provides a direct measure of J. In some systems, J may be negative (antiferromagnetic case) and the antiparallel configuration has a lower energy.

The simple analysis presented here captures some of the basic features of few-electron semiconducting quantum dots [8] in which the charge states to which the levels belong, can be identified. The complete electron spectrum has also been determined in metallic CNT quantum dots [5,9]. Although, for a nanotube, many densely spaced excitations occur, level spectroscopy is possible as the regularly spaced levels are well separated from each other with $E_C \approx \Delta$. Careful inspection of the excitation and addition spectra of CNTs shows that the exchange coupling J is ferromagnetic and that it is small, of the order of a few meV, or less. Further

identification of the states can be performed in a magnetic field, using the Zeeman effect as a diagnostic tool. Douplet states are expected to split into two levels, and triplet states into three.

One final remark concerns the $N = 0$ diamond. In systems such as semiconducting quantum dots, where there is a gap separating the ground state from the first excited state, Δ_1 may be of the order of hundreds of meV, and in that case no electronic excitations are expected to end up in this diamond.

2.5.2 Including Vibrational States

An interesting phenomenon in molecular transport occurs when the molecular vibrations couple to the electrons, giving rise to excitations available for transport (as mentioned above). This phenomenon has been studied quite extensively in recent years, and the basics will briefly be discussed at this point (for further details, see Refs. [10,11]).

Molecules are rather "floppy" in nature, and from classical mechanics it is known that small deformations of a molecule with respect to its lowest energy conformation can be described in terms of *normal modes*. These are excitations in which all nuclei oscillate with the *same* frequency ω (although some nuclei may stand still). In particular, these excitations have the form

$$R_{i,\alpha}^{(l)}(t) = X_{i,\alpha}^{(l)} \exp(i\omega^{(l)} t),$$

where $R_{i\alpha}^{(l)}$ is the Cartesian coordinate $\alpha = x, y, z$ of nucleus i; l labels the normal mode; $X_{i,\alpha}^{(l)}$ is a fixed vector which determines the amplitudes of the oscillation for the degree of freedom labeled by i,α. The vibrations are described by a harmonic oscillator, which has a spectrum with energy levels separated by an amount $\hbar\omega^{(l)}$:

$$E_v^{(l)} = \hbar\omega^{(l)}(v + 1/2), \quad v = 0, 1, 2, \ldots.$$

For molecular systems, the normal modes are often referred to as *vibrons* (in analogy with phonons in a periodic solid). These modes couple with the electrons as the electrons feel a change in the electrostatic potential when the nuclei move in a normal mode. The coupling is determined by the electron–vibron coupling constant, called λ.

The presence of vibrational excitations can be detected in transport measurements. However, it should be noted that for this to happen, the vibrational modes must be excited, which can occur for two reasons: (i) the thermal fluctuations excite these modes; or (ii) they can be excited through the electron–vibron coupling.

In order to study the effect of electron–vibron coupling on transport, for simplicity the discussion is restricted to a single vibrational mode and a single electronic level. The nuclear part of the Hamiltonian is

$$H = \frac{P^2}{2M} + \frac{1}{2} M\omega^2 X^2$$

where P, X and M represent the momentum, position and mass of the oscillator.

It twins out [11] that the electron-vibron coupling has the form:

$$H_{e-v} = \lambda\hbar\omega\hat{n}X/u_0,$$

where $\hat{n}$ is the number operator, which takes on the values 0 or 1 depending on whether there is an electron in the orbital under consideration; $u_0 = \sqrt{\hbar/(2M\omega)}$ is the zero-point fluctuation associated with the ground state of the harmonic oscillator. The electron–vibron coupling λ is given as (φ is the electronic orbital):

$$\lambda = \frac{1}{\hbar\omega}\sqrt{\frac{\hbar}{2\omega}}\frac{1}{\sqrt{M}}\left\langle \varphi \left| \frac{\partial H_{el}}{\partial X} \right| \varphi \right\rangle.$$

When the charge in the state φ increases from 0 to 1, the equilibrium position of the harmonic oscillator (i.e., the minimum of the potential energy) is shifted over a distance $-2\lambda u_0$ along X, and it is also shifted down in energy (see Figure 2.11a). Fermi's "golden rule" states that the transition rate for going from the neutral island in the conformational ground state to a charged island in some

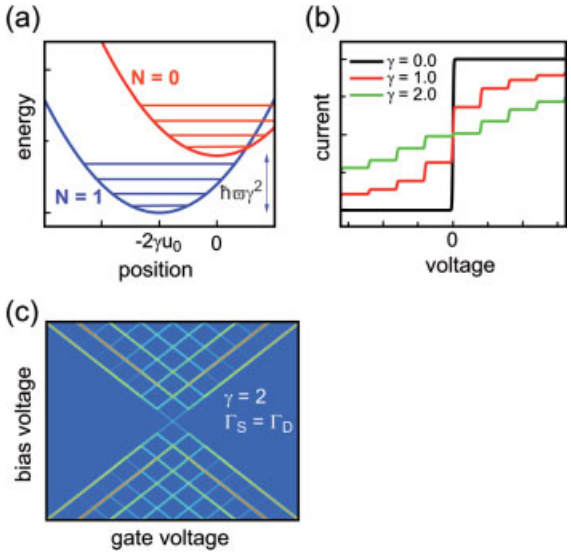


Figure 2.11 (a) Potential of the harmonic oscillator for the empty (red) and occupied state (blue). When an electron tunnels onto the island, the position of the potential minimum is shifted in space and energy. (b) Current–voltage characteristics calculated for three different values of the electron–phonon coupling constant. For non-zero coupling steps appear, which are equally spaced in the voltage (harmonic spectrum). (c) Differential conductance plotted in a stability diagram for an island coupled to a single vibrational mode. Lines running parallel to the diamond edges correspond to the steps, forming a lozenge pattern of excitation lines in (b). Around zero bias the current is suppressed for this rather large electron–phonon coupling (phonon blockade).

excited vibrational state is proportional to the square of the overlap between the initial and final states. Hence, this rate is proportional to the overlap of the ground state of the harmonic oscillator corresponding to the higher parabola and the excited state of the oscillator corresponding to the shifted parabola (to be multiplied by the coupling between lead and island). This overlap is known as the Frank–Condon factor. It is clear that for large displacements, this overlap may be larger for passing to a vibrationally excited state than for passing to the vibrational ground state of the shifted oscillator. The Franck–Condon factors may be calculated analytically (see for example Ref. [10]).

The sequential tunneling regime, which corresponds to weak coupling, can be described in terms of a rate equation: the *master equation*. This describes the time evolution of the probability densities for the possible states on the molecular island. The master equation can be used for any sequential tunneling process and is particularly convenient when vibrational excitations play a role. The details of formulating and solving master equations are beyond the scope of this chapter, but the interested reader is referred to Refs. [11,12] for further details.

Figure 2.11b and c were prepared using such a master equation analysis. For sufficient electron–phonon coupling, steps appear in the current–voltage characteristics (Figure 2.11b), which for $\Gamma_D = \Gamma_S$ leads to the lozenge pattern in a stability diagram, as illustrated in Figure 2.11c. It should be noted that, if the vibrational modes are excited, they may in turn lose their energy through coupling to the leads or other parts of the device. This can be represented by an effective damping term for the nuclear degrees of freedom. For actual molecules, solving the master equations by using Frank–Condon factors obtained from quantum chemical calculations may be used to compare theory with experiment. This is especially useful because the observed vibrational frequencies can be used as a "fingerprint" of the molecule under study [7,13–15] (see also Figure 2.8a).

2.6 Second-Order Processes

In the analysis presented so far, sequential tunneling events do not contribute to the current inside Coulomb diamonds as they are blocked in these regions. However, it should be realized that elastic co-tunneling processes (as depicted in the upper part of Figure 2.3) always take place, albeit that the current levels are generally very small. For second-order processes, the current is proportional to $\Gamma_S\Gamma_D$ instead of showing a linear dependence (on $\Gamma_S\Gamma_D/(\Gamma_S + \Gamma_D)$) as for first-order processes. Consequently, co-tunneling becomes more important for larger Γ-values. In some cases, higher order coherent processes involving virtual states give rise to observable features inside Coulomb diamonds. In the following section, two examples are discussed; namely, the Kondo effect in quantum dots, which is an elastic co-tunneling process conserving the dot energy, and inelastic co-tunneling, which leaves the dot in an excited state.

2.6.1
The Kondo Effect in a Quantum Dot with an Unpaired Electron

The Kondo effect has long been known to cause a resistance increase at low temperatures in metals with magnetic impurities [16]. However, in recent years Kondo physics has also been observed in semiconducting [17], nanotube [18] and single-molecule quantum dots [19]. It arises when a localized unpaired spin interacts by antiferromagnetic exchange with the spin of the surrounding electrons in the leads (see Figure 2.12a). The Heisenberg uncertainty principle allows the electron to tunnel out for only a short time of about $\hbar/\Delta E$, where ΔE is the energy of the electron relative to the Fermi energy and is taken as positive. During this time, another electron from the Fermi level at the opposite lead can tunnel onto the dot, thus conserving the total energy of the system (elastic co-tunneling). The exchange interaction causes the majority spin in the leads to be opposite to the original spin of the dot. Therefore, the new electron entering from these leads is more likely to have the opposite spin. This higher-order process gives rise to a so-called *Kondo resonance* centered around the Fermi level. The width of this resonance is proportional to the characteristic energy scale for Kondo physics, T_K. For $\Delta E \gg \Gamma$, T_K is given by:

$$k_B T_K = \frac{\sqrt{\Gamma U}}{2} \exp\left[\frac{\pi \Delta E(\Delta E + U)}{\Gamma U}\right]. \tag{2.11}$$

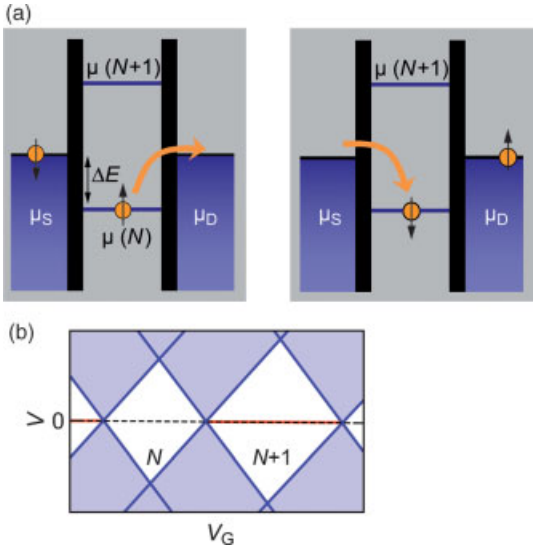


Figure 2.12 (a) A schematic drawing of the two-step Kondo process which occurs for odd occupancy of the island. (b) The Kondo effect leads to a zero bias conductance peak (red lines) in the differential conductance plots. *N* is even in this case.

Typical values for T_K are 1 K for semiconducting quantum dots, 10 K for CNTs, and 50 K for molecular junctions. This increase of T_K with decreasing dot size can be understood from the prefactor, which contains the charging energy ($U = e^2/C$).

In contrast to bulk systems, the Kondo effect in quantum dots leads to an *increase* of the conductance, as exchange makes it easier for the spin states belonging to the two electrodes to mix with the state (of opposite spin) on the dot, thereby facilitating transport through the dot. The conductance increase occurs only for small bias voltages, and the characteristic feature is a peak in the trace of the differential conductance versus bias voltage (see Figure 2.10b, red lines). The peak occurs at zero bias inside the diamond corresponding to an odd number of electrons. (For zero spin, no Kondo is expected; for $S = 1$ a Kondo resonance may be possible, but the Kondo temperature is expected to be much smaller.) The full width at half maximum (FWHM) of this peak is proportional to T_K: FWHM $\approx 2k_B T_K/e$. Equation 2.11 indicates that T_K is gate-dependent because ΔE can be tuned by the gate voltage. Consequently, the width of the resonance is the smallest in the middle of the Coulomb blockade valley and increases towards the degeneracy point on either side.

Another characteristic feature of the Kondo resonance is the logarithmic decrease in peak height with temperature. In experiments, this logarithmic dependence of the conductance maximum is often used for diagnostic means, and in the middle of the Coulomb blockade valley it is given by:

$$G(T) = \frac{G_C}{[1 + (2^{1/S} - 1)(T/T_K)^2]^S}, \tag{2.12}$$

where $S = 0.22$ for spin-1/2 impurities and $G_C = 2e^2/h$ for symmetric barriers. For asymmetric barriers, G_C is lower than the conductance quantum. Equation 2.12 shows that for low temperatures, the maximum conductance of the Kondo peak saturates at G_C while at the Kondo temperature it reaches a value of $G_C/2$.

2.6.2 Inelastic Co-Tunneling

The inelastic co-tunneling mechanism becomes active above a certain bias voltage, which is independent of the gate voltage. At this point, the current increases stepwise because an additional transport channel opens up. In the stability diagram, this results in a horizontal line inside the Coulomb-blockaded regime. This conductance feature appears symmetrically around zero at a source-drain bias of $\pm\Delta/e$ for an excited level that lies at an energy Δ above the ground state. Co-tunneling spectroscopy therefore offers a sensitive measure of excited-state energies, which may be either electronic or vibrational. Often, in combination with Kondo peaks, inelastic co-tunneling lines are commonly observed in semiconducting, nanotube and molecular quantum dots. An example of inelastic co-tunnel lines (dashed horizontal lines) for a metallic nanotube quantum dot is shown in Figure 2.13a.

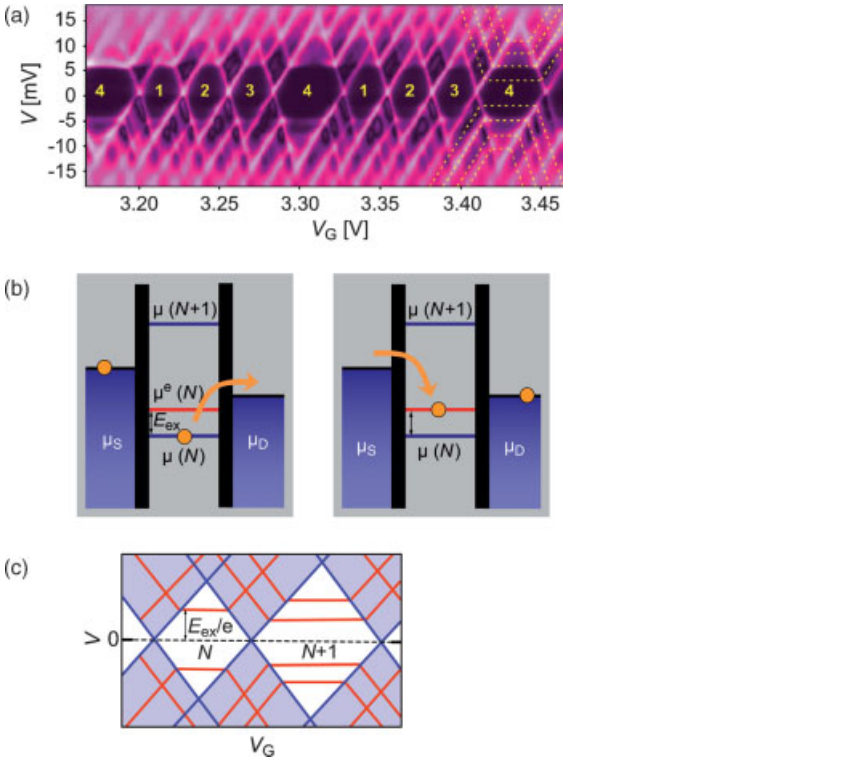


Figure 2.13 Inelastic co-tunneling. (a) A measured stability diagram of a metallic, single-walled carbon nanotube (from Ref. [5]). The dashed horizontal lines indicate the presence of inelastic co-tunnel lines. (b) Schematic drawing of this two-step tunneling process, leaving the dot in an excited state. (c) Inelastic co-tunneling gives rise to horizontal lines in the blocked current region. The energy of the excitation can directly be determined from these plots, as indicated in the figure.

Figure 2.13b shows the mechanism of inelastic co-tunneling. An occupied state lies below the Fermi level, and this can only virtually escape for a small time, as governed by the Heisenberg uncertainty relation. If, in the meantime, an electron from the left lead tunnels onto the dot in the excited level (red), then effectively one electron has been transported from left to right. The dot is left in an excited level, and the energy difference E_{ex} must be paid by the bias voltage; hence, this two-step process is only possible for $|V| > E_{ex}/e$. Relaxation inside the dot may then lead to the dot to decay into the ground state.

Acknowledgments

The authors thank Menno Poot for his critical reading of the manuscript.

References

1 Feynman, R.P. (1961) There is plenty of room at the bottom, in *Miniaturization*, (ed. H.D. Hilbert), Reinhold, New York, pp. 282–286.

2 Datta, S. (1995) *Electronic transport in mesoscopic systems*, Cambridge University Press, Cambridge.

3 Beenakker, C.W.J. (1991) Theory of Coulomb-blockade oscillations in the conductance of a quantum dot. *Physical Review B-Condensed Matter*, **44**, 1646–1656.

4 Foxman, E.B. *et al.* (1994) Crossover from single-level to multilevel transport in artificial atoms. *Physical Review B-Condensed Matter*, **50**, 14193–14199.

5 Sapmaz, S., Jarillo-Herrero, P., Kong, J., Dekker, C., Kouwenhoven, L.P. and van der Zant, H.S.J. (2005) Electronic excitation spectrum of metallic nanotubes. *Physical Review B-Condensed Matter*, **71**, 153402.

6 Oreg, Y., Byczuk, K. and Halperin, B.I. (2000) Spin configurations of a carbon nanotube in a nonuniform external potential. *Physical Review Letters*, **85**, 365–368.

7 Park, H., Park, J., Lim, A.K.L., Anderson, E.H., Alivisatos, A.P. and McEuen, P.L. (2000) Nanomechanical oscillations in a single-C_{60} transistor. *Nature*, **407**, 57–60.

8 Kouwenhoven, L.P., Austing, D.G. and Tarucha, S. (2001) Few-electron quantum dots. *Reports on Progress in Physics*, **64**, 701–736.

9 Liang, W., Bockrath, M. and Park, H. (2002) Shell filling and exchange coupling in metallic single-walled carbon nanotubes. *Physical Review Letters*, **88**, 126801.

10 Flensberg, K. and Braig, S. (2003) Incoherent dynamics of vibrating single-molecule transistors. *Physical Review B-Condensed Matter*, **67**, 245415.

11 Mitra, A., Aleiner, I. and Millis, A.J. (2004) Phonon effects in molecular transistors: Quantum and classical treatment. *Physical Review B-Condensed Matter*, **69**, 245302.

12 Koch, J. and von Oppen, F. (2005) Franck–Condon blockade and giant fano factors in transport through single molecules. *Physical Review Letters*, **94**, 206804.

13 Djukic, D., Thygesen, K.S., Untiedt, C., Smit, R.H.M., Jacobsen, K.W. and van Ruitenbeek, J.M. (2005) Stretching dependence of the vibration modes of a single-molecule Pt-H_2-Pt bridge. *Physical Review B-Condensed Matter*, **71**, 161402.

14 Pasupathy, A.N., Park, J., Chang, C., Soldatov, A.V., Lebedkin, S., Bialczak, R.C., Grose, J.E., Donev, L.A.K., Sethna, J.P., Ralph, D.C. and McEuen, P.L. (2005) Vibration-assisted electron tunneling in C140 single-molecule transistors. *Nano Letters*, **5**, 203–207.

15 Osorio, E.A., O'Neill, K., Stuhr-Hansen, N., Faurskov Nielsen, O., Bjørnholm, T. and van der Zant, H.S.J. (2007) Addition energies and vibrational fine structure measured in electromigrated single-molecule junctions based on an oligophenylenevinylene derivative. *Advanced Materials*, **19**, 281–285.

16 Kondo, J. (1964) Resistance minimum in dilute magnetic alloys. *Progress of Theoretical Physics*, **32**, 37.

17 (a) Goldhaber-Gordon, D., Shtrikman, H., Mahalu, D., Abusch-Magder, D., Meirav, U. and Kastner, M.A. (1998) Kondo effect in a single-electron transistor. *Nature*, **391**, 157–159.(b) Cronenwett, S.M., Oosterkamp, T.H. and Kouwenhoven, L.P. (1998) A tunable Kondo effect in quantum dots. *Science*, **281**, 540–544.

18 Nygård, J., Cobden, D.H. and Lindelof, P.E. (2000) Kondo physics in carbon nanotubes. *Nature*, **408**, 342–346.

19 (a) Park, J. *et al.* (2002) Coulomb blockade and the Kondo effect in single-atom transistors. *Nature*, **417**, 722–725. (b) Liang, W., Shores, M.P., Bockrath, M.,

Long, J.R. and Park, H. (2002) Kondo resonance in a single-molecule transistor. *Nature*, **417**, 725–729. (c) Yu, L.H., Keane, Z.K., Ciszek, J.W., Cheng, L., Tour, J.M., Baruah, T., Pederson, M.R. and Natelson, D. (2005) Kondo resonances and anomalous gate dependence in the electrical conductivity of single-molecule transistors. *Physical Review Letters*, **95**, 256803.

3
Spin Injection–Extraction Processes in Metallic and Semiconductor Heterostructures

Alexander M. Bratkovsky

3.1
Introduction

Spin transport in metal-, metal-insulator, and semiconductor nanostructures holds promise for the next generation of high-speed, low-power electronic devices [1–10]. Amongst important spintronic effects already used in practice are included a giant magnetoresistance (MR) in magnetic multilayers [11] and tunnel MR (TMR) in ferromagnet-insulator-ferromagnet (FM-I-FM) structures [12–19]. The injection of spin-polarized electrons into semiconductors is of particular interest because of relatively large spin relaxation time ($\sim$1 ns in semiconductors, $\sim$1 ms in organics) [2] during which the electron can travel over macroscopic distances without losing polarization, or stay in a quantum dot/well. This also opens up possibilities, albeit speculative ones, for quantum information processing using spins in semiconductors.

The potential of spintronic devices is illustrated most easily with a simple spin-dependent transport process, which is a tunneling magnetoresistance in FM-I-FM structure (see the next section). The effect is a simple consequence of the golden rule that dictates a dependence of the tunnel current on the density of initial and final states for tunneling electron. Most of the results for tunnel spin junctions may be reused later in describing the spin injection from ferromagnets into semiconductors (or vice versa) in the later sections.

It is worth noting from the outset that there are two major characteristics of the spin transport processes that will define the outcome of a particular measurement, namely *spin polarization* and *spin injection efficiency*. These may be very different from each other, and this may lead (and frequently does) to a confusion among researchers. The spin polarization measures the imbalance in the *density* of electrons with opposite spins (spin accumulation/depletion),

$$P = \frac{n_\uparrow - n_\downarrow}{n_\uparrow + n_\downarrow}, \tag{3.1}$$

Nanotechnology. Volume 3: Information Technology I. Edited by Rainer Waser

ISBN: 978-3-527-31738-7

while the injection *efficiency* is the *polarization* of injected *current J*

$$\Gamma = \frac{J_\uparrow - J_\downarrow}{J_\uparrow + J_\downarrow}, \tag{3.2}$$

where $\uparrow(\downarrow)$ refers to the electron spin projection on a quantization axis. In case of ferromagnetic materials, the axis is antiparallel to the magnetization moment $\vec{M}$. Generally, $P \neq \Gamma$, but in some cases they can be close. Since in the ferromagnets the spin density is constant, a reasonable assumption can be made that the current is carried independently by two electron fluids with opposite spins (Mott's two-fluid model [20]). Then, in the FM bulk the injection efficiency parameter is

$$\Gamma = \Gamma_F = \frac{(\sigma_\uparrow - \sigma_\downarrow)}{\sigma}, \tag{3.3}$$

where $\sigma_{\uparrow(\downarrow)}$ are the conductivities of up-, (down)spin electrons in ferromagnet, $\sigma = \sigma_\uparrow + \sigma_\downarrow$.

In the case of spin tunneling, it is found that Γ characterizes the value of MR in magnetic *tunnel* junctions, which is quite obvious as there one measures the difference between *currents* in two configurations: with parallel (P) and antiparallel (AP) moments on electrodes. The tunnel current is small; hence the injected spin density is minute compared to metallic carrier densities. At the same time, in experiments where one injects spin (creates a non-equilibrium spin population) in a quantum well, this results in the emission of polarized light (spinLED), the measured intensity of which is, obviously, proportional to the spin polarization *P* (see below).

We will outline the major spin-transport effects here in the Introduction, starting with an analysis of tunnel magnetoresistance (TMR), followed by giant magnetoresistance (GMR) and spin-torque (ST) switching in magnetic nanopillars. We will then outline the spin-orbit effects in three-dimensional (3-D) and two-dimensional (2-D) semiconductor systems (Dresselhaus and Vasko-Rashba effects) and use this for further discussion of Datta–Das interference device and the Spin-Hall effect. A brief assessment will then be made of spin logic devices, showing why they are impractical because of difficulty in precise manipulation of individual spins, in addition to practically gapless excitation of the spin waves that easily destroy a particular spin configuration of the multispin system. Spin ensemble-based quantum computing – that is, coherent manipulation of spin ensembles over a number of steps – is, even less likely than using classical spin logic.

We then turn to spin injection/extraction effects in ferromagnetic metal-semiconductor heterojunctions. It is shown that efficient spin injection is possible with modified Schottky junctions, and is a strongly non-linear effect. A brief discussion is then provided of a few possible spin-injection effects and devices, some of which are likely to be demonstrated in the near future. The final section is devoted to a complementary topic of spin injection in degenerate semiconductors. These processes are significantly different from the case of non-degenerate semiconductors so as to warant a separate discussion.

3.2
Main Spintronic Effects and Devices

3.2.1
TMR

Tunnel magnetoresistance is observed in metal–insulator–metal magnetic tunnel junctions (MTJ), usually with Ni–Fe, Co–Fe electrodes and (amorphous) Al_2O_3 tunnel barrier where one routinely observes upward of 40–50% change in conductance as a result of the changing relative orientation of magnetic moments on electrodes. A considerably larger effect, about 200% TMR, is found in Fe/MgO/Fe junctions with an epitaxial barrier, that may be related to surface states and/or peculiarities of the band structure of the materials. As will be seen shortly, TMR is basically a simple effect of an asymmetry between densities of spin-up and -down (initial and final tunneling) states. TMR is intimately related to giant magnetoresistance [11]; that is, a giant change in conductance of magnetic multilayers with relative orientation of magnetic moments in the stack.

We can estimate the TMR using the golden rule which states that the tunnel current at small bias voltage V is $J_\sigma = G_\sigma V$, $G_\sigma \propto |M|^2 g_{i\sigma}\, g_{f\sigma}$, where $g_{i(f)\sigma}$ is the density of initial (final) tunneling states with a spin projection σ, and M is the tunneling matrix element. Consider the case of electrodes made from the same material. It is clear from the band schematic shown in Figure 3.1 that the total rates of tunneling in parallel (a) and antiparallel (b) configurations of moments on electrodes are different.

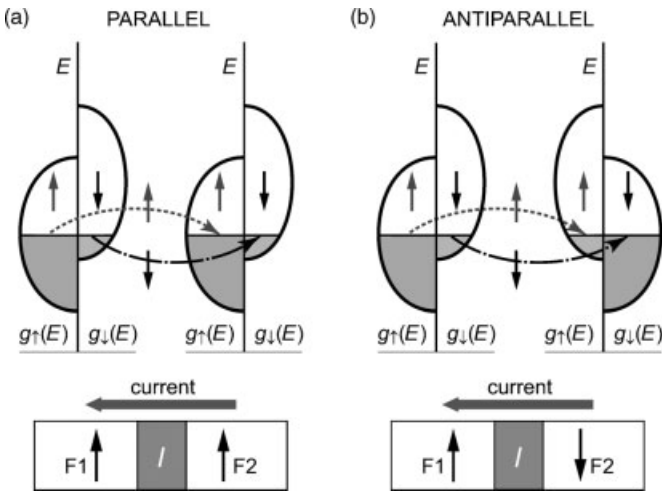


Figure 3.1 Schematic illustration of spin tunneling in FM-I-FM junction for (a) parallel (P) and (b) antiparallel (AP) configuration of moments on ferromagnetic electrodes. Top panels: band diagram; bottom panels: schematic of the corresponding magnetic configuration in the junction with regards to current direction.

Indeed, denoting $D = g_\uparrow$ and $d = g_\downarrow$ as the partial densities of states (DOS), we can write down the following golden rule expression for parallel and antiparallel moments on the electrodes:

$$G_P \propto D^2 + d^2, \quad G_{AP} \propto 2Dd, \tag{3.4}$$

and arrive at the expression for TMR first derived by Jullieres [12]

$$\mathrm{TMR} \equiv \frac{G_P - G_{AP}}{G_{AP}} = \frac{(D-d)^2}{2Dd} = \frac{2P^2}{1-P^2}, \tag{3.5}$$

where we have introduced a polarization P, which fairly approximates the polarization introduced in Equation 3.1, at least for narrow interval of energies:

$$P = \frac{D-d}{D+d} \equiv \frac{g_\uparrow - g_\downarrow}{g_\uparrow + g_\downarrow}. \tag{3.6}$$

Below, we shall see that the "polarization" entering expression for a particular process, depends on particular physics and also on the nature of the electronic states involved. It should be noted that, for instance, the DOS entering the above expression for TMR, is not the *total* DOS but rather the one for states that contribute to tunneling current. Thus, Equation 3.6 may lead one to believe that the *tunnel* polarization in elemental Ni should be negative, as there is a sharp peak in the minority carrier density of states at the Fermi level. The data, however, suggest unambiguously that the tunnel polarization in Ni is positive [14], $P > 0$. This finds a simple explanation in a model by M.B. Stearns, who highlighted the presence in elemental 3d metals parts of Fermi surface with almost 100% d-character and a small effective mass close to one of a free electron [21]. A detailed discussion of TMR effects is provided below.

3.2.2 GMR

There are important differences between TMR and GMR processes. Indeed, GMR is most reminiscent of TMR for current-perpendicular-to-planes (CPP) geometry in FM-N-FM-... stacks, where N stands for normal metal spacer (Figure 3.2a). In the CPP geometry, the spins cross the nanometer-thin normal spacer layer (N) without spin flip, similarly to tunneling through the oxide barrier, but the elastic mean free path is comparable or smaller than the N thickness, so that a drift-diffusive electrons transport takes place in metallic GMR stacks. In commonly used current-in-plane geometry (CIP), the electrons bounce between different ferromagnetic layers, (Figure 3.2b) effectively showing the same motif in transport across the layers as in the CPP geometry. Comparing with TMR, the latter (and spin injection efficiency) depends on the difference between the densities of states g_σ, spin $\sigma = \uparrow(\downarrow)$ at the Fermi level, while GMR depends on relative conductivity

$$\sigma_\sigma = e^2 \langle g_\sigma v_\sigma^2 \tau_\sigma \rangle_F,$$

where the angular brackets indicate an average over the Fermi surface that involves the Fermi velocity v_σ and the momentum relaxation time τ_σ One can still use the

Mott's two independent spin fluid picture, but as one is dealing with metallic heterostructure, the continuity (or Boltzmann) equations must be solved for a periodic FM-N-FM-... stack to find the ramp of an electrochemical potential that defines the total current. Neglecting any slight imbalance of electrochemical potentials for two spins in the N regions (spin accumulation), one may construct an equivalent circuit model for CPP stack in the spirit of Mott's model, and thus qualitatively explain the GMR. The parallel "spin" layer resistances would be $R_{\uparrow(\downarrow)}$ $\sigma_{\uparrow(\downarrow)}L_F$ for the FM layers, and $r = \sigma_N^{-1}L_N$ in the normal N regions with thicknesses $L_F(L_N)$, respectively. For conductances in two configurations of the moments, we then obtain (Figure 3.2c):

$$G_{\mathrm{P}} = \frac{1}{R_{\mathrm{P}}} = \frac{1}{2R_{\uparrow} + r} + \frac{1}{2R_{\downarrow} + r}, \tag{3.7}$$

$$G_{\mathrm{AP}} = \frac{1}{R_{\mathrm{AP}}} = \frac{2}{R_{\uparrow} + R_{\downarrow} + r}, \tag{3.8}$$

and the GMR simply becomes

$$\mathrm{GMR} = \frac{G_{\mathrm{P}} - G_{\mathrm{AP}}}{G_{\mathrm{AP}}} = \frac{(R_{\downarrow} - R_{\uparrow})^2}{(2R_{\uparrow} + r)(2R_{\downarrow} + r)} \tag{3.9}$$

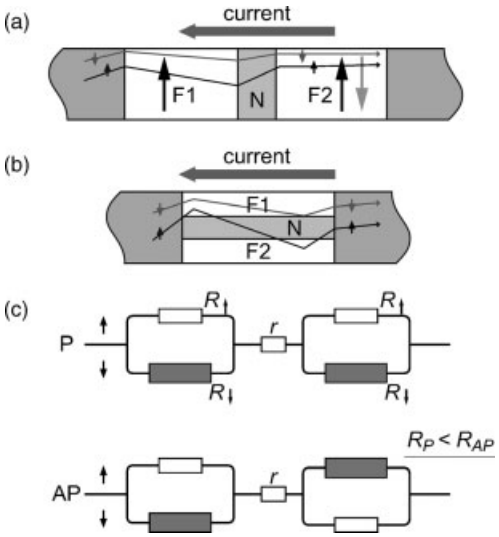


Figure 3.2 Schematic of giant magnetoresistance (GMR) in (a) current perpendicular to plane (CPP) and (b) current in-plane (CIP) geometries. Electron scattering in the GMR valve depends on the configuration of moments and spin of traversing electrons in both configurations. (c) Equivalent circuit model for GMR. Two-fluid spin model that is valid in the absence of spin relaxation, leads to higher resistivity for antiparallel arrangement of magnetic moments in the spin valve, $R_{AP} > R_P$.

which we can rewrite as

$$\mathrm{GMR} = \frac{\Gamma_F^2}{(1 + r/R)^2 - \Gamma_F^2} \approx \frac{\Gamma_F^2}{1 - \Gamma_F^2}, \tag{3.10}$$

where $R = R_\uparrow + R_\downarrow$. The latter is very similar to the expression TMR [Equation 3.5], but there is an absence of a factor two in the numerator. It can be seen that the polarization entering GMR is different from that entering TMR. Even in more common CIP geometry, the electrons certainly do scatter across the interfaces; hence, this equation can also be used for semi-quantitative estimates of the GMR effect in CIP geometry (Figure 3.2c). Obviously, the effective circuit model [Equations 3.7 and 3.8] remains exactly the same because it simply reflects the two-fluid approximation for contributions of both spins. However, all effective resistances depend in rather nontrivial manner on the geometry (CPP or CIP) and electronic structure of the metals involved that is particularly complicated in magnetic transition metals.

In terms of applications, the TMR effect is used in non-volatile magnetic random access memory (MRAM) devices and as a field sensor, while GMR is widely used in magnetic read heads as field sensors. The MRAM devices are in a tight competition with semiconductor memory cards (FLASH), since MRAM is technologically more involved (and hence more expensive) than standard silicon technology. It is problematic to use those effects in building three-terminal devices with gain that would show any advantage over standard CMOS transistors.

3.2.3
(Pseudo)Spin-Torque Domain Wall Switching in Nanomagnets

Magnetic memory based on TMR is non-volatile, may be rather fast (~1 ns), and can be scaled down considerably toward paramagnetic limits observed in nanomagnets. Switching, however, requires the MTJ to be placed at a crosspoint of the bit and word wires carrying current that produces a sufficient magnetic field for switching the domain orientation in "free" (unpinned) MTJ ferromagnetic electrodes. The undesirable side effect is a crosstalk between cells, a rather complex layout, and a power budget. Alternatively, one may take a GMR multilayer in a nanopillar form with antiparallel orientation of magnetic moments and run a current through it (this would obviously correspond to a CPP geometry) (Figure 3.3). In this case, there will be a spin accumulation in the drain layer; that is, the accumulation of *minority* spins in the drain electrode for *antiparallel* configuration of moments on electrodes (for the electrodes made out the same material. This formally means a transfer of spin (angular) moment across the space layer. Injection of angular momentum means that there is a change in the spin momentum on the drain electrode, the spin projection on the quantization axis then evolves with time simply because of an influx of electrons with a different projection of polarization on the (drain electrode) quantization axis, $d\mathcal{P}_{\mathrm{Rz}}/dt \neq 0$, $\mathcal{P}_{\mathrm{Rz}} = n_+ - n_-$, where $\pm$ marks along (against) the quantization axis on the right electrode with $n_{+(-)}$ the densities of majority (minority) electrons. One may call this a (pseudo) torque effectively acting on a moment in the

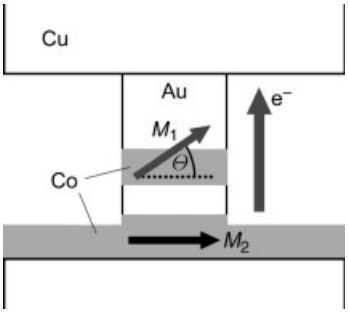


Figure 3.3 Schematic of a nanopillar device for spin-torque measurements. Current through the device results in change in angle θ between magnetic moments in the free layer in the middle and the bottom ferromagnetic electrode.

drain electrode, although this is obviously not a good term. This simple effect was predicted in Refs. [22, 23], and observed experimentally in nanopillars multilayer stacks by Tsoi *et al.* [24], and also in other studies.

The spin *accumulation* is proportional to density J of the spin-polarized current through the drain magnet or magnetic particle, that carries $(-\hbar J/2q)$ spin moment with it per second. The change in the longitudinal component of spin polarization $\mathcal{P}_{\mathrm{Rz}}$ in the right electrode next to the interface is then simply

$$\begin{aligned} d\mathcal{P}_{\mathrm{Rz}} &= -\frac{\hbar(J_+ - J_-)}{2q} = -\frac{\hbar V(\mathcal{G}_+ - \mathcal{G}_-)}{2q} \\ &= -\frac{qV(T_+ - T_-)}{4} \\ &= -\frac{\hbar V}{4q}(G_{+\uparrow} + G_{+\downarrow} - G_{-\uparrow} - G_{-\downarrow}), \end{aligned} \tag{3.11}$$

in a linear regime, where $\mathcal{G}_\pm = (q^2/\hbar)T_\pm$ are the conductances for spin-up and -down channels, expressed through the transmission probabilities $T_\pm$ and partial spin conductances G. There is an angle θ between the quantization axes on the source and drain electrode, $\cos\theta = \vec{\mathcal{P}}_{\mathrm{L}} \cdot \vec{\mathcal{P}}_{\mathrm{R}}/(\mathcal{P}_{\mathrm{L}}\mathcal{P}_{\mathrm{R}})$. The partial spin conductances G are in turn given by the standard Landauer expression through the transmission coefficients $t_{\sigma\sigma'}$ as $T_{\sigma=\pm} = \sum_{\sigma'=\uparrow,\downarrow}|t_{\sigma\sigma'}|^2$. The above expression can be expressed through the partial conductances for arbitrary configuration of spins on the electrodes $G_{+\uparrow} = (q^2/\hbar)|t_{+\uparrow}|^2$, $G_{-\uparrow} = (q^2/\hbar)|t_{-\uparrow}|^2, \ldots$, where the orientation along (against) the spin on the right electrode, $\mathcal{P}_{\mathrm{R}}$, is marked by the subscript $+(-)$. Assuming that there is no spin-flip in the oxide (non-magnetic metal) spacer, we can express the transmission amplitudes through those calculated for parallel/antiparallel configuration of spins on the electrodes with the use of a standard rule for spin wave function in a rotated frame.

Finally, the rate of change in polarization on the drain electrode due to influx of polarized current is simply

$$\frac{d\mathcal{P}_{Rz}}{dt} = -\frac{\hbar V}{4q}\left[(G_{\uparrow\uparrow} - G_{\downarrow\downarrow})\cos^2\frac{\theta}{2} + (G_{\uparrow\downarrow} - G_{\downarrow\uparrow})\sin^2\frac{\theta}{2}\right], \tag{3.12}$$

and the term on the right-hand side should be added to the driving force in the Landau–Lifshitz (LL) equation on the right-hand side. This expression should be better suited for MTJs, as in metallic spin valves one must consider spin accumulation in a metallic spacer. Note that Slonczewski obtains $d\mathcal{P}_{Rz}/dt \propto \sin\theta$ for MTJs [25], which may be inaccurate. Indeed, consider the antiparallel configuration ($\theta = \pi$) of *unlike* electrodes. Then, $d\mathcal{P}_{Rz}/dt \propto G_{\uparrow\downarrow} - G_{\downarrow\uparrow} \neq 0$, since for the unlike electrodes $G_{\uparrow\downarrow} \neq G_{\downarrow\uparrow}$, and there obviously will be a change in the spin density in the right electrode because the influx of spins into majority states would not be equal to the influx into minority states. To handle the resulting spin dynamics properly, one needs to write down the continuity equation for the spin, similar to Equation 3.23 below, with Equation 3.12 as the boundary condition at the interface.

Time-resolved measurements of current-induced reversal of a free magnetic layer in permalloy/Cu/permalloy elliptical nanopillars at temperatures from 4.2 to 160 K can be found in Ref. [26]. There is considerable device-to-device variation in the ST attributed to presence of an antiferromagnetic oxide layer around the perimeter of the Permalloy free layer (and some ambiguity in an expression used for the torque itself). Obviously, controlling this layer would be very important for the viability of the whole approach for memory applications, and so on. There are reports about the activation character of switching that may be related to pinning of the domain walls at the side walls of the pillar. The injected DC polarized current may also induce a magnetic vortex oscillation, when vortex may be formed in a magnetic island in, for example, a pillar-like spin valve. These induced oscillations have recently been found [27]. It is worth noting that the agreement between theory and experiment may be fortuitous: thus, in permalloy nanowires the speed of domain wall has substantially exceeded the rate of spin angular momentum transfer rate [28].

3.3
Spin-Orbital Coupling and Electron Interference Semiconductor Devices

In most cases of interest, such as direct band semiconductors near high-symmetry points, and a two-dimensional electron gas [29, 30], the spin-orbital (SO) coupling effects can be treated fairly well within the Kane's or Luttinger–Kohn's models in so-called *kp* method (see e.g. Refs. [31, 32]). The SO interaction is given by the term in electron Hamiltonian

$$H_{SO} = \frac{\hbar}{4m_0^2c^2}\left[\vec{\nabla}\ U \times\ \vec{p}\right]\ \vec{\sigma}\,, \tag{3.13}$$

where $\vec{p} = -i\hbar\nabla$ the electron momentum operator.

SO interactions lead in some cases, as for semiconductor 2-D channels, to various effects that can be used to build electron interference devices, at least as a matter of principle. As an interesting (yet unrealized at the time of writing) example of such a

three-terminal spintronic device, it is worth describing Datta–Das ballistic spintronic modulator/switch [3]. This is a quantum interference device with FET-like layout where a 2-D electron gas (2DEG) has a ferromagnetic source and drain. The asymmetric confinement potential induces precession of spins injected into the 2DEG channel due to specific low-symmetry SO effect (Vasko–Rashba spin splitting) [29, 30]. The resulting angle may become large, $\sim\pi$ in channels $\gtrsim \mu$m long and made of narrow-gap semiconductors with strong so coupling. Since the ferromagnetic drain works as a spin filter, one hopes that changing the gate voltage would change the shape of the confinement potential and the Vasko–Rashba coupling constant α. As a result, one may be able to change the precession angle of ballistic electrons and the current through the structure (yet to be observed). To appreciate the situation, we need to describe the SO effects in a simple *kp*-model for semiconductors. As we shall see, the SO effects are expectedly weak, being of relativistic nature, and so in general are the effects. It is difficult to expect that SO-based devices can outperform any of the conventional electronics devices in standard applications.

In MOSFET structures the confining potential is asymmetric, so there appears an inversion asymmetry term derived by Vasko [28] (Vasko–Rashba or simply Rashba term, see also Ref. [29]). The only contribution coming from the confinement field in SO Hamiltonian is $\propto\langle\nabla_z V\rangle$ giving the Vasko–Rashba term,

$$H_{\mathrm{R}} = \alpha(k_y\sigma_x - k_x\sigma_y). \tag{3.14}$$

The magnitude of the coupling constant α depends on the confining potential, and this can in principle be modified by gating. It also defines the spin-precession wave vector $k_\alpha = \alpha m/\hbar^2$. Such a term, H_{R}, is also present in cubic systems with strain [35]. The Vasko–Rashba Hamiltonian for heavy holes is cubic in k, and, generally, very small.

Electric fields due to impurities (and external field) lead to extrinsic contributions of the spin-orbit coupling in the standard form

$$H_{\mathrm{ext}} = \lambda[\,\vec{k}\times\nabla U]\,\vec{\sigma}\,, \tag{3.15}$$

where U is the potential due to impurities and an externally applied field, with the coupling constant λ derived from 8×8 Kane Hamiltonian in third-order perturbation theory [33, 34]

$$\lambda = \frac{P^2}{3}\left[\frac{1}{E_g^2} - \frac{1}{(E_g+\Delta)^2}\right], \tag{3.16}$$

where P is the matrix element of momentum found from $\langle S|p_x|X\rangle = \langle S|p_y|Y\rangle = \langle S|p_z|Z\rangle = iPm_0/\hbar$. This is the same analytical form as the vacuum spin-orbit coupling but, for $\Delta>0$ the coupling has the *opposite sign*. Numerically, $\lambda = 5.3$ Å^2 for GaAs and 120 Å^2 for InAs, that is, spin-orbit coupling in *n*-GaAs is by *six orders* of magnitude larger than in vacuum [31]. This helps to generate the relatively large extrinsic spin currents observed in the spin-Hall effect (see below). In 2-D, $H_{\mathrm{ext}} = \lambda(\vec{k}\times\nabla U)_z\cdot\sigma_z$.

Now, we can analyze the behavior of polarized electrons injected into the 2DED channel. A free-electron VR Hamiltonian H_{VR} has two eigenstates with the momenta $k_{\pm}(\epsilon)$ for opposite spins for each energy ϵ, with $k_- - k_+ = 2m\alpha/\hbar^2$. Datta and Das [3] have noted that the conductivity of the device depends on the phase difference $\Delta\theta = (k_- - k_+)L = 2m\alpha L/\hbar^2$ between electron carriers after crossing a ballistic channel of a length L and oscillates with a period defined by the interference condition $(k_- - k_+)L = 2\pi n$, with n as integer. An equivalent description of the same phenomenon is the precession of an electron spin in an effective magnetic field $\vec{B}_{\mathrm{so}} = (2\alpha/g\mu_B)(\vec{k}\times\hat{\mathbf{z}})$, with μ_B the Bohr magneton, g the gyromagnetic ratio. The device is supposed to be controlled by the gate voltage V_g that modulates the SO coupling constant α, $\alpha = \alpha(V)_g$. This pioneering report generated much attention, yet to date the device appears not to have been demonstrated. Using typical parameters from Refs. [36, 37], $\hbar\alpha \approx 1 \times 10^{-11}\,\mathrm{eV}\cdot\mathrm{m}$ and $m = 0.1\,m_0$ for the effective carrier mass, current modulation would be observable in channels with relatively large length $L \gtrsim 1\,\mu\mathrm{m}$. Given the above, the observation of the effect would need: (i) efficient spin injection into channel from the FM source, which is tricky and requires a modified Schottky barrier (see below); (ii) the splitting should well exceed the bulk inversion asymmetry effect; (iii) the inhomogeneous broadening of α due to impurities, that mask the Vasko–Rashba splitting, should be small; and (iv) one should be able to gate control α. All these represent great challenges for building a room-temperature interference device, where one needs to use narrow-gap semiconductors and structures with *ballistic* channels. The device is not efficient is a diffusive regime. The gate control of α has been demonstrated (see Refs. [36, 38] and references therein).

3.3.1
Spin-Hall Effect (SHE) and Magnetoresistance due to Edge Spin Accumulation

Recently, there has been a resurgence of interest in the spin-Hall effect (SHE), which is another general consequence of the spin-orbital interaction, predicted by Dyakonov and Perel in 1971 [40]. These authors found that because of the spin-orbital interaction the electric and spin currents are intertwined: an electrical current produces a transverse spin current, and *vice versa*. In the case when impurity scattering dominates, which is quite often, the transverse is caused by the Mott skew scattering of spin-polarized carriers due to the SO interaction, [see Eq. (3.13)]. Since the current drags along the polarization of the carriers, the *spin accumulation at the edges* occurs, a so-called spin-Hall effect (SHE). In ferromagnets, the appearing Hall current is termed *anomalous*, and is always accompanied by the SHE. Importantly, the edge spin accumulation results in a slight *decrease* in sample resistance. External magnetic field would destroy the spin accumulation (Hanle effect) and lead to a *positive magnetoresistance*, recently identified by Dyakonov [41].

We present here a simple phenomenological description of spin-Hall effects (direct and inverse) and Dyakonov's magnetoresistance [42]. To this end, we introduce the electron charge flux $\vec{f}$ related to the current density as $\vec{J} = -q\vec{f}$, where q is

the elementary charge. For parts not related to the SO interaction, we have the usual drift-diffusion expression:

$$\vec{f}^{(0)} = -\mu n \vec{E} - D\vec{\nabla} n, \tag{3.17}$$

where μ and D are the usual electron mobility and diffusion coefficient, connected by the Einstein relation, $\vec{E}$ the electric field, and n is the electron density. The spin polarization flux t_{ij} is a tensor characterizing the flow of the jth component of the polarization *density* $\mathcal{P}_j = n_{j\uparrow} - n_{j\downarrow}$ in the direction i (spin density is $s_j(\mathbf{r}) = \frac{1}{2}\mathcal{P}_j$). It is non-zero even in the absence of spin–orbit interaction, simply because the spins are carried by electron flux, and we mark the corresponding quantity $t_{ij}^{(0)}$. Then, we have

$$t_{ij}^{(0)} = -\mu E_i \mathcal{P}_j - D\partial_i \mathcal{P}_j, \tag{3.18}$$

where $\partial_i = \partial/\partial x_i$, and one can add other sources of current, such as temperature gradient, in Equations 3.17 and 3.18. Spin–orbit interaction couples the charge and spin currents. For a material with an inversion symmetry, we have [42]:

$$f_i = f_i^{(0)} + \gamma \varepsilon_{ijk} t_{jk}^{(0)}, \tag{3.19}$$

$$t_{ij} = t_{jk}^{(0)} - \gamma \varepsilon_{ijk} f_k^{(0)}, \tag{3.20}$$

where ϵ_{ijk} is the unit antisymmetric tensor and $\gamma \ll 1$ is a dimensionless coupling constant proportional to the spin–orbit interaction λ [Equation 3.16] (typically, $\gamma \sim 10^{-2} - 10^{-3}$). The difference in signs in Equations 3.19 and 3.20 is consistent with the Onsager relations, and is due to the different properties of $\vec{f}$ and t_{ij} with respect to time inversion. Explicit phenomenological expressions for the two currents follow from Equations 3.17–3.20 [40, 41]:

$$\vec{J}/q = \mu n \vec{E} + D\vec{\nabla} n + \beta \vec{E} \times \vec{P} + \delta \vec{\nabla} \times \vec{\mathcal{P}}, \tag{3.21}$$

$$t_{ij} = -\mu E_i \mathcal{P}_j - D\partial_i \mathcal{P}_j + \varepsilon_{ijk}(\beta n E_k + \delta \partial_k n), \tag{3.22}$$

where the parameters $\beta = \gamma\mu$, $\delta = \gamma D$, satisfy the same Einstein relation, as do μ and D. The spin polarization vector evolves with time in accordance with the continuity equation [41, 42]:

$$\partial_t \mathcal{P}_j + \partial_i t_{ij} + |\vec{\Omega} \times \vec{\mathcal{P}}|_j + \mathcal{P}_j/\tau_s = 0, \tag{3.23}$$

where the vector $\Omega = g\mu_B H/\hbar$ is the spin precession frequency in the applied magnetic field $\vec{H}$, and τ_s the spin relaxation time. The term $\beta \vec{E} \times \vec{P}$ describes the *anomalous Hall effect*, where the spin polarization plays the role of the magnetic field. We ignore the action of magnetic field on the particle dynamics, which is justified if $\omega_c \tau \ll 1$, where ω_c is the cyclotron frequency and τ is the momentum relaxation time. Since normally $\tau_s \gg \tau$, it is possible to have both $\Omega\tau_s \gg 1$ and $\omega_c\tau \ll 1$ in a certain range of magnetic fields. It is also assumed that the equilibrium spin polarization in the applied magnetic field is negligible. The fluxes [Equations 3.21 and 3.22] need to be modified for an inhomogeneous magnetic field by adding a counter-term

proportional to $\partial B_j/\partial x_i$, which takes care of the force acting on the electron with a given spin in an inhomogeneous magnetic field $\vec{H}(r)$.

Equations 3.21–3.23 derived in Ref. [41] fully describe all physical consequences of spin–charge current coupling. For instance, the term $\delta\,\vec{\nabla}\times\vec{\mathcal{P}}$ describes an electrical current induced by an inhomogeneous spin density (so-called Inverse spin-Hall Effect) found experimentally for the first time by Bakun *et al.* [42] under the conditions of optical spin orientation. The term $\beta n\epsilon_{ijk}E_k$ (and its diffusive counterpart $\delta\epsilon_{ijk}\partial n/\partial x_k$) in Equation 3.22, describes what is now called the spin-Hall effect: an electrical current induces a transverse spin current, resulting in spin accumulation near the sample boundaries [41]. Recently, a spin-Hall effect was detected optically in thin films of n-doped GaAs and In GaAs [44] (with bulk electrons) and 2-D holes [45]. All of these phenomena are closely related and have their common origin in the coupling between spin and charge currents given by Equations 3.21 and 3.22. Any mechanism that produces the anomalous Hall effect will also lead to the spin-Hall effect, and vice versa. Remarkably, there is a single dimensionless parameter, γ, that governs the resulting physics.

It was found recently by Dyakonov that the spin-Hall effect is accompanied by a *positive* magnetoresistance due to spin accumulation near the sample boundaries [41]. The spin accumulation occurs over the spin diffusion length $L_s = \sqrt{D\tau_s}$. Therefore, it depends on the ratio L_s to the sample width L and vanishes in wide samples with $L_s/L \ll 1$. In a stripe sample with the width L (in xy plane), the z-component of $\mathcal{P}$ varies across the stripe (y-axis), $\vec{\nabla}\times\vec{\mathcal{P}} \neq 0$, this creates a positive correction to a current compared to a hypothetical case of an absent spin–orbit coupling, and the sample resistivity goes down. By applying the magnetic field in xy plane, one may destroy the spin polarization (the Hanle effect) and observe the *positive* (Dyakonov's) magnetoresistance in magnetic fields at the scale $\Omega\tau_s \sim 1$.

The data for 3-D [44] and 2-D [46] GaAs suggest the estimate $\gamma \sim 10^{-2}$, for platinum at room temperature [48] one finds $\gamma = 3.7\times10^{-3}$, so in these cases a magnetoresistance due to spin accumulation is on the order of 10^{-4} and 10^{-5}, respectively. It should be possible to find this MR due to its characteristic dependence on the field and the width of the sample, when it becomes comparable to the spin diffusion length. Because of the high sensitivity of electrical measurements, magnetoresistance might provide a useful tool for studying the spin–charge interplay in semiconductors and metals.

3.3.2
Interacting Spin Logic Circuits

There are various suggestions of more exoteric devices based on, for example, arrays of spin-polarized quantum dots with exchange coupled single spins with typical exchange coupling energy $\delta \sim 1$ meV [47], or magnetic quantum cellular automata [48]. It is assumed that one can apply a local field or short magnetic π-pulse to flip the "input" spin that would result in nearest-neighbor spins to flip in accordance with the new ground state (of the antiferromagnetically coupled circuit of quantum dots). The idea is that those spin arrays (no quantum coherence is required) can be used to

perform classical logic on bits represented by spins pointing along or against the quantizations axis, $|+z\rangle \to 1$, $|-z\rangle \to 0$. However, there are problems with using those schemes. Indeed, the standard Zeeman splitting for electron in the field of 1 T is only 0.5 K *in vacuo*, so that one needs the field of *at least* ∼150 T to flip the spin (or use materials with unusually large gyromagnetic factors), or one can apply ∼1 T transversal *B*-field for some 30 ps to do the same. The practicalities of building such a control system at a nanoscale is a major challenge, and would require a steep power budget. The other challenge is that instead of *nearest-neighbor* spins falling into a new shallow ground state with the directions of all other spins fixed, the initial flip would trigger spin wave(s) in the circuit, thus destroying the initial set-up. Indeed, the spin wave spectrum in large coupled arrays of N spins is almost gapless, with the excited state just $\sim\delta/N$ above the ground state (see e.g., Ref. [49]). Additionally, the spins are subject to a fluctuating external (effective) magnetic field that tends to excite the spin waves and destroy the direction of the spins along set quantization axis $\pm z$. For the same reason, keeping the spins in a *coherent* superposition state is unlikely, so quantum computing with coupled spins is even less practical [50].

It is clear from the above discussion, however, that it is unlikely that the Datta–Das or any other interference devices can offer any advantages over standard MOSFETs, especially as they do not have any *gain*, should operate in a ballistic regime (i.e., at low temperatures in clean systems), and require new fabrication technology.

3.4 Tunnel Magnetoresistance

Here, we describe some important aspects of TMR on the basis of simple microscopic model for elastic, impurity-assisted, surface state-assisted and inelastic contributions. Most of these results are generic, and some will be useful later to analyze room-temperature spin injection into semiconductor through a modified Schottky barrier. A model for spin tunneling has been formulated by Julliere [12], and further developed in Refs. [17, 18, 21]. It is expected to work rather well for iron-, cobalt-, and nickel-based metals, according to theoretical analysis [21] and experiments [15]. However, it disregards important points such as an impurity scattering and a reduced effective mass of carriers inside the barrier. Both issues have important implications for magnetoresistance and will be considered here, along with proposed novel half-metallic systems which should, in principle, show the ultimate performance. Enhanced performance is also found in MTJ with MgO epitaxial oxide barrier, which may be a combination of band-structure and surface effects [16, 51]. In particular, Zhang and Butler [51] predicted a very large TMR in Fe/MgO/Fe, bcc Co/MgO/Co, and FeCo/MgO/FeCo tunnel junctions, having to do with peculiar band matching for majority spin states in a metal with that in MgO tunnel barrier.

We shall describe electrons in ferromagnet-insulating barrier-ferromagnet (FM–I-FM) systems by the Schrödinger equation [17] $[-(\hbar^2/2m_i)\nabla^2 + U_i - \frac{1}{2}\vec{\Delta}_{xc}\vec{\sigma}]\psi = E\psi$ with $U(\mathbf{r})$ the potential (barrier) energy, $\vec{\Delta}_{xc}(\mathbf{r})$ the exchange splitting of, for example,

d-states in 3d ferromagnets (=0 inside the barrier), $\vec{\sigma}$ stands for the Pauli matrices; index $i = 1(3)$ for left (right) ferromagnetic electrode FM1(2) and $i = 2$ for tunneling barrier (quantities for the tunnel barrier also marked t), respectively.

We start with the expression for a direct tunnel current density of spin σ from FM1 to FM2 [52]

$$J_{\sigma 0} = \frac{q}{h}\int dE[f(E - F_{\sigma 0}^{\mathrm{FM2}}) - f(E - F_{\sigma 0}^{\mathrm{FM1}})]\int \frac{d^2 k_{\parallel}}{(2\pi)^2} T_{\sigma}(E,\ \vec{k}_{\parallel}), \tag{3.24}$$

where $f(x)$ is the Fermi–Dirac distribution function with local Fermi level $F_{\sigma 0}^{\mathrm{FM1(2)}}$ for ferromagnetic electrode FM1(2), $T_{\sigma} = \sum_{\sigma'} T_{\sigma\sigma'}$ the transmission probability from majority (minority) spin subband in FM1 $\sigma = \uparrow$ or $\downarrow$ into majority (minority) spin subband in FM2, $\sigma' = \uparrow (\downarrow)$. It has a particularly simple form for a square barrier and *collinear* [parallel (P) or antiparallel (AP)] moments on electrodes:

$$T_{\sigma\sigma'} = \frac{16 m_1 m_3 m_2^2 k_1 k_3 \kappa^2}{(m_2^2 k_1^2 + m_1^2\kappa^2)(m_2^2 k_3^2 + m_3^2\kappa^2)} e^{-2\kappa w}, \tag{3.25}$$

where κ is the attenuation constant for the wavefunction in the barrier $k_1 \equiv k_{1\sigma}$, $k_2 = i\kappa$, $k_3 \equiv k_{3\sigma'}$ are the momenta normal to the barrier for the corresponding spin subbands, w is the barrier width, and we have used a limit of T at $\kappa w \gg 1$ [18]. With the use of Equations 3.17 and 3.18, and accounting for the misalignment of magnetic moments in ferromagnetic terminals (given by the mutual angle θ), we obtain following expression for the junction conductance per unit area, assuming $m_1 = m_3$,

$$G = G_0(1 + P_1 P_2 \cos\theta), \tag{3.26}$$

$$\begin{aligned} P_{1(2)} &= \frac{k_{\uparrow} - k_{\downarrow}}{k_{\uparrow} + k_{\downarrow}} \frac{\kappa^2 - m_2^2 k_{\uparrow} k_{\downarrow}}{\kappa^2 + m_2^2 k_{\uparrow} k_{\downarrow}} \\ &= \frac{(v_{\uparrow} - v_{\downarrow})(v_t^2 - v_{\uparrow} v_{\downarrow})}{(v_{\uparrow} + v_{\downarrow})(v_t^2 + v_{\uparrow} v_{\downarrow})}, \end{aligned} \tag{3.27}$$

where $P_{1(2)}$ is the effective polarization of the FM1(2) electrode, $\kappa = [2m_2(U_0 - E)/\hbar^2]^{1/2}$, and U_0 is the top of the barrier. Equation 3.26 correct an expression derived earlier [17] for the effective mass of the carriers in the barrier. To obtain the last simple expression for tunnel current polarization Equation 3.27, which has exactly the same form also for FM-semiconductor modified Schottky junctions (below), we have introduced the carrier band velocities $v_{\uparrow(\downarrow)} = \hbar k_{\uparrow(\downarrow)}/m$, that are generally different for FM1(2), and "tunneling" velocity $v_t = \hbar\kappa/m_2$. These relations between the velocity an momentum are equivalent to an *effective mass* approximation.

By taking a typical value of $G_0 = 4$–$5\ \Omega^{-1}\,\mathrm{cm}^{-2}$ (Ref. [15] $k_{\uparrow} = 1.09\ \text{Å}^{-1}$, $k_{\downarrow} = 0.42\ \text{Å}^{-1}$, $m_1 \approx 1$ (for itinerant d electrons in Fe) [21] and a typical barrier height for Al_2O_3 (measured from the Fermi level μ) $\phi = U_0 - \mu \approx 3$ eV, and the thickness $w \approx 20$ Å, one arrives at the following estimate for the effective mass in the barrier: $m_2 \approx 0.4$ [53]. These values give $P_{\mathrm{Fe}} = 0.28$, which is noticeably smaller than the experimental value of 0.4–0.5 (note that neglect of the mass correction, $m_2 < 1$, as in Ref. [17], would give a negative value of the effective polarization). Below, we shall see that tunneling

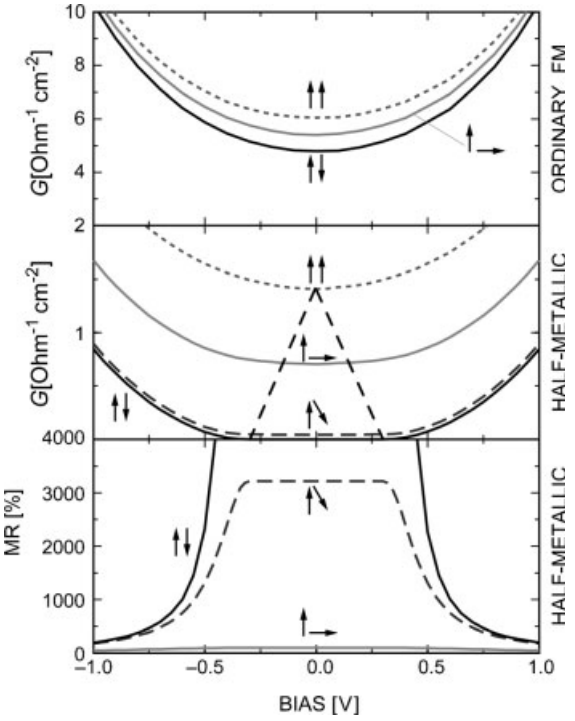


Figure 3.4 Conductance and magnetoresistance of tunnel junctions versus bias. Top panel: conventional (Fe-based) tunnel junction (for parameters, see text). Middle panel: half-metallic electrodes. Bottom panel: magnetoresistance for the half-metallic electrodes. The dashed line shows schematically a region where a gap in the minority spin states is controlling the transport. Even for imperfect antiparallel alignment ($\theta = 160^\circ$, marked $\uparrow\searrow$), the magnetoresistance for half-metallics (bottom panel) exceeds 3000% at biases below the threshold V_c. All calculations have been performed at 300 K, with inclusion of multiple image potential and exact transmission coefficients. Parameters are described in the text.

assisted by polarized surface states may lead to much larger TMR, this may relevant to observed large values of TMR [19].

The most striking feature of Equation 3.26 is that MR tends to infinity for vanishing $k_\downarrow$; that is when the electrodes are made of a 100% spin-polarized material ($P = P' = 1$) because of a gap in the density of states (DOS) for minority carriers up to their conduction band minimum $E_{\mathrm{CB}\downarrow}$. Then G^{AP} vanishes together with the transmission probability Equation 3.25, as there is a zero DOS at $E = \mu$ for both spin directions. Such a half-metallic behavior is rare, but some materials possess this amazing property, most interestingly the oxides CrO_2 and Fe_3O_4 (e.g., see recent discussion in Ref. [2]). These oxides are very interesting for future applications in combination with matching materials, as will be seen below.

Remarkably, for $|V| < V_c$ in the AP geometry one has $MR = \infty$. From the middle and the bottom panels in Figure 3.4 we see that even at 20° deviation from the AP configuration, the value of MR exceeds 3000% in the interval $|V| < V_c$, and this is indeed a very large value.

3.4.1
Impurity Suppression of TMR

An important aspect of spin-tunneling is the effect of tunneling through the defect states in the (amorphous) oxide barrier. Dangling bonds and random trap states may play the role of defects in an amorphous barrier. Since the contacts under consideration are typically short, their current–voltage (I–V) curve and MR should be very sensitive to defect resonant states in the barrier with energies close to the Fermi level, forming "channels" with the nearly periodic positions of impurities. Generally, channels with one impurity (most likely to dominate in thin barriers) would result in a monotonous behavior of the I–V curve, whereas channels with two or more impurities would produce intervals with negative differential conductance. Impurity-assisted spin tunneling at zero temperature [the general case of non-zero temperature would require integration with the Fermi factors as in Equation 3.24] can be written in the standard form [54]:

$$G_\sigma = \frac{2e^2}{\pi\hbar}\sum_i \frac{\Gamma_{L\sigma}\Gamma_{R\sigma}}{(E_i-\mu)^2+\Gamma^2}, \tag{3.28}$$

where $\Gamma_\sigma = \Gamma_{L\sigma} + \Gamma_{R\sigma}$ is the total width of a resonance given by a sum of the partial widths $\Gamma_{L(R)}$ corresponding to electron tunneling from the impurity state at the energy E_i to the left (right) terminal. It is easiest to analyze the case of parallel (P) and antiparallel (AP) mutual orientation of magnetic moments M_1 and M_2 on electrodes with the angle θ between them. In this case, one looks at tunneling of majority (maj) and minority (min) carriers from the left electrode $L_\sigma = (L_{maj}, L_{min})$ into states $R_\sigma = (R_{maj}, R_{min})$ for parallel orientation ($\theta = 0$) or $R_\sigma = (R_{min}, R_{maj})$ in antiparallel orientation ($\theta = \pi$), respectively. The general case is then easily obtained from standard spinor algebra for spin projections. The tunnel widths can be evaluated analytically for a rectangular barrier, $\Gamma_{L\sigma} \sim g_{L\sigma}\Omega\exp[-\kappa(w+2z_i)]$, where z_i is the coordinate of the impurity with respect to the center of the barrier, $g_{L\sigma}$ the density of states in the (left) electode, Ω [18].

The resonant conductance Equation 3.28 has a sharp maximum [$= e^2/(2\pi\hbar)$] when $\mu = E_i$ and $\Gamma_L = \Gamma_R$, that is for the symmetric position of the impurity in the barrier for parallel configuration. For antiparallel configuration, most effective impurities will be positioned somewhat off-center since the DOS for the majority and minority spins may be quite different. An asymmetric position of effective impurities in the AP orientation immediately suggests smaller conductance G_{AP} than G_P and *positive* ("normal") impurity TMR > 0. This result is confirmed by direct calculation. Indeed, if we assume that we have ν defect/localized levels in a unit volume and unit energy interval in a barrier, then, replacing the sum by an integral in Equation 3.28, and considering a general configuration of the magnetic moments on terminals, we obtain the following formula for impurity-assisted conductance per unit area in leading order in $\exp(-\kappa w)$:

$$G_1 = g_1(1 + \Pi_L\Pi_R\cos\theta), \tag{3.29}$$

where we have introduced the quantities

$$g_1 = \frac{e^2}{\pi\hbar} N_1, \quad N_1 = \pi^2 \nu \Gamma_1/\kappa,$$

$$\Pi_{L(R)} = \frac{(r_\uparrow - r_\downarrow)}{(r_\uparrow + r_\downarrow)}, \tag{3.30}$$

N_1 being the effective number of one-impurity channels per unit area, and one may call Π_F a "polarization" of the impurity channels, defined by the factor $r_\sigma = [m_2 \kappa k_\sigma/(\kappa^2 + m_2^2 k_\sigma^2)]^{1/2}$ with momenta k_σ for left (right) [L (R)] electrode.

Comparing direct and impurity-assisted contributions to conductance, we see that the latter dominates when the density of localized states $\nu \gtrsim (\kappa/\pi)^3 \epsilon_i^{-1} \exp(-\kappa w)$, and in our example a crossover takes place at the density of localized states $\nu \gtrsim 10^{17}\,\mathrm{eV}^{-1}\,\mathrm{cm}^{-3}$. When resonant transmission dominates, the magnetoresistance will be given by

$$MR_1 = \frac{2\Pi_L \Pi_R}{(1 - \Pi\Pi')}, \tag{3.31}$$

which is only 4% in the case of Fe. We see that indeed MR_1 is suppressed yet remains positive (unless the polarization of tunnel carriers is opposite to the magnetization direction on one of the electrodes, in this case MR is obviously inverted for trivial reasons). There are speculations about a possibility of negative MR_1, which is analyzed below in the following subsection.

We have estimated the above critical DOS for localized states for the case of Al_2O_3 barrier, in systems such as amorphous Si the density of localized states is higher because of considerably smaller band gap, and estimated as $8 \times 10^{18}\,\mathrm{eV}^{-1}\,\mathrm{cm}^{-3}$, mainly due to dangling bonds and band edge smearing because of disorder [55]. One can appreciate that in junctions with thin Al_2O_3 amorphous barriers (<20–25 Å) of practical interest the impurity-assisted tunneling is not the major effect, so the above consideration of elastic tunneling applies. In this seminal work, Beasley have studied a-Si barriers with wide variety of thicknesses $w = 30$–1000 Å and obtained detailed data on crossover from direct tunneling to directed inelastic hopping along statistically rare, yet highly conductive, chains of localized states. The crossover thicknesses depend heavily on the materials parameters of the barrier. The above-described suppression of TMR by impurities was confirmed experimentally for magnetic tunnel junctions in Ref. [56].

3.4.2 Negative Resonant TMR?

It should be noted that the MR becomes suppressed yet remains positive. Indeed, the conductance is dominated, but can it change sign, or become *inverted* in the case of impurity assisted tunneling? It was shown above that the asymetry of polarized DOS in contacts gives *positive* resonant tunneling TMR. Negative (inverse MR_1) can only appear if the dominating impurity levels were lined up with the Fermi level [Equation 3.28] and positioned in asymmetric positions in the barrier, with for

example, the "right" width of the resonance much larger than the "left" width, $\Gamma_{R\sigma} \equiv \Gamma_\sigma \gg \Gamma_{L\sigma} \equiv \gamma_\sigma$

$$G_\sigma \approx \frac{e^2}{\pi\hbar}\frac{\gamma_\sigma}{\Gamma_\sigma},$$

$$\text{TMR} \sim -P_1 P_2, \ (?) \tag{3.32}$$

the latter was noted in Ref. [57]. However, the required coincidence is statistically very unlikely in tunnel junctions, where the number of impurity states involved is $\gg 1$, as in all usual situations with a possible exception of very small area tunnel junctions. Indeed, the data in Ref. [57] suggest that, in a tiny percentage of small area junctions, the TMR is negative. The attempts to simulate the amorphous barrier that may produce such a result in resonant tunneling regime showed, however, that one needs an unphysically large amount of disorder in the barrier to obtain traces of negative TMR. Indeed, for the barrier with height $U = 1.5$ eV, an unphysical amount of onsite disorder $\gamma = 4U = 6$ eV should be assumed. It must be concluded that the speculations about negative resonant TMR in Ref. [57] have nothing to do with most observations of inverse TMR. Averaging over disorder suppresses TMR, as predicted in Ref. [18] and observed in for example Ref. [56]. It is noted, however, that is the case when the impurity states are located at a particular interface in the barrier, perhaps as in tunnel junctions with composite barriers MgO/NiO [58], there may be a suppression and a slight inversion of TMR in a certain window of bias voltages, given by the energy interval occupied by the interfacial states, as described elsewhere [59].

3.4.3
Tunneling in Half-Metallic Ferromagnetic Junctions

Now we shall discuss a couple of systems with half-metallic behavior, CrO_2/TiO_2 and CrO_2/RuO_2 (Figure 3.5). These are based on half-metallic CrO_2, and all species have the rutile structure type with almost perfect lattice matching, which should yield a good interface and should help in keeping the system at the desired stoichiometry. TiO_2 and RuO_2 are used as the barrier/spacer oxides. The electronic structure of CrO_2/TiO_2 is truly stunning in that it has a half-metallic gap which is 2.6 eV wide and extends on both sides of the Fermi level, where there is a gap either in the minority *or* majority spin band. Thus, a huge magnetoresistance should, in principle, be seen not only for electrons at the Fermi level biased up to 0.5 eV, but also for *hot* electrons starting at about 0.5 eV above the Fermi level. We note that states at the Fermi level are a mixture of Cr(d) and O($2p$) states, so that p–d interaction within the first coordination shell produces a strong hybridization gap, and the Stoner spin-splitting moves the Fermi level right into the gap for minority carriers (Figure 3.5). It is worth noting that CrO_2 and RuO_2 are very similar in terms of a paramagnetic band structure, but the difference in the number of conduction electrons and exchange splitting results in a usual metallic behavior of RuO_2 as compared to the half-metallic ferromagnet CrO_2.

An important difference between two spacer oxides is that TiO_2 is an insulator whereas RuO_2 is a good metallic conductor. Thus, the former system can be used in a

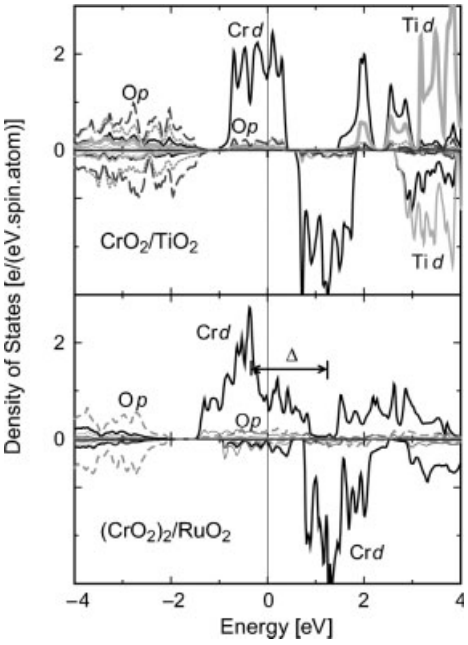


Figure 3.5 Density of states of CrO_2/TiO_2 (top panel) and $(CrO_2)_2/RuO_2$ (bottom panel) half-metallic multilayers calculated with the use of the LMTO method. The partial contributions are indicated by letters. The zero of energy corresponds to the Fermi level. Δ indicates a spin-splitting of the Cr *d* band near E_F (schematic). Note a strong hybridization of Cr *d* with O 2*p* states at E_F and below the hybridization gap. Growth direction is [001].

tunnel junction, whereas the latter will form a metallic multilayer. In the latter case the physics of conduction is different from tunnelling, but the effect of vanishing phase volume for transmitted states still works when current is passed through such a system *perpendicular to planes*. One interesting possibility is to form three-terminal devices with these systems, like a spin-valve transistor [60], and check the effect in a hot-electron region. CrO_2/TiO_2 seems to a be a natural candidate to check the present predictions about half-metallic behavior and for a possible record tunnel magnetoresistance. One important advantage of these systems is an almost perfect lattice match at the oxide interfaces. The absence of such a match of the conventional Al_2O_3 barrier with Heussler half-metallics (NiMnSb and PtMnSb) may have been among other reasons for their unimpressive performance [2]. The main concerns for achieving a very large value of magnetoresistance will be spin-flip centers, magnon-assisted events, and imperfect alignment of moments. As for conventional tunnel junctions, the present results show that presence of defect states in the barrier, or a resonant state, as in a resonant tunnel diode-type of structure, reduces their magnetoresistance several fold but may dramatically increase the current through the structure.

3.4.4
Surface States Assisted TMR

Direct tunneling, as we have seen, gives a TMR of about 30%, whereas in recent experiments TMR is well above this value, approaching 40–50% in systems with Al_2O_3 amorphous barrier, and 200% in systems with epitaxial MgO barriers [16, 52]. It will become clear below, that this enhancement is unlikely to come from the inelastic processes. Until now, we have disregarded the possibility of localized states at metal–oxide interfaces. Bearing in mind that the usual barrier AlO_x is amorphous, the density of such surface states may be high, and we must take into account tunneling into/from those states. The results for Tamm states that may exist at clean interfaces, are similar. The corresponding tunneling conductance per unit area is [19]:

$$G_s(\theta) = \frac{e^2}{\pi\hbar} B\bar{D}_s(1 + P_F P_s \cos\theta),$$

$$P_s = \frac{D_{s\uparrow} - D_{s\downarrow}}{D_{s\uparrow} + D_{s\downarrow}}, \quad \bar{D}_s = \frac{1}{2}(D_{s\uparrow} + D_{s\downarrow}), \tag{3.33}$$

where P_s is the polarization and $\bar{D}s$ is the average density of surface states, and θ is the mutual angle between moments on electrodes. The parameter $B \sim [2\pi\hbar^2 m\kappa/(m_2^2 w)]\exp(-2\kappa w)$, where w is the barrier width, κ is the absolute value of electron momentum under the barrier, m and m_2 are the free electron mass and the effective mass in the barrier, respectively. The corresponding magnetoresistance would be $MR_s = 2P_F P_s/(1 - P_F P_s)$. It is easy to show that the bulk-to-surface conductance exceeds the bulk-to-bulk one at densities of surface states $D_s > D_{sc} \sim 10^{13}\,\mathrm{cm}^{-2}\,\mathrm{eV}^{-1}$ per spin, comparable to those found at some metal–semiconductor interfaces. Since this result was obtained, various groups confirmed this with *ab-initio* calculations, usually without mentioning the original result obtained in Ref. [19].

If on both sides of the barrier the density of surface states is above the critical value D_{sc}, the magnetoresistance would be due to surface-to-surface tunneling with a value given by $MR_{ss} = 2P_{s1}P_{s2}/(1 - P_{s1}P_{s2})$. If the polarization of surface states is larger than that of the bulk, as is often the case even for imperfect surfaces [61], then it would result in enhanced TMR.

3.4.5
Inelastic Effects in TMR

Inelastic processes with excitation of magnon or phonon modes introduce new energy scales into the problem (30–100 meV) which correspond to a region where unusual I–V tunnel characteristics are seen (Figure 3.6). One can describe their effect on TMR fairly well within the tunnel Hamiltonian approach [19]. We obtain for magnon-assisted inelastic tunneling current at $T = 0$ with the use of tunnel Hamiltonian formalism [62]:

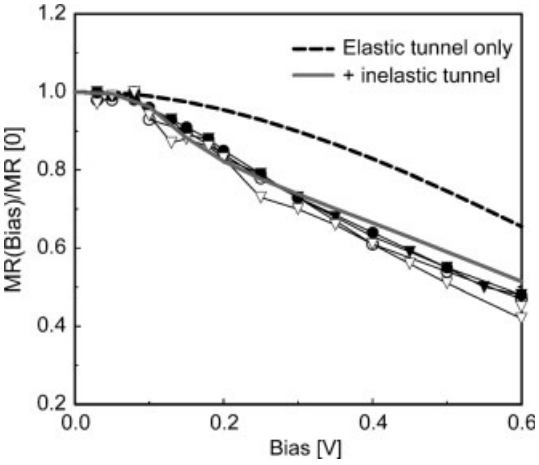


Figure 3.6 Fit to experimental data for the magnetoresistance of $CoFe/Al_2O_3/NiFe$ tunnel junctions [9] with inclusion of elastic and inelastic (magnons and phonons) tunneling. The fit gives for magnon DOS $\propto \omega^{0.65}$, which is close to a standard bulk spectrum $\propto \omega^{1/2}$.

$$
\begin{aligned}
I^x_{\mathrm{P}} &= \frac{2\pi e}{\hbar}\sum_{\alpha} X^{\alpha} g^L_{\downarrow} g^R_{\uparrow} \int d\omega \rho^{\mathrm{mag}}_{\alpha}(\omega)(eV-\omega)\theta(eV-\omega),\\
I^x_{\mathrm{AP}} &= \frac{2\pi e}{\hbar}\left[X^R g^L_{\uparrow} g^R_{\uparrow} \int d\omega \rho^{\mathrm{mag}}_{\mathrm{R}}(\omega)(eV-\omega)\theta(eV-\omega)\right.\\
&\quad \left. + X^L g^L_{\downarrow} g^R_{\downarrow} \int d\omega \rho^{\mathrm{mag}}_{\mathrm{L}}(\omega)(eV-\omega)\theta(eV-\omega)\right],
\end{aligned}
\tag{3.34}
$$

where X is the incoherent tunnel exchange vertex, $\rho^{\mathrm{mag}}_{\alpha}(\omega)$ is the magnon density of states that has a general form $\rho^{\mathrm{mag}}_{\alpha}(\omega) = (\nu+1)\omega^{\nu}/\omega_0^{\nu+1}$, the exponent ν depends on a type of spectrum, ω_0 is the maximum magnon frequency, $g_{\mathrm{L\ (R)}}$ marks the corresponding electron density of states on left (right) electrode, $\theta(x)$ is the step function, $\alpha = L,R$. The analogous expressions can be written down for phonon contribution with the important distinction that electron–phonon interaction does *not* affect spin (if one ignores any small magnetoelastic contribution). The elastic and inelastic contributions together will define the total junction conductance $G = G(V,T)$ as a function of the bias V and temperature T. We find that the inelastic contributions from magnons Equation 3.34 and phonons grow as $G^x(V,0) \propto (|eV|/\omega_0)^{\nu+1}$ and $G^{ph}(V,0) \propto (eV/\omega_D)^4$ at low biases. These contributions saturate at higher biases: $G^x(V,0) \propto 1 - \frac{i+1}{\nu+2}\frac{\omega_0}{|eV|}$ at $|eV| > \omega_0$; $G^{ph}(V,0) \propto 1 - \frac{4}{5}\frac{\omega_D}{|eV|}$ at $|eV| > \omega_D$. This behavior would lead to sharp features in the I–V curves on a scale of 30–100 mV (Figure 3.6).

It is important to highlight the opposite effects of phonons and magnons on the TMR. If we take the case of the same electrode materials and denote $D = g_{\uparrow}$ and $d = g_{\downarrow}$,

then we see that $G^{x}_{\mathrm{P}}(V,0) - G^{x}_{\mathrm{AP}}(V,0) \propto -(D-d)^2(|eV|/\omega_0)^{\nu+1} < 0$, whereas $G^{ph}_{\mathrm{P}}(V,0) - G^{ph}_{\mathrm{AP}}(V,0) \propto +(D-d)^2(eV/\omega_D)^4 > 0$; that is, spin-mixing due to magnons *kills* the TMR, whereas the phonons tend to reduce the negative effect of magnon emission [63]. Different bias and temperature dependence can make possible a separation of these two contributions, which are of opposite sign. At finite temperatures we obtain the contributions of the same respective sign as above. For magnons: $G^{x}_{\mathrm{P}}(0,T) - G^{x}_{\mathrm{AP}}(0,T) \propto -(D-d)^2(-TdM/dT) < 0$, where $M = M(T)$ is the magnetic moment of the electrode at a given temperature T. The phonon contribution is given by a standard Debye integral with the following results: $G^{ph}_{\mathrm{P}}(0,T) - G^{ph}_{\mathrm{AP}}(0,T) \propto +(D-d)^2(T/\omega_D)^4 > 0$ at $T \ll \omega_D$, and $G^{ph}_{\mathrm{P}}(0,T) - G^{ph}_{\mathrm{AP}}(0,T) \propto +(D-d)^2(T/\omega_D)$ at $T \gtrsim \omega_D$. It is worth mentioning that the magnon excitations are usually cut off, for example, by the anisotropy energy K_{an} at some ω_c. Therefore, at low temperatures the conductance at small biases will be almost constant. We conclude that the inelastic processes are responsible for TMR diminishing with bias voltage, the unusual shape of the I–V curves at low biases, and their temperature behavior, which is also affected by impurity-assisted tunneling. The surface states-assisted tunneling may lead to enhanced TMR, if their polarization is higher than that of the bulk. This could open up ways to improving performance of ferromagnetic tunnel junctions.

3.5
Spin Injection/Extraction into (from) Semiconductors

Much attention has been devoted recently to exploring the possibility of a three-terminal spin injection device where spin is injected into semiconductor from either metallic ferromagnetic electrode [67, 68], or from magnetic semiconductor electrode, as demonstrated in Ref. [64]. However, the magnetization in FMS usually vanishes or is too small at room temperature. Relatively high spin injection from ferromagnets (FM) into non-magnetic semiconductors (S) has recently been demonstrated at low temperatures [65], and the attempts to achieve an efficient room-temperature spin injection have faced substantial difficulties [66]. Theoretical studies of the spin injection from ferromagnetic metals, as initiated in Refs. [68, 69], have been subject of extensive research in Refs. [5–10, 69–77] that has gained much insight into the problem of spin injection/accumulation in semiconductors. As a consequence, some suggestions for *spin transistors* and other spintronic devices have appeared that are experimentally realizable, can work at room temperatures, and exceed the parameters of standard semiconductor analogues [5, 6].

As an important distinction with spin transport in magnetic tunnel junctions, one would like to create non-equilibrium spin polarization and manipulate it with external fields in semiconductors, with a possible advantage of long spin relaxation time in comparison with mean collision time. In order to be interesting for applications, there should be a straightforward method of creating substantial non-equilibrium spin polarization *density* in semiconductor. This is different from

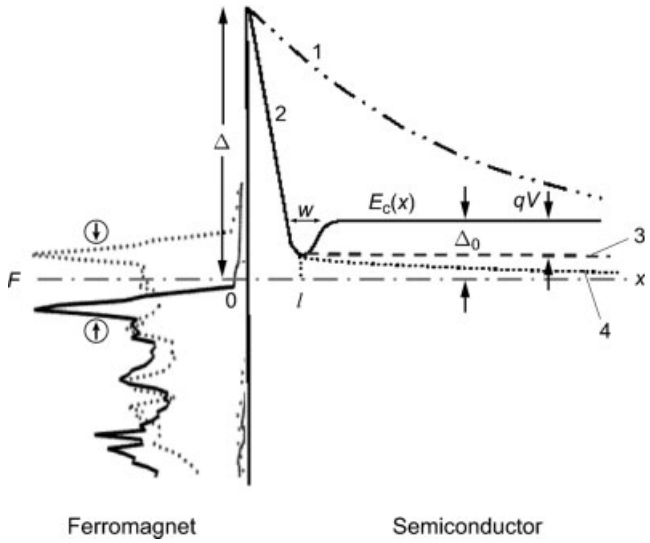


Figure 3.7 Energy diagrams of ferromagnet-semiconductor heterostructure with δ-doped layer (F is the Fermi level; Δ the height and l the thickness of an interface potential barrier; Δ_0 the height of the thermionic barrier in n-semiconductor). The standard Schottky barrier (curve 1); $E_c(x)$ the bottom of conduction band in n-semiconductor in equilibrium (curve 2), under small (curve 3), and large (curve 4) bias voltage. The spin-polarized density of states in Ni is shown at $x < 0$.

tunnel junctions, where one is interested in large spin injection efficiency; that is, in large resistance change with respect to magnetic configuration of the electrodes. Obviously, a spin imbalance in the drain ferromagnetic electrode is created in MTJ, proportional to the current density, but relatively minute, given a huge density of carriers in a metal. The principal difficulty of *spin injection* in semiconductor from ferromagnet is that the materials in the FM-S junctions usually have very different electron affinity and, therefore, high Schottky barrier forms at the interface [78] (Figure 3.7, curve 1). Thus, in GaAs and Si the barrier height $\Delta \simeq 0.5$–0.8 eV with practically all metals, including Fe, Ni, and Co [65, 78], and the barrier width is large, $l \gtrsim 100$ nm for doping concentration $N_d \lesssim 10^{17}\,\mathrm{cm}^{-3}$. The spin-injection corresponds to a reverse current in the Schottky contact, which is saturated and usually negligible due to such large l and Δ [78]. Therefore, a thin heavily doped $n^+ - \mathrm{S}$ layer between FM metal and S is used to increase the reverse current [78] determining the spin-injection [6, 8, 65, 72]. This layer sharply reduces the thickness of the barrier, and increases its tunneling transparency [6, 78]. Thus, a substantial spin injection has been observed in FM-S junctions with a thin n^+–layer [65].

One usually overlooked formal paradox of spin injection is that a current through Schottky junctions (as derived in textbooks) depends solely on parameters of a semiconductor [78], and cannot formally be spin-polarized. Some authors even

emphasize that in Schottky junctions "spin-dependent effects do not occur" [70]. In earlier reports [67–76], spin transport through FM-S junction, its spin-selective properties, and non-linear I–V characteristics have not been actually calculated. They were described by various, often contradictory, boundary conditions at the FM-S interface. For example, Aronov and Pikus assumed that *spin injection efficiency* of a FM-S interface is a constant equal to that in the ferromagnet FM, $\Gamma = \Gamma_F$, [Equation 3.3], and then studied non-linear spin accumulation in S considering spin diffusion and drift in an electric field [67]. The authors of Refs. [68–72] assumed a continuity of both the currents and the electrochemical potentials for both spins, and found that a spin polarization of injected electrons depends on a ratio of conductivities of a FM and S (the so-called "conductivity mismatch" problem). At the same time, it has been asserted in Refs. [73–76] that the spin injection becomes appreciable when the electrochemical potentials have a substantial discontinuity (produced e.g., by a tunnel barrier [74]). The effect, however, was described by the unknown spin-selective interface conductances $G_{i\sigma}$, which cannot be found within those theories.

We have developed a microscopic theory of the spin transport through ferromagnet-semiconductor junctions, which include an ultrathin heavily doped semiconductor layer (δ-doped layer) between FM and S [6, 8]. We have studied the non-linear effects of spin accumulation in S near reverse-biased modified FM-S junctions with the δ-doped layer [6] and spin extraction from S near the modified forward-biased FM-S junctions [8]. We found conditions for the most efficient spin injection, which are *opposite* to the results of previous phenomenological theories. We show, in particular, that: (i) the current of the FM-S junction does depend on spin parameters of the ferromagnetic metal but *not* its conductivity, so, contrary to the results [68–72, 74–76], the "conductivity mismatch" problem *does not exist* for the Schottky FM-S junctions. We find also that: (ii) a spin injection efficiency Γ of the FM-S junction depends strongly on the current, contrary to the assumptions in [67–72, 74–76]; and (iii) the highest spin polarization of both the injected electrons P and spin injection efficiency can be realized at room temperatures and relatively small currents in *high-resistance* semiconductors, *contrary* to claims in Ref. [71], which are of most interest for spin injection devices [3, 4, 6]. We also show that: (iv) tunneling resistance of the FM-S junction must be relatively small, which is *opposite* to the condition obtained in linear approximation in Ref. [74]; and that (v) the spin-selective interface conductances $G_{i\sigma}$ are not constants, as was assumed in Ref. [73–76], but vary with a current J in a strongly non-linear fashion. We have suggested a new class of spin devices on the basis of the present theory.

Below, we describe a general theory of spin current, spin injection and extraction in Section 3.5.1, followed by the discussion of the conditions of an efficient spin injection and extraction in Section 3.5.2. Further, we turn to the discussion of high-frequency spin valve effect in a system with two δ-doped Schottky junctions in Section 3.5.3 A new class of spin devices is detailed in Section 3.5.3 field detector, spin transistor, and square-law detector. The efficient spin injection and extraction may be a basis for efficient sources of (modulated) polarized radiation, as mentioned in Section 3.5.4.

3.5.1
Spin Tunneling through Modified (Delta-Doped) Schottky Barrier

The modified FM-S junction with transparent Schottky barrier is produced by δ-doping the interface by sequential donor and acceptor doping. The Schottky barrier is made very thin by using large donor doping N_d^+ in a thin layer of thickness l. For reasons which will become clear shortly, we would like to have a narrow spike followed by the narrow potential well with in the width w and the depth $\sim rT$, where T is the temperature in units of $k_B = 1$ and $r \sim 2$–3, produced by an acceptor doping N_a^+ of the layer w (Figure 3.7). Here $l \lesssim l_0$, where $l_0 = \sqrt{\hbar^2/[2m_*(\Delta - \Delta_0)]}$ ($l_0 \lesssim 2$ nm), the remaining low (and wide) barrier will have the height $\Delta_0 = (E_{c0} - F) > 0$, where E_{c0} is the bottom of the conduction band in S in equilibrium, q the elementary charge, and ϵ (ϵ_0) the dielectric permittivity of S (vacuum). A value of Δ_0 can be set by choosing a donor concentration in S,

$$N_d = N_c \exp\left[\frac{(F^S - E_{c0})}{T}\right] = N_c \exp\left(\frac{-\Delta_0}{T}\right) = n, \tag{3.35}$$

where F^S is the Fermi level in the semiconductor bulk, $N_c = 2M_c(2\pi m_* T)^{3/2}h^{-3}$ the effective density of states and M_c the number of effective minima of the semiconductor conduction band; n and m_* the concentration and effective mass of electrons in S [79]. Owing to small barrier thickness l, the electrons can rather easily tunnel through the δ-spike, but only those with an energy $E \geq E_c$ can overcome the wide barrier Δ_0 due to thermionic emission, where $E_c = E_{c0} + qV$. We assume here the standard convention that the bias voltage $V < 0$ and current $J < 0$ in the reverse-biased FM-S junction and $V > 0$ ($J > 0$) in the forward-biased FM-S junction [78]. At positive bias voltage $V > 0$, we assume that the bottom of conduction band shifts upwards to $E_c = E_{c0} + qV$ with respect to the Fermi level of the metal. Presence of the mini-well allows to keep the thickness of the δ-spike barrier equal to $l \lesssim l_0$ and its transparency high at voltages $qV \lesssim rT$ (see below).

The calculation of current through the modified barrier is rather similar to what has been done in the case of tunnel junctions above, with a distinction that in the present case the barrier is triangular, (Figure 3.7). We again assume elastic coherent tunneling, so that the energy E, spin σ, and $\vec{k}_{||}$ (the component of the wave vector $\vec{k}$ parallel to the interface) are conserved. The exact current density of electrons with spin $\sigma = \uparrow, \downarrow$ through the FM-S junction containing the δ-doped layer (at the point $x = l$; Figure 3.7) is written similarly to Equaion 3.24 as:

$$J_{\sigma 0} = \frac{q}{h}\int dE[f(E - F^S_{\sigma 0}) - f(E - F^{FM}_{\sigma 0})] \int \frac{d^2 k_{\|}}{(2\pi)^2} T_\sigma, \tag{3.36}$$

where $F^S_{\sigma 0}(F^{\mathrm{FM}}_{\sigma 0})$ are the spin quasi-Fermi levels in the semiconductor (ferromagnet) near the FM-S interface, and the integration includes a summation with respect to a band index. Note that here we study a strong *spin accumulation* in the semiconductor. Therefore, we use *nonequilibrium* Fermi levels, $F^{\mathrm{FM}}_{\sigma 0}$ and $F^S_{\sigma 0}$, describing distributions

of electrons with spin $\sigma = \uparrow,\downarrow$ in the FM and the S, respectively, which is especially important for the semiconductor. In reality, due to very high electron density in FM metal in comparison with electron density in S, $F_{\sigma 0}^{FM}$ differs negligibly from the equilibrium Fermi level F for currents under consideration; therefore, we can assume that $F_{\sigma 0}^{FM} = F$, as in Refs. [18, 52] (see discussion below).

The current Equation 3.36 should generally be evaluated numerically for a complex band structure $E_{k\sigma}$ [79]. The analytical expressions for $T_\sigma(E, k_{||})$ can be obtained in an effective mass approximation, $\hbar k_\sigma = m_\sigma v_\sigma$, where $v_\sigma = |\nabla E_{k\sigma}|/\hbar$ is the band velocity in the metal. The present Schottky barrier has a "pedestal" with a height $(E_c - F) = \Delta_0 + qV$, which is opaque at energies $E < E_c$. For $E > E_c$ we approximate the δ-barrier by a triangular shape and one can use an analytical expression for $T_\sigma(E, k_{||})$ [5] and find the spin current at the bias $0 < -qV \lesssim rT$, including at room temperature,

$$J_{\sigma 0} = j_0 d_\sigma \left[\frac{2n_{\sigma 0}(V)}{n} - \exp\left(-\frac{qV}{T} \right) \right], \tag{3.37}$$

$$j_0 = \alpha_0 n q v_T \exp(-\eta \kappa_0 l). \tag{3.38}$$

with the most important spin factor

$$d_\sigma = \frac{v_T v_{\sigma 0}}{v_{t0}^2 + v_{\sigma 0}^2}. \tag{3.39}$$

Here $\alpha_0 = 1.2(\kappa_0 l)^{1/3}$, $\kappa_0 \equiv 1/l_0 = (2m_*/\hbar^2)^{1/2}(\Delta - \Delta_0 - qV)^{1/2}$, $v_{t0} = \sqrt{2(\Delta - \Delta_0 - qV)/m_*}$ is the characteristic "tunnel" velocity, $v_\sigma = v_\sigma(E_c)$ the velocity of polarized electrons in FM with energy $E = E_c$, $v_T = \sqrt{3T/m_*}$ the thermal velocity. At larger reverse bias the miniwell on the right from the spike in Figure 3.7 disappears and the current practically saturates. Quite clearly, the tunneling electrons incident almost normally at the interface and contribute most of the current (a more careful sampling can be done numerically [79]).

One can see from Equation 3.30 that the total current $J = J_{\uparrow 0} + J_{\downarrow 0}$ and its spin components $J_{\sigma 0}$ depend on a conductivity of a semiconductor but *not* a ferromagnet, as in usual Schottky junction theories [79]. On the other hand, $J_{\sigma 0}$ is proportional to the spin factor d_σ and the coefficient $j_0 d_\sigma \propto v_T^2 \propto T$, but not the usual Richardson's factor T^2 [78]. Equation 3.37, for current in the FM-S structure, is valid for any sign of the bias voltage V. Note that at $V > 0$ (forward bias) it determines the spin current from S into FM. Hence, it describes *spin extraction* from S [8].

Following the pioneering studies of Aronov and Pikus [67], one customarily assumes a boundary condition $J_{\uparrow 0} = (1 + \Gamma_F)J/2$. Since there is a spin accumulation in S near the FM-S boundary, the density of electrons with spin σ in the semiconductor is $n_{\sigma 0} = n/2 + \delta n_{\sigma 0}$, where $\delta n_{\sigma 0}$ is a non-linear function of the current J, and $\delta n_{\sigma 0} \propto J$ at small current [67] (see also below). Therefore, the larger J the higher the $\delta n_{\sigma 0}$ and the smaller the current $J_{\sigma 0}$ [see Equation 3.37]. In other words, a type of negative feedback is realized, which decreases the spin injection efficiency Γ and

makes it a non-linear function of J (see below). We show that the spin injection efficiency, Γ_0, and the polarization, $P_0 = [n_\uparrow(0) - n_\downarrow(0)]/n$ in the semiconductor near FM-S junctions essentially differ, and that both are small at small bias voltage V (and current J) but *increase* with the current up to P_F. Moreover, P_F can essentially differs from Γ_F, and may ideally approach 100%.

The current in a spin channel σ is given by the standard drift-diffusion approximation [67, 76]:

$$J_\sigma = q\mu n_\sigma E + qD\nabla n_\sigma, \tag{3.40}$$

where E is the electric field; and D and μ are the diffusion constant and mobility of the electrons respectively. D and μ do not depend on the electron spin σ in the non-degenerate semiconductors. From current continuity and electroneutrality conditions

$$J(x) = \sum_\sigma J_\sigma = \text{const}, \quad n(x) = \sum_\sigma n_\sigma = \text{const}, \tag{3.41}$$

we find

$$E(x) = J/q\mu n = \text{const}, \quad \delta n_\downarrow(x) = -\delta n_\uparrow(x). \tag{3.42}$$

Since the *injection* of spin-polarized electrons from FM into S corresponds to a reverse current in the Schottky FM-S junction, one has $J < 0$, and $E < 0$ Figure 3.7. The spatial distribution of density of electrons with spin σ in the semiconductor is determined by the continuity equation [67, 71]

$$\nabla J_\sigma = \frac{q\delta n_\sigma}{\tau_s}, \tag{3.43}$$

where in the present one-dimensional case $\nabla = d/dx$. With the use of Equations 3.40 and 3.42, we obtain the equation for $\delta n_\uparrow(x) = -\delta n_\downarrow(x)$ [67, 76]. Its solution satisfying a boundary condition $\delta n_\uparrow \to 0$ at $x \to \infty$, is

$$\delta n_\uparrow(x) = C\frac{n}{2}\exp\left(-\frac{x}{L}\right) \equiv P_{n0}\exp\left(-\frac{x}{L}\right), \tag{3.44}$$

$$L_{\text{inject (extract)}} = \frac{1}{2}\left[\sqrt{L_E^2 + 4L_s^2} + (-)L_E\right] = \frac{L_s}{2}\left(\sqrt{\frac{J^2}{J_S^2} + 4} - \frac{J}{J_S}\right), \tag{3.45}$$

where the plus (minus) sign refers to forward (reverse) bias on the junction, $L_s = \sqrt{D\tau_s}$ is the usual spin-diffusion length, $L_E = \mu|E|\tau_s = L_s|J|/J_s$ the spin-drift length. Here we have introduced the characteristic current density

$$J_S \equiv qDn/L_s, \tag{3.46}$$

and the plus and minus signs in the expression for the spin penetration depth L Equation 3.45 refer to the spin *injection* at a reverse bias voltage, $J < 0$, and spin *extraction* at a forward bias voltage, $J > 0$, respectively. Note that $L_{\text{inject}} > L_{\text{extract}}$, and the spin penetration depth for injection increases with current, at large currents, $|J| \gg J_S$, $L_{\text{inject}} = L_s|J|/J_s \gg L_s$, whereas $L_{\text{extract}} = L_sJ_s/J \ll L_s$.

The degree of spin polarization of non-equilibrium electrons (i.e., a spin *accumulation* in the semiconductor near the interface) is given simply by the parameter C in Equation 3.37:

$$C = \frac{n_\uparrow(0) - n_\downarrow(0)}{n} = P(0) \equiv P_0. \tag{3.47}$$

By substituting Equation 3.44 into Equations 3.40 and 3.37, we find

$$J_{\uparrow 0} = \frac{J}{2}\left(1 + P_0 \frac{L}{L_E}\right) = \frac{J}{2}\frac{(1+P_F)(\gamma - P_0)}{\gamma - P_0 P_F}, \tag{3.48}$$

where $\gamma = \exp(-qV/T) - 1$. From Equation 3.41, one obtains a quadratic equation for $P_n(0)$ with a physical solution that can be written fairly accurately as

$$P_0 = \frac{P_F \gamma L_E}{\gamma L + L_E}. \tag{3.49}$$

By substituting Equation 3.49 into Equation 3.37, we find for the total current $J = J_{\uparrow 0} + J_{\downarrow 0}$:

$$J = -J_m \gamma = -J_m(e^{-qV/T} - 1), \tag{3.50}$$

$$J_m = \alpha_0 n q v_T (1 - P_F^2)(d_{\uparrow 0} + d_{\downarrow 0}) e^{-\eta \kappa_0 l}, \tag{3.51}$$

for the bias range $|qV| \lesssim rT$. The sign of the Boltzmann exponent is unusual because we consider the tunneling thermoemission current in a modified barrier. Obviously, we have $J > 0$ (<0) when $V > 0$ (<0) for forward (reverse) bias.

We notice that at a reverse bias voltage $-qV \simeq rT$ the shallow potential mini-well vanishes, and $E_c(x)$ takes the shape shown in Figure 3.7 (curve 3). For $-qV > rT$, a wide potential barrier at $x > l$ (in S behind the spike) remains flat (characteristic length scale $\gtrsim 100\,\text{nm}$ at $N_d \lesssim 10^{17}\,\text{cm}^{-3}$), as in usual Schottky contacts [78]. Therefore, the current becomes weakly dependent on V, since the barrier is opaque for electrons with energies $E < E_c - rT$ (curve 4). Thus, Equation 3.50 is valid only at $-qV \lesssim rT$ and the reverse current at $-qV \gtrsim rT$ practically saturates at the value

$$J_{\text{sat}} = qn\alpha_0 v_T (d_{\uparrow 0} + d_{\downarrow 0})(1 - P_F^2)\exp(r - \eta\kappa_0 l). \tag{3.52}$$

With the use of Equations 3.50 and 3.45, we obtain from Equation 3.42 the spin polarization of electrons near FM-S interface,

$$P_0 = -P_F \frac{2J}{2J_m + \sqrt{J^2 + 4J_S^2} - J}. \tag{3.53}$$

The spin injection efficiency at FM-S interface is, using Equations 3.30, 3.41, 3.38 and 3.46,

$$\Gamma_0 \equiv \frac{J_{\uparrow 0} - J_{\downarrow 0}}{J_{\uparrow 0} + J_{\downarrow 0}} = P_0 \frac{L}{L_E} = -P_F \frac{\sqrt{4J_S^2 + J^2} - J}{2J_m + \sqrt{J^2 + 4J_S^2} - J}. \tag{3.54}$$

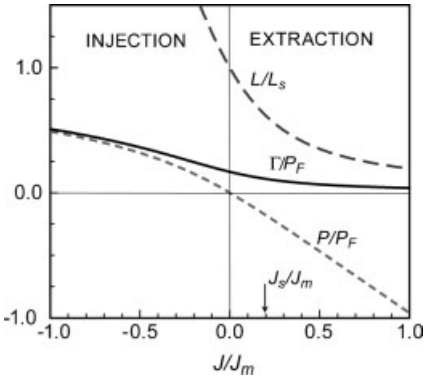


Figure 3.8 The spin accumulation $P = (n_\uparrow - n_\downarrow)/n$, the spin polarization of a current $\Gamma = (J_\uparrow - J_\downarrow)/J$, and the relative spin penetration depth L/L_s (broken line) in the semiconductor as the functions of the relative current density J/J_s for spin injection ($J < 0$) and spin extraction ($J > 0$) regimes. P_F is the spin polarization in the ferromagnet, the ratio $J_s/J_m = 0.2$, L_s is the usual spin diffusion depth. The spin penetration depth considerably exceeds L_s for the injection and smaller than L_s for the extraction.

One can see that Γ_0 strongly differs from P_0 at small currents. As expected, $P_n \approx P_F|J|/J_m \to 0$ vanishes with the current (Figure 3.8), and the prefactor differs from those obtained in Refs. [67, 71, 73, 75, 76].

These expressions should be compared with the results for the case of a degenerate semiconductor, Ref. [10], for the polarization

$$P_0 = -\frac{6P_F J}{3\left(\sqrt{J^2 + 4J_S^2} - J\right) + 10J_m}, \tag{3.55}$$

and the spin injection efficiency

$$\Gamma_0 = -P_F \frac{\sqrt{4J_S^2 + J^2} - J}{2J_m + \sqrt{J^2 + 4J_S^2} - J}. \tag{3.56}$$

In spite of very different statistics of carriers in a degenerate and non-degenerate semiconductor, the accumulated polarization as a function of current behaves similarly in both cases. The important difference comes from an obvious fact that the efficient spin accumulation in degenerate semiconductors may proceed at and below room temperature, whereas in present design an efficient spin accumulation in FM-S junctions with non-degenerate S can be achieved at around room temperature only.

In the reverse-biased FM-S junctions the current $J < 0$ and, according to Equations 3.46 and 3.47, sign $(\delta n_{\uparrow 0}) = \text{sign}\ (P_F)$. In some realistic situations, like elemental Ni, the polarization at energies $E \approx F + \Delta_0$ would be negative, $P_F < 0$ and, therefore, electrons with spin $\sigma = \downarrow$ will be accumulated near the interface. For large currents $|J| \gg J_S$, the spin penetration depth L in Equation 3.45 increases with

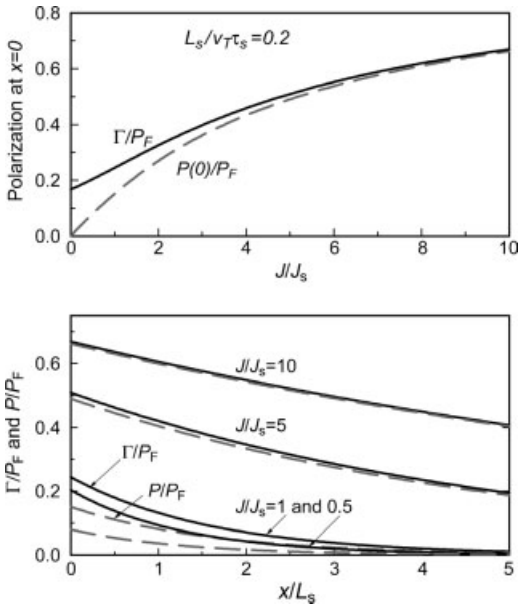


Figure 3.9 Spin polarization of a current $\Gamma = (J_\uparrow - J_\downarrow)/J$ (solid line) and spin accumulation $P = (n_\uparrow - n_\downarrow)/n$ (broken line) in the semiconductor as the functions of the relative current density J/J_s (top panel) and their spatial distribution for different densities of total current J/J_s (bottom panel) at $L_s/v_T\tau_s = 0.2$ where $J_s = qnL_s/\tau_s$, P_F is the spin polarization in the ferromagnet (see text).

current J and the spin polarization (of electron density) approaches the maximum value P_F. Unlike the spin accumulation P_0, the spin injection efficiency (polarization of current) Γ_0 does not vanish at small currents, but approaches the value $\Gamma_0^0 = P_F J_S/(J_S + J_m) \ll P_F$ in the present system with transparent tunnel δ-barrier. There is an important difference with the magnetic tunnel junctions, where the tunnel barrier is relatively opaque and the injection efficiency (polarization of current) is high, $\Gamma \approx P_F$ [18]. However, the polarization of carriers P_0, measured in, for example, spin-LED devices [66], would be minute (see below). Both P_0 and Γ_0 approach the maximum P_F only when $|J| \gg J_S$, (Figure 3.9). The condition $|J| \gg J_S$ is fulfilled at $qV \simeq rT \gtrsim 2T$, when $J_m \gtrsim J_S$.

Another situation is realized *in the forward-biased FM-S junctions* when $J > 0$. Indeed, according to Equations 3.53 and 3.54 at $J > 0$ the electron density distribution is such that sign $(\delta n_{\uparrow 0}) = -$sign (P_F). If a system like elemental Ni is considered (Figure 3.7), then $P_F(F + \Delta_0) < 0$ and $\delta n_{\uparrow 0} > 0$; that is, the electrons with spin $\sigma = \uparrow$ would be accumulated in a non-magnetic semiconductor (NS), whereas electrons with spin $\sigma = \downarrow$ would be extracted from NS (the opposite situation would take place for P_F $(F + \Delta_0) > 0$). One can see from Equation 3.46 that $|P_0|$ one can reach a

maximum P_F only when $J \gg J_s$. According to Equation 3.50, the condition $J \gg J_s$ can only be fulfilled when $J_m \gg J_s$. In this case Equation 3.46 reduces to

$$P_0 = \frac{-P_F J}{J_m} = -P_F(1 - e^{-qV/T}). \tag{3.57}$$

Therefore, the absolute magnitude of a spin polarization approaches its maximal value $|P_0| \simeq P_F$ at $qV \gtrsim 2T$ linearly with current (Figure 3.8). The maximum is reached when J approaches the value J_m, which depends weakly on bias V (see below). In this case $\delta n_{\uparrow(\downarrow)}(0) \approx \mp P_F n/2$ at $P_F > (d_\uparrow > d_\downarrow)$, so that the electrons with spin $\sigma = \uparrow$ are *extracted*, $n_\uparrow(0) \approx (1 - P_F)n/2$, from a semiconductor, while the electrons with spin $\sigma = \downarrow$ are *accumulated* in a semiconductor, $n_\uparrow(0) \approx (1 + P_F)n/2$, near the FM-S interface. The penetration length of the accumulated spin Equation 3.45 at $J \gg J_s$ is

$$L = \frac{L_s^2}{L_E} = \frac{L_s J_S}{J \ll L_s} \ll L_s \quad \text{at} \quad J \gg J_S, \tag{3.58}$$

that is, it decreases as $L \propto 1/J$ (Figure 3.8). We see from Equation 3.54 that at $J \gg J_s$

$$\Gamma_0 = \frac{P_F J_S^2}{J_m J} \to 0. \tag{3.59}$$

Hence, the behavior of the spin injection efficiency at forward bias (extraction) is very different from a spin injection regime, which occurs at a reverse bias voltage: here, the spin injection efficiency Γ_0 remains $\ll P_F$ and vanishes at large currents as $\Gamma_0 \propto J_s/J$. Therefore, we come to an unexpected conclusion that *the spin polarization of electrons,* accumulated in a non-magnetic semiconductor near forward biased FM-S junction can be relatively large for the parameters of the structure when the spin injection efficiency is actually very *small* [8]. Similar, albeit much weaker phenomena are possible in systems with *wide opaque* Schottky barriers [80] and have been probably observed [81]. Spin extraction may also be observed at low temperature in FMS-S contacts [82]. A proximity effect leading to polarization accumulation in FM-S contacts [83] may be related to the same mechanism.

3.5.2
Conditions for Efficient Spin Injection and Extraction

According to Equations 3.51 and 3.46, the condition for maximal polarization of electrons P_n can be written as

$$J_m \gtrsim J_s, \tag{3.60}$$

or, equivalently, as a condition

$$\beta \equiv \frac{\alpha_0 \upsilon_T (d_{\uparrow 0} + d_{\downarrow 0})(1 - P_F^2) e^{-\eta l/l_0} \tau_s}{L_s} \gtrsim 1. \tag{3.61}$$

Note that when $l \lesssim l_0$, the spin injection efficiency at small current is small $\Gamma_0^0 = P_F/(1+\beta) \gg P_F$, since in this case the value $\beta \simeq (d_{\uparrow 0} + d_{\downarrow 0})\alpha_0 \nu_T \tau_s / L_s \gg 1$ for

real semiconductor parameters. The condition $\beta \gg 1$ can be simplified and rewritten as a requirement for the spin-relaxation time

$$\tau_s \gg D\left(\frac{\Delta - \Delta_0}{2\alpha_0 \nu_{\sigma 0}^2 T}\right)^2 \exp\frac{2\eta l}{l_0}. \tag{3.62}$$

It can be met only when the δ-doped layer is very thin, $l \lesssim l_0 \equiv \kappa_0^{-1}$. With typical semiconductor parameters at $T \simeq 300\,\mathrm{K}$ ($D \approx 25\,\mathrm{cm}^2\ \mathrm{s}^{-1}$, $(\Delta - \Delta_0) \simeq 0.5\,\mathrm{eV}$, $\upsilon_{\sigma 0}$ $10^8 \mathrm{cm}^2\ \mathrm{s}^{-1}$ [78]) the condition in Equation 3.62 is satisfied at $l \lesssim l_0$ when the spin-coherence time $\tau_s \gg 10^{-12}\,\mathrm{s}$. It is worth noting that it can certainly be met: for instance, τ_s can be as large as ~1 ns even at $T \simeq 300\,\mathrm{K}$ (e.g., in ZnSe [84]).

Note that the higher the semiconductor conductivity, $\sigma_s = q\mu n \propto n$, the larger the threshold current $J > J_m \propto n$ [Equation 3.51]for achieving the maximal spin injection. In other words, the polarization P_0 reaches the maximum value P_F at *smaller* current in *high-resistance* lightly doped semiconductors compared to heavily doped semiconductors. Therefore, the "conductivity mismatch" [70, 74, 75] is actually irrelevant for achieving an efficient spin injection.

The necessary condition $|J| \gg J_s$ can be rewritten at small voltages, $|qV| \ll T$, as

$$r_c \ll \frac{L_s}{\sigma_s}, \tag{3.63}$$

where $r_c = (dJ/\mathrm{d}V)^{-1}$ is the tunneling contact resistance. Here, we have used the Einstein relation $D/\mu = T/q$ for non-degenerate semiconductors. We emphasize that Equation 3.63 is *opposite* to the condition found by Rashba in Ref. [74] for small currents.

We also emphasize that the spin injection in structures considered in the literature [4, 65–76] has been dominated by electrons at the Fermi level and, according to calculation [85], $g_\downarrow(F)$ and $g_\uparrow(F)$ are such that $P_F \lesssim 40\%$. We also notice that the condition in Equation 3.61 for parameters of the Fe/AlGaAs heterostructure studied in Refs. [65] ($l \simeq 3\,\mathrm{nm}$, $l_0 \simeq 1\,\mathrm{nm}$ and $\Delta_0 = 0.46\,\mathrm{eV}$) is satisfied when $\tau_s \gtrsim 5 \times 10^{-10}\,\mathrm{s}$ and can be fulfilled only at low temperatures. Moreover, for the concentration $n = 10^{19}\,\mathrm{cm}^{-3}$ E_c lies below F, so that the electrons with energies $E \simeq F$ are involved in tunneling, but for these states the polarization is $P_F \lesssim 40\%$. Therefore, the authors of Ref. [65] were indeed able to estimate the observed spin polarization as being ≈32% at low temperatures.

Better control of the injection can be realized in heterostructures where a δ-layer between the ferromagnet and the *n*-semiconductor layer is made of very thin heavily doped n^+-semiconductor with larger electron affinity than the *n*-semiconductor. For instance, FM–n^+-GaAs–*n*-$Ga_{1-x}Al_xAs$, FM–n^+-Ge_xSi_{1-x}–*n*-Si or FM–n^+-Zn_{1-x} Cd_x Se–*n*-ZnSe heterostructures can be used for this purpose. The GaAs, Ge_xSi_{1-x} or $Zn_{1-x}Cd_xSe$$n^+$–layer must have the width $l < 1\,\mathrm{nm}$ and the donor concentration $N_d^+ > 10^{20}\mathrm{cm}^{-3}$. In this case, the ultrathin barrier forming near the ferromagnet-semiconductor interface is transparent for electron tunneling. The barrier height Δ_0 at Ge_xSi_{1-x}–Si, GaAs–$Ga_{1-x}Al_xAs$ or $Zn_{1-x}Cd_xSe$–ZnSe interface is controlled by the composition x, and can be selected as $\Delta_0 = 0.05$–$0.15\,\mathrm{eV}$. When the donor concentration in Si, $Ga_{1-x}Al_xAs$, or ZnSe layer is $N_d < 10^{17}\,\mathrm{cm}^{-3}$, the injected electrons cannot penetrate relatively low and wide barrier Δ_0 when its width $l_0 > 10\,\mathrm{nm}$.

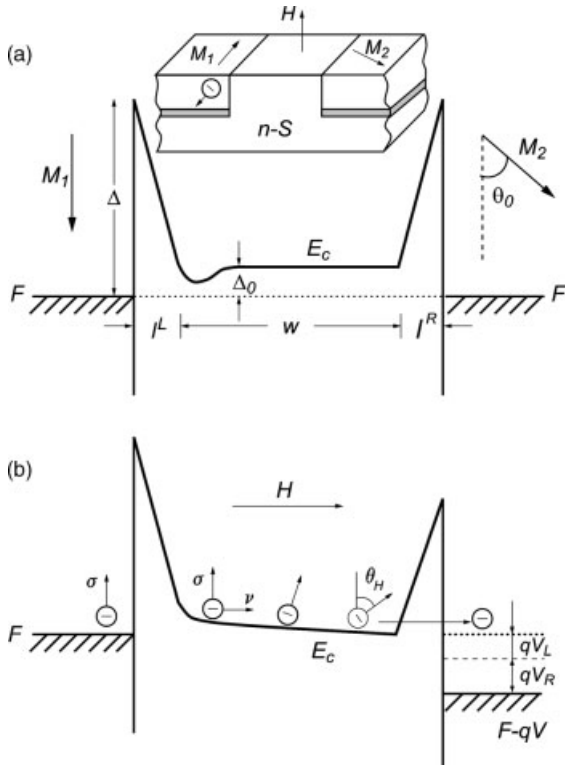


Figure 3.10 Energy diagram of a the FM-S-FM heterostructure with δ-doped layers in equilibrium (a) and at a bias voltage V (b), with V_L (V_R) the fraction of the total drop across the left (right) δ-layer. F marks the Fermi level, Δ the height, $l^{L(R)}$ the thickness of the left (right) δ-doped layer, Δ_0 the height of the barrier in the n-type semiconductor (n-S), E_c the bottom of conduction band in the n-S, w the width of the n-S part. The magnetic moments on the FM electrodes $\vec{M}_1$ and $\vec{M}_2$ are at some angle θ_0 with respect to each other. The spins, injected from the left, drift in the semiconductor layer and rotate by the angle θ_H in the external magnetic field H. Inset: schematic of the device, with an oxide layer separating the ferromagnetic films from the bottom semiconductor layer.

3.5.3
High-Frequency Spin-Valve Effect

Here we describe a new high-frequency spin valve effect that can be observed in a FM-S-FM device with two back-to-back modified Schottky contacts (Figure 3.10). We find the dependence of current on a magnetic configuration in FM electrodes and an external magnetic field. The spatial distribution of spin-polarized electrons is determined by the continuity [Equation 3.43] and the current in spin channel σ is given by Equation 3.40. Note that $J < 0$, thus $E < 0$ in a spin injection regime. With the use of the kinetic equation and Equation 3.40, we obtain the equation for $\delta n_\uparrow$, Equation 3.43 [68]. Its general solution is

$$\delta n_{\uparrow}(x) = \frac{n}{2}(c_1 e^{-x/L_1} + c_2 e^{-(w-x)/L_2}), \tag{3.64}$$

where $L_{1(2)} = (1/2)[\sqrt{L_E^2 + 4L_s^2} + (-)L_E]$ is the same as found earlier in Equation 3.45. Consider the case when $w \ll L_1$ and the transit time $t_{tr} \simeq w^2/(D + \mu|E|w)$ of the electrons through the *n*-semiconductor layer is shorter than τ_s. In this case, a spin ballistic transport takes place; that is, the spin of the electrons injected from the FM_1 layer is conserved in the semiconductor layer, $\sigma' = \sigma$. Probabilities of the electron spin $\sigma = \uparrow$ to have the projections along $\pm\vec{M}_2$ are $\cos^2(\theta/2)$ and $\sin^2(\theta/2)$, respectively, where θ is the angle between vectors $\sigma = \uparrow$ and $\vec{M}_2$. Accounting for this, we find that the resulting current through the structure saturates at bias voltage $-qV > T$ at the value

$$J = J_0 \frac{1 - P_R^2 \cos^2\theta}{1 - P_L P_R \cos\theta}, \tag{3.65}$$

where J_0 is the prefactor similar to Equation 3.38. For the *opposite* bias the total current J is given by Equation 3.65 with the replacement $P_L \leftrightarrow P_R$. The current J is minimal

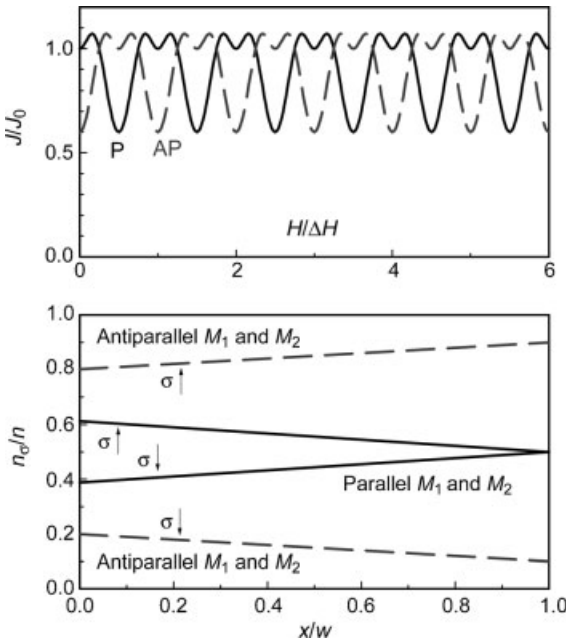


Figure 3.11 Oscillatory dependence of the current J through the structure on the magnetic field H (top panel) for parallel (P) and antiparallel (AP) moments M_1 and M_2 on the electrodes, Figure 3.10, and $P_L = P_R = 0.5$. Spatial distribution of the spin polarized electrons $n_{\uparrow(\downarrow)}/n$ in the structure for different configurations of the magnetic moments M_1 and M_2 in the limit of saturated current density J, $w = 60$ nm, $L_2 = 100$nm (bottom panel).

for antiparallel (AP) moments $\vec{M}_1$ and $\vec{M}_2$ in the electrodes when $\theta = \pi$ and near maximal for parallel (P) magnetic moments $\vec{M}_1$ and $\vec{M}_2$.

The present heterostructure has an additional degree of freedom, compared to tunneling FM-I-FM structures that can be used for *magnetic sensing*. Indeed, spins of the injected electrons can precess in an external magnetic field H during the transit time t_{tr} of the electrons through the semiconductor layer ($t_{tr} < \tau_s$). The angle between the electron spin and the magnetization $\vec{M}_2$ in the FM_2 layer in Equation 3.65 is in general $\theta = \theta_0 + \theta_H$ where θ_0 is the angle between the magnetizations M_1 and M_2, and $\theta_H = \gamma_0 g H t_{tr}(m_0/m_*)$ is the spin rotation angle. Here, H is the magnetic field normal to the spin direction, $\gamma = qg/(m_* c)$ is the gyromagnetic ratio, g is the g-factor. According to Equation 3.65, with increasing H the current *oscillates* with an amplitude $(1 + P_L P_R)/(1 - P_L P_R)$ and period $\Delta H = (2\pi m_*)(\gamma_0 g m_0 t_{tr})^{-1}$ (Figure 3.11, top panel). The maximum operating speed of the field sensor is very high, since redistribution of nonequilibrium-injected electrons in the semiconductor layer occurs over the transit time $t_{tr} = w/\mu|E| = J_s w \tau_s/(J L_s)$, $t_{tr} \lesssim 10^{-11}$ s for $w \lesssim 200_*$ nm, $\tau_s \sim 3 \times 10^{-10}$ s, and $J/J_s \gtrsim 10$ ($D \approx 25\ \mathrm{cm^2\ s^{-1}}$) at $T \simeq 300$ K [78]). Therefore, the operating frequency $f = 1/t_{tr} \gtrsim 100$ GHz may be achievable at room temperature. We see that: (i) the present heterostructure can be used as a sensor for an ultrafast nanoscale reading of an inhomogeneous magnetic field profile; (ii) it includes two FM-S junctions and can be used for measuring the spin polarizations of these junctions; and (iii) it is a *multifunctional* device where current depends on the mutual orientation of the magnetizations in the ferromagnetic layers, an external magnetic field, and a (small) bias voltage. Thus, it can be used as a logic element, a magnetic memory cell, or an ultrafast read head.

3.5.4
Spin-Injection Devices

The high-frequency spin-valve effect, described above, can be used for designing a new class of ultrafast spin-injection devices such as an amplifier, a frequency multiplier, and a square-law detector [6]. Their operation is based on the injection of spin-polarized electrons from one ferromagnet to another through a semiconductor layer, and spin precession of the electrons in the semiconductor layer in a magnetic field induced by a (base) current in an adjacent nanowire. The base current can control the emitter current between the magnetic layers with frequencies up to several 100 GHz. Here, we shall describe a spintronic mechanism of ultrafast amplification and frequency conversion, which can be realized in heterostructures comprising a metallic ferromagnetic nanowire surrounded by a semiconductor (S) and a ferromagnetic (FM) thin shells (Figure 3.12a). Practical devices may have various layouts, with two examples shown in Figure 3.12b and c.

Let us consider the principle of operation of the spintronic devices shown in Figure 3.12a. When the thickness w of the n-type semiconductor layer is not very small ($w \gtrsim 30$ nm), tunneling through this layer would be negligible. The base voltage V_b is applied between the ends of the nanowire. The base current J_b, flowing through the nanowire, induces a cylindrically symmetric magnetic field $H_b = J_b/2\pi\rho$ in the S layer, where ρ is the distance from the center of nanowire. When the emitter

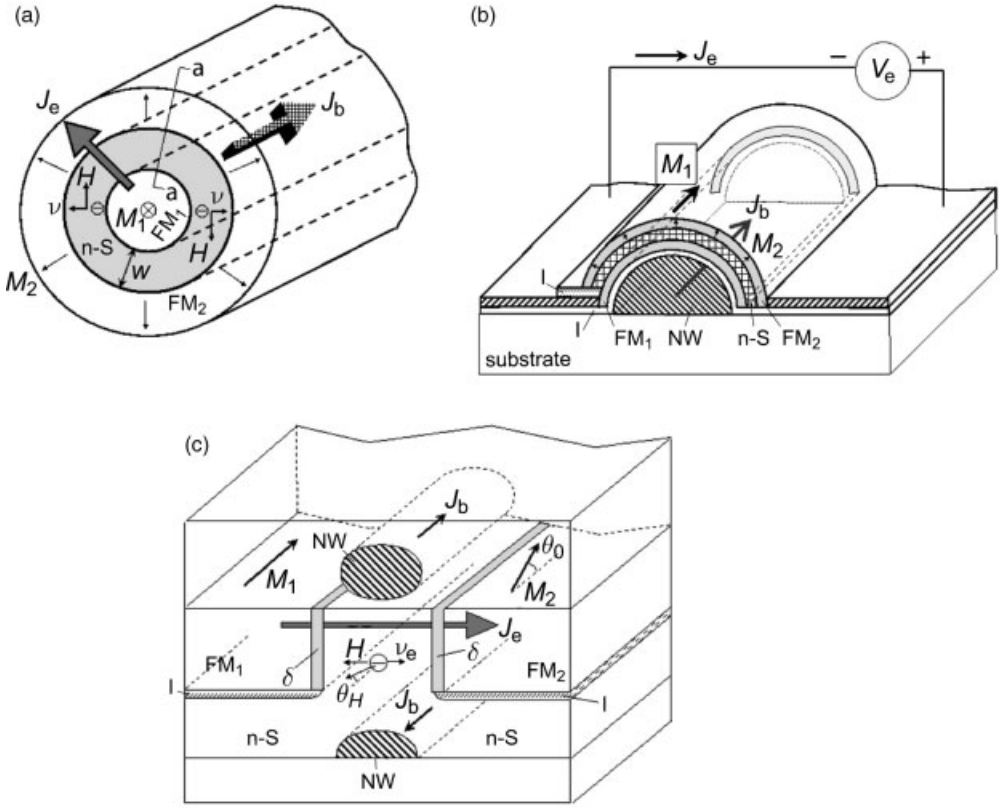


Figure 3.12 Schematic of the spin injection-precession devices having (a) cylindrical, (b) semi-cylindrical, and (c) planar shape. Here, FM_1 and FM_2 are the ferromagnetic layers; *n*-S the *n*-type semiconductors layer; *w* the thickness of the *n*-S layer; δ the δ-doped layers; NW the highly conductive nanowires; I the insulating layers. The directions of the magnetizations $\vec{M}_1$ and $\vec{M}_2$ in the FM_1 and FM_2 layers, as well as the electron spin σ, the magnetic field H_b, and the angle of spin rotation θ in S are also shown.

voltage V_e is applied between FM layers, the spin-polarized electrons are injected from the first layer (nanowire FM_1) through the semiconductor layer into the second (exterior) ferromagnetic shell, FM_2. The FM_1-S and FM_2-S junctions are characterized by the spin injection efficiencies P_1 and P_2, respectively. We assume that the transit time t_{tr} of the electrons through the S layer is less than the spin relaxation time, τ_s (i.e., we consider the case of a spin ballistic transport). The exact calculation gives a current J_e, through the structure as a function of the angle θ between the magnetization vectors $\vec{M}_1$ and $\vec{M}_2$ in the ferromagnetic layers. At small angles θ or $P_1 = P_{\mathrm{L}}$ or $P_2 = P_{\mathrm{R}}$

$$J_e = J_{0e}(1 + P_L P_R \cos\theta), \tag{3.66}$$

where $\theta = \theta_0 + \theta_{\mathrm{H}}$, θ_0 is the angle between $\vec{M}_1$ and $\vec{M}_2$, and θ_H is the angle of the spin precesses with the frequency $\Omega = \gamma H_\perp$, where $H_\perp$ is the magnetic field component

normal to the spin and γ is the gyromagnetic ratio. One can see from Figure 3.12a that $H_\perp = H_b = J_b/(2\pi\rho)$. Thus, the angle of the spin rotation is equal to $\theta_H = \gamma H_b t_{tr} = t_{tr} J_b/2\pi\rho_s$, where ρ_s is the characteristic radius of the S layer. Then, according to Equation 3.57,

$$J_e = J_{e0}[1 + P_1 P_2 \cos(\theta_0 + k_j J_b)], \tag{3.67}$$

where $k_j = \gamma t_{tr}/2\pi\rho_s = \gamma/\omega\rho_s$ and $\omega = 2\pi/t_{tr}$ is the frequency of a variation of the base current, $J_b = J_s \cos(\omega t)$.

Equation 3.67 shows that, when the magnetization M_1 is perpendicular to M_2, $\theta_0 = \pi/2$, and $\theta_H \ll \pi$,

$$J_e = J_{e0}(1 + k_j P_1 P_2 J_b),\ G = dJ_e/dJ_b = J_{e0} k_j P_1 P_2. \tag{3.68}$$

Hence, the *amplification* of the base current occurs with the gain G, which can be relatively high even for $\omega \gtrsim 100$ GHz. Indeed, $\gamma = q/(m_* c) \approx 2.2(m_0/m_*)10^5\ m/(\text{A}\cdot\text{s})$, where m_0 is the free electron mass, m_* the effective mass of electrons in the semiconductor, and c the velocity of light. Thus, the factor $k_j \simeq 10^3\,\text{A}^{-1}$ when $\rho_s \simeq 30$ nm, $m_0/m_* = 14$ (GaAs) and $\omega = 100$ GHz, so that $G > 1$ at $J_{e0} > 0.1$ mA/$(P_1 P_2)$.

When M_1 is collinear with M_2 ($\theta_0 = 0,\pi$) and $\theta_H \ll \pi$, then, according to Equation 3.67, the emitter current is

$$J_e = J_{e0}(1 \pm P_1 P_2) \mp \frac{1}{2} J_{e0} P_1 P_2 k_j^2 J_b^2. \tag{3.69}$$

Therefore, the time-dependent component of the emitter current $\delta J_e(t) \propto J_b^2(t)$, and the device operates as a square-law detector. When $J_b(t) = J_{b0}\cos(\omega_0 t)$, the emitter current has a component $\delta J_e(t) \propto \cos(2\omega_0 t)$, and the device operates as a *frequency multiplier*. When $J_b(t) = J_h \cos(\omega_h t) + J_s \cos(\omega_s t)$, the emitter current has the components proportional to $\cos(\omega_h \pm \omega_s)t$; that is, the device can operate as a high-frequency *heterodyne detector* with the conversion coefficient $K = J_{e0} J_h P_1 P_2 k_j^2/4$. For $k_j = 10^3\text{A}^{-1}$, one obtains $K > 1$ when $J_{e0} J_h > 4\ (\text{mA})^2/(P_1 P_2)$.

3.5.5
Spin Source of Polarized Radiation

The spin extraction effect can be used for making an efficient source of (modulated) polarized radiation. Consider a structure containing a FM-S junction with the δ-doped layer and a double p–n'–n heterostructure where the n'-region is made from narrower gap semiconductor (Figure 3.13). We show that the following effects can be realized in the structure when both FM-S junction and the heterostructure are biased in the forward direction, and the electrons are injected from n-semiconductor region into FM and p-region. Due to a spin selection property of FM-S junction [7], spin-polarized electrons appear in n-region with a spatial extent $L \lesssim L_s$ near the FM-S interface, where L_s is the spin diffusion length in NS. When the thickness of the n-region w is smaller than L, the spin-polarized electron from the n-region and

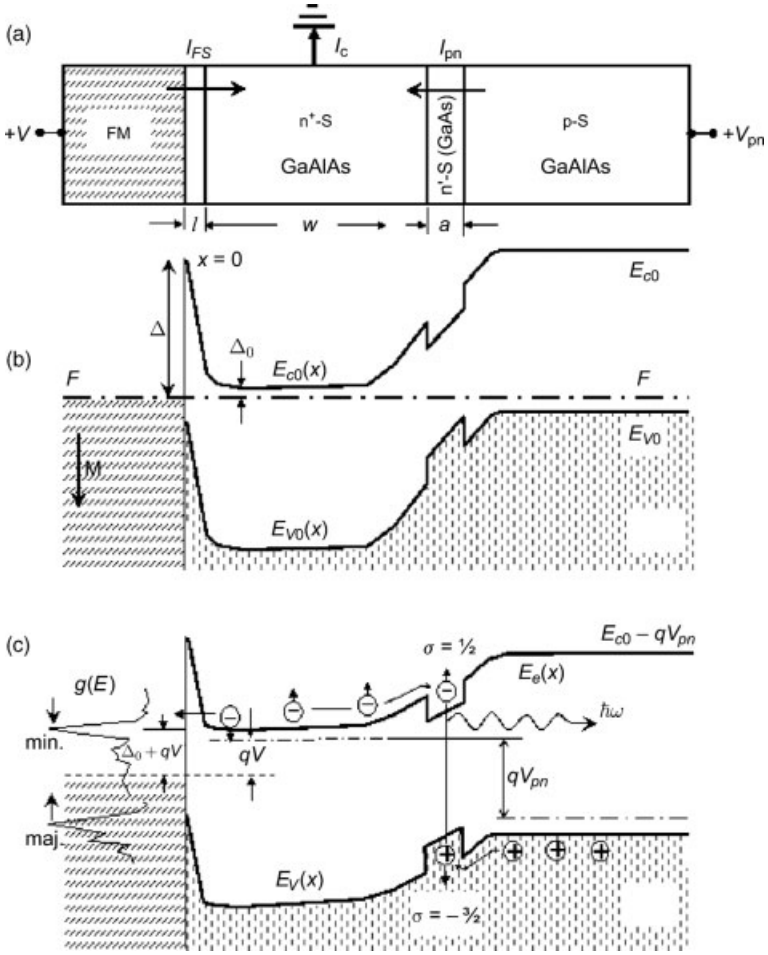


Figure 3.13 Schematic (a) of structure and the band diagram of polarized photon source containing FM-S junction with δ-doped layer and a double n^+–n'–p heterostructure without (b) and under the bias voltage V. Minority spin electrons are extracted from n^+–S semiconductor layer and the remaining (majority) electrons are recombined in n'–S quantum well. F is the Fermi level, Δ the height and l the thickness of the δ-doped layer, Δ_0 the height of a barrier in the n-type semiconductor, $E_c(x)$ the bottom of conduction band in the semiconductor. The spin density of states is shown at $x < 0$ with a high peak in minority states at $E = F + \Delta_0$, typical of elemental Ni, as an example.

holes from p-region are injected and accumulated in a thin narrow-gap n'-region (quantum well) where they recombine and emit polarized photons.

The conditions for maximal polarization are obtained as follows. When the thickness of n-region is $w < L$, we can assume that $\delta n_\uparrow(x) \simeq \delta n_{\uparrow 0}$ and $P_n \simeq P_0$. In this case, integrating Equation 3.43 over the volume of the n-semiconductor region (with area S and thickness w), we obtain

$$I_{\uparrow FS} + I_{\uparrow pn} - I_{c\uparrow} = q\delta n_{\uparrow 0} wS/\tau_s = P_0 I_S w/2L_s. \tag{3.70}$$

Here, $I_s = J_s S$; $I_{\uparrow FS} = J_{\uparrow 0} S$ and $I_{\uparrow pn}$ are the electron currents with spin $\sigma = \uparrow$ flowing into the n-region from FM and the p-region, respectively; $I_{\uparrow c}$ is the spin current out of the n-region in a contact (Figure 3.13a). The current $I_{\uparrow pn}$ is determined by injection of electrons with $\sigma = \uparrow$ from the n-region into the p-region through the crossectional area S, equal to $I_{\uparrow pn} = I_{pn} n_{\uparrow 0}/n = I_{pn}(1 + P_0)/2$, where I_{pn} is the total current in the p–n junction. The current of metal contact I_c is not spin-polarized; hence $I_{c\uparrow} = (I_{pn} + I_{FS})/2$, where I_{FS} is the total current in the FM-S junction. The current in the FM-S junction I_{FS} approaches a maximal value $I_m = J_m S$ at rather small bias, $qV_{FS} > 2T$. When $I_{pn} \ll I_{FS} \simeq I_m$ and $I_m \gg I_S w/L_s$, we get $P_0 \simeq -P_F$. The way to maximize polarization is by adjusting V_{FS}. The maximal $|P_0|$ can be achieved for the process of electron tunneling through the δ-doped layer when the bottom of conduction band in a semiconductor $E_c = F + \Delta_0 + qV_{FS}$ is close to a peak in the density of states of minority carriers in the elemental ferromagnet (Figure 3.13c, curve g).

The rate of polarized radiation recombination is $R_\sigma = qn'_\sigma d/\tau_R$ and the polarization of radiation is $p = (R_\uparrow - R_\downarrow)/(R_\uparrow + R_\downarrow) = 2\delta n'_\uparrow/n'$. Since $I_{pn} = qn'd/\tau n$, we find $2\delta n'_\uparrow/n' = P_0 \tau'_s (\tau'_s + \tau_n)^{-1}$, so that $p = P_0 \tau'_s (\tau'_s + \tau_n)^{-1}$. Thus, the radiation polarization p can approach maximum $p \simeq |P_F|$ at large current $I \simeq I_m$ when $\tau < \tau'_s$. The latter condition can be met at high concentration n' when the time of radiation recombination $\tau_R \simeq \tau_n < \tau'_s$. For example, in GaAs $\tau_R \simeq 3 \times 10^{-10}$ s at $n \gtrsim 5 \times 10^{17}\,\mathrm{cm}^{-3}$ [86] and τ'_s can be larger than τ_R [84]. We emphasize that spin injection efficiency near a forward-biased FM-S junction is very small.

Practical structures may have various layouts, with one example shown in Figure 3.14. It is clear that the distribution of $\delta n_\uparrow(r\rightharpoonup)$ in such a 2-D structure is characterized by the length $L \lesssim L_s$ in the direction x where the electrical field E can be strong, and by the diffusion length L_s in the (y, z) plane where the field is weak. Therefore, the spin density near FM and p–n junctions will be close to $\delta n_{\uparrow 0}$ when the size of the p-region is $d < L_s$. Thus, the above results for one-dimensional structure

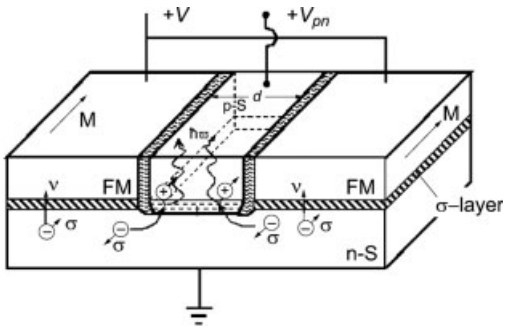


Figure 3.14 Layout of a structure from Figure 3.13, including FM layers and semiconductor n- and p-regions. Here, n' made from narrower gap semiconductor, δ-doped layers are between the FM layers and the n-semiconductor. FM layers are separated by thin dielectric layers from the p-region.

(Figure 3.13) are also valid for more complex geometry shown in Figure 3.14. The predicted effect should also exist for a reverse-biased FM-S junction where the radiation polarization p can approach $+P_F$.

3.6 Conclusions

In this chapter we have described a variety of heterostructures where the spin degree of freedom can be used to efficiently control the current: magnetic tunnel junctions, metallic magnetic multilayers exhibiting giant magnetoresistance, spin-torque effects in magnetic nanopillars. We also described a method of facilitating an efficient spin injection/accumulation in semiconductors from standard ferromagnetic metals at room temperature. The main idea is to engineer the band structure near the ferromagnet-semiconductor interface by fabricating a delta-doped layer there, thus making the Schottky barrier very thin and transparent for tunneling. A long spin lifetime in a semiconductor allows the suggestion of some interesting new devices such as field detectors, spin transistors, square-law detectors, and sources of the polarized light described in the present text. This development opens up new opportunities in potentially important novel spin injection devices. We also discussed a body of various spin-orbit effects and systems of interacting spins. In particular, Spin Hall effects result is positive magnetoresitance due to spin accumulation that may be used to extract the coeefficints for spin-orbit transport. We notice, however, that Datta–Das spinFET would have inferior characteristics compared to MOSFET. We also discussed the severe challenges facing *single-spin logic and, especially, spin-based quantum computers*.

References

1 (a) Wolf, S.A., Awschalom, D.D., Buhrman, R.A., Daughton, J.M., von Molnar, S., Roukes, M.L., Chtchelkanova, A.Y. and Treger, D.M. (2001) *Science*, **294**, 1488; (b) Awschalom D.D., Loss D. and Samarth N. (eds) (2002) *Semiconductor Spintronics and Quantum Computation*, Springer, Berlin.

2 Žutić I., Fabian, J. and Das Sarma, S. (2004) *Reviews of Modern Physics*, **76**, 323.

3 (a) Datta, S. and Das, B. (1990) *Applied Physics Letters*, **56**, 665. (b) Gardelis, S., Smith, C.G., Barnes, C.H.W., Linfield, E.H. and Ritchie, D.A. (1999) *Physical Review B-Condensed Matter*, **60**, 7764.

4 (a) Sato, R. and Mizushima, K. (2001) *Applied Physics Letters*, **79**, 1157; (b) Jiang, X., Wang, R., van Dijken, S., Shelby, R., Macfarlane, R., Solomon, G.S., Harris, J. and Parkin, S.S.P. (2003) *Physical Review Letters*, **90**, 256603.

5 Bratkovsky, A.M. and Osipov, V.V. (2004) *Physical Review Letters*, **92**, 098302.

6 Osipov, V.V. and Bratkovsky, A.M. (2004) *Applied Physics Letters*, **84**, 2118.

7 Osipov, V.V. and Bratkovsky, A.M. (2004) *Physical Review B-Condensed Matter*, **70**, 235302.

8 Bratkovsky, A.M. and Osipov, V.V. (2004) *Journal of Applied Physics*, **96**, 4525.

9 Bratkovsky, A.M. and Osipov, V.V. (2005) *Applied Physics Letters*, **86**, 071120.

10 Osipov, V.V. and Bratkovsky, A.M. (2005) *Physical Review B-Condensed Matter*, **72**, 115322.

11 (a) Baibich, M.N., Broto, J.M., Fert, A., Nguyen Van Dau, F., Petroff, F., Etienne, P., Creuzet, G., Friederich, A. and Chazelas, J. (1988) *Physical Review Letters*, **61**, 2472; (b) Berkowitz, A.E., Mitchell, J.R., Carey, M.J., Young, A.P., Zhang, S., Spada, F.E., Parker, F.T., Hutten, A. and Thomas, G. (1992) *Physical Review Letters*, **68**, 3745.

12 Julliere, M. (1975) *Physics Letters*, **54A**, 225.

13 Maekawa, S. and Gäfvert, U. (1982) *IEEE Transactions on Magnetics*, **18**, 707.

14 Meservey, R. and Tedrow, P.M. (1994) *Physics Reports*, **238**, 173.

15 Moodera, J.S., Kinder, L.R., Wong, T.M. and Meservey, R. (1995) *Physical Review Letters*, **74**, 3273.

16 (a) Yuasa, S., Nagahama, T., Fukushima, A., Suzuki, Y. and Ando, K. (2004) *Nature Materials*, **3**, 858; (b) Parkin, S.S.P., Kaiser, C., Panchula, A., Rice, P.M., Hughes, B., Samant, M. and Yang, S.-H. (2004) *Nature Materials*, **3**, 862.

17 Slonczewski, J.C. (1989) *Physical Review B-Condensed Matter*, **39**, 6995.

18 Bratkovsky, A.M. (1997) *Physical Review B-Condensed Matter*, **56**, 2344.

19 Bratkovsky, A.M. (1998) *Applied Physics Letters*, **72**, 2334.

20 Mott, N.F. (1936) *Proceedings of the Royal Society of London. Series A*, **153**, 699.

21 Stearns, M.B. (1977) *Journal of Magnetism and Magnetic Materials*, **5**, 167.

22 Berger, L. (1996) *Physical Review B-Condensed Matter*, **54**, 9353.

23 Slonczewski, J.C.(1996) *Journal of Magnetism and Magnetic Materials*, **159**, L1.

24 Tsoi, M.V. *et al*. (1998) *Physical Review Letters*, **80**, 4281.

25 (a) Slonczewski, J.C. (2002) *Journal of Magnetism and Magnetic Materials*, **247**, 324; (b) Slonczewski, J.C. (2005) *Physical Review B-Condensed Matter*, **71**, 024411; (c) Slonczewski, J.C., Sun, J.Z. (2007) *J. Magn. Magn. Mater.*, **310**, 169.

26 Emley, N.C. *et al*. (2006) *Physical Review Letters*, **96**, 247204. They have added a phenomenological spin-torque (ST) factor $\eta = A/(1 + B\cos\theta)$ to fit the data for the torque.

27 Pribiag, V.S. *et al*. (2007) *Nature Physics*, **3**, 498.

28 Hayashi, M., Thomas, L., Rettner, C., Moriya, R., Bazaliy, Ya.B. and Parkin, S.S.P. (2007) *Physical Review Letters*, **98**, 037204.

29 (a) Vasko, F.T. (1979) *Pisma Zh Eksp Teor Fiz*, **30**, 574; (b) Vasko, F.T. (1979) *JETP Letters*, **30**, 541.

30 (a) Yu. A., Bychkov and Rashba, E.I. (1984) *Pisma Zh Eksp Teor Fiz*, **39**, 66; (1984) *JETP Letters*, **39**, 78; (b) Yu. A., Bychkov and Rashba, E.I. (1984) *Journal of Physics C*, **17**, 6039.

31 Anselm, A. (1981) *Introduction to Semiconductor Theory*, Prentice-Hall, New Jersey.

32 Engel, H.A., Rashba, E.I. and Halperin, B.I. (2006) cond-mat/0603306.

33 Winkler, R. (2003) *Spin-Orbit Coupling Effects in Two-Dimensional Electron and Hole System*, Springer, Berlin.

34 Nozieres, P. and Lewiner, C. (1973) *Journal of Physics (Paris)*, **34**, 901.

35 Winkler, R. (2000) *Physical Review B-Condensed Matter*, **62**, 4245.

36 Pikus, G.E. and Titkov, A.N. (1984) in *Optical Orientation*, (eds F. Meier and B.P. Zakharchenya), North Holland, Amsterdam, p. 73.

37 Nitta, J., Akazaki, T., Takayanagi, H. and Enoki, T. (1997) *Physical Review Letters*, **78**, 13351338.

38 Koga, T., Nitta, J., Akazaki, T. and Takayanagi, H. (2002) *Physical Review Letters*, **89**, 046801.

39 Grundler, D. (2000) *Physical Review Letters*, **84**, 60746077.

40 (a) Dyakonov, M.I. and Perel, V.I. (1971) *JETP Letters*, **13**, 467; (b) Dyakonov, M.I. and Perel, V.I. (1971) *Physics Letters A*, **35**, 459.

41 Dyakonov, M.I. (2007) *Physical Review Letters*, **99**, 126601.

42 Bakun, A.A., Zakharchenya, B.P., Rogachev, A.A., Tkachuk, M.N. and Fleisher, V.G. (1984) *Pisma Zh Eksp Teor Fiz*, **40**, 464.

43 Kato, Y.K. *et al.* (2004) *Science*, **306**, 1910.

44 Wunderlich, J. *et al.* (2005) *Physical Review Letters*, **94**, 047204.

45 Liu, B., Shi, J., Wang, W., Zhao, H., Li, D., Zhang, S., Xue Q. and Chen, D. (2006) arXiv:cond-mat/0610150.

46 Kimura, T., Otani, Y., Sato, T., Takahashi, S. and Maekawa, S. (2007) *Phy Rev. Lett.*, **98**, 156601.

47 Bandyopadhay, S., Das, B. and Miller, A.E. (1994) *Nanotechnology*, **5**, 113.

48 Cowburn, R.P. and Welland, M.E. (2000) *Science*, **287**, 1466.

49 Rakhmanova, S. and Mills, D.L. (1996) *Physical Review B-Condensed Matter*, **54**, 9225.

50 Dyakonov, M.I. quant-ph/0610117.

51 Zhang, X.-G. and Butler, W.H. (2004) *Physical Review B-Condensed Matter*, **70**, 172407.

52 Duke, C.B. (1969) *Tunneling in Solids*, Academic Press, New York.

53 Ma, W.G. (1992) *Appl. Phys. Lett.*, **61**, 2542. Even smaller $m_2 = 0.2$ has been used by Q. Q. Shu and for Al-Al_2O_3-metal junctions.

54 Larkin, A.I. and Matveev, K.A. (1987) *Zh Eksp Teor Fiz*, **93**, 1030. (1987) *Soviet Physics JETP*, **66**, 580.

55 Xu, Y., Ephron, D. and Beasley, M.R. (1995) *Physical Review B-Condensed Matter*, **52**, 2843.

56 (a) Jansen, R. and Moodera, J.S. (2000) *Physical Review B-Condensed Matter*, **61**, 9047; (b) Jansen, R. and Moodera, J.S. (1998) *Journal of Applied Physics*, **83**, 6682.

57 Tsymbal, E.Y., Sokolov, A., Sabirianov, I.F. and Doudin, B. (2003) *Physical Review Letters*, **90**, 186602.

58 Parkin, S.S.P. private communication.

59 Bratkovsky, A.M. to be published.

60 Monsma, D.J. *et al.* (1995) *Physical Review Letters*, **74**, 5260.

61 Smirnov, A.V. and Bratkovsky, A.M. (1997) *Physical Review B-Condensed Matter*, **55**, 14434.

62 Mahan, G.D. (1990) *Many-Particle Physics*, 2nd edn, Plenum Press, New York, Chapter 9.

63 Zhang, S. *et al.* (1997) *Physical Review Letters*, **79**, 3744. These authors have assumed that surface magnons are excited, and did not consider phonons and bias dependence of direct tunneling.

64 (a) Osipov, V.V., Viglin, N.A. and Samokhvalov, A.A. (1998) *Physics Letters A*, **247**, 353; (b) Ohno, Y., Young, D.K., Beschoten, B., Matsukura, F., Ohno, H. and Awschalom, D.D. (1999) *Nature*, **402**, 790; (c) Fiederling, R., Keim, M., Reuscher, G., Ossau, W., Schmidt, G., Waag, A. and Molenkamp, L.W. (1999) *Nature*, **402**, 787.

65 (a) Hanbicki, A.T. Jonker, B.T. Itskos, G. Kioseoglou, G. and Petrou, A. (2002) *Applied Physics Letters*, **80**, 1240; (b) Hanbicki, A.T., van't Erve, O.M.J., Magno, R., Kioseoglou, G., Li, C.H. and Jonker, B.T. (2003) *Applied Physics Letters*, **82**, 4092; (c) Adelmann, C., Lou, X., Strand, J., Palmstrøm, C.J. and Crowell, P.A. (2005) *Physical Review B-Condensed Matter*, **71**, 121301.

66 (a) Hammar, P.R., Bennett, B.R., Yang, M.J. and Johnson, M. (1999) *Physical Review Letters*, **83**, 203; (b) Zhu, H.J., Ramsteiner, M., Kostial, H., Wassermeier, M., Schönherr, H.-P. and Ploog, K.H. (2001) *Physical Review Letters*, **87**, 016601; (c) Lee, W.Y., Gardelis, S., Choi, B.-C., Xu, Y.B., Smith, C.G., Barnes, C.H.W., Ritchie, D.A., Linfield, E.H. and Bland, J.A.C. (1999) *Journal of Applied Physics*, **85**, 6682; (d) Manago T. and Akinaga, H. (2002) *Applied Physics Letters*, **81**, 694. (e) Motsnyi, A.F., De Boeck, J., Das, J., Van Roy, W., Borghs, G., Goovaerts, E. and Safarov, V.I. (2002) *Applied Physics Letters*, **81**, 265; (f) Ohno, H., Yoh, K., Sueoka, K., Mukasa, K., Kawaharazuka, A. and Ramsteiner, M.E. (2003) *Japanese Journal of Applied Physics*, **42**, L1.

67 Aronov A.G. and Pikus, G.E. (1976) *Fiz Tekh Poluprovodn*, **10**, 1177. (1976) *Soviet Physics Semiconductors-USSR*, **10**, 698.

68 (a) Johnson M. and Silsbee, R.H. (1987) *Physical Review B-Condensed Matter*, **35**, 4959; (b) Johnson M. and Byers, J. (2003) *Physical Review B-Condensed Matter*, **67**, 125112.

69 (a) van Son, P.C., van Kempen, H. and Wyder, P. (1987) *Physical Review Letters*, **58**, 2271; (b) Schmidt, G., Richter, G., Grabs, P., Gould, C., Ferrand, D. and Molenkamp, L.W. (2001) *Physical Review Letters*, **87**, 227203.

70 Schmidt, G., Ferrand, D., Molenkamp, L.W., Filip A.T. and van Wees, B.J. (2000) *Physical Review B-Condensed Matter*, **62**, R4790.

71 Yu Z.G. and Flatte, M.E. (2002) *Physical Review B-Condensed Matter*, **66**, R201202.

72 Albrecht J.D. and Smith, D.L. (2002) *Physical Review B-Condensed Matter*, **66**, 113303.

73 Hershfield, S. and Zhao, H.L. (1997) *Physical Review B-Condensed Matter*, **56**, 3296.

74 Rashba, E.I. (2000) *Physical Review B-Condensed Matter*, **62**, R16267.

75 Fert A. and Jaffres, H. (2001) *Physical Review B-Condensed Matter*, **64**, 184420.

76 Yu Z.G. and Flatte, M.E. (2002) *Physical Review B-Condensed Matter*, **66**, 235302.

77 (a) Shen, M., Saikin, S. and Cheng, M.-C. (2005) *IEEE Transactions on Nanotechnology*, **4**, 40; (b) Shen, M., Saikin, S. and Cheng, M.-C. (2004) *Journal of Applied Physics*, **96**, 4319.

78 (a) Sze, S.M. (1981) *Physics of Semiconductor Devices*, Wiley, New York; (b) Monch, W. (1995) *Semiconductor Surfaces and Interfaces*, Springer, Berlin; (c) Tung, R.T. (1992) *Physical Review B-Condensed Matter*, **45**, 13509.

79 (a) Sanvito, S., Lambert, C.J., Jefferson, J.H. and Bratkovsky, A.M. (1999) *Physical Review B-Condensed Matter*, **59**, 11936; (b) Wunnicke, O., Mavropoulos, Ph., Zeller, R., Dederichs, P.H. and Grundler, D. (2002) *Physical Review B-Condensed Matter*, **65**, 241306.

80 (a) Ciuti, C., McGuire, J.P. and Sham, L.J. (2002) *Applied Physics Letters*, **81**, 4781; (b) Ciuti, C., McGuire, J.P. and Sham, L.J. (2002) *Physical Review Letters*, **89**, 156601.

81 Stephens, J., Berezovsky, J., Kawakami, R.K., Gossard, A.C. and Awschalom, D.D. (2004) cond-mat/0404244.

82 Žutić, I., Fabian, J. and Das Sarma S. (2002) *Physical Review Letters*, **88**, 066603.

83 Epstein, R.J., Malajovich, I., Kawakami, R.K., Chye, Y., Hanson, M., Petroff, P.M., Gossard, A.C. and Awschalom, D.D. (2002) *Physical Review B-Condensed Matter*, **65**, 121202.

84 (a) Kikkawa, J.M., Smorchkova, I.P., Samarth, N. and Awschalom, D.D. (1997) *Science*, **277**, 1284; (b) Kikkawa J.M. and Awschalom, D.D. (1999) *Nature*, **397**, 139; (c) Malajovich, I., Berry, J.J., Samarth, N. and Awschalom, D.D. (2001) *Nature*, **411**, 770; (d) Hagele, D., Oestreich, M., Rühle, W.W., Nestle, N. and Eberl, K. (1998) *Applied Physics Letters*, **73**, 1580.

85 (a) Mazin, I.I. (1999) *Physical Review Letters*, **83**, 1427; (b) Moruzzi, V.L., Janak, J.F. and Williams, A.R. (1978) *Calculated Electronic Properties of Metals*, Pergamon, New York.

86 Levanyuk A.P. and Osipov, V.V. (1981) *Soviet Physics Uspekhi*, **24**, 3. *Usp Fiz Nauk*, (1981) **133**, 427.

4
Physics of Computational Elements

Victor V. Zhirnov and Ralph K. Cavin

4.1
The Binary Switch as a Basic Information-Processing Element

4.1.1
Information and Information Processing

Information can be defined as a technically quantitative measure of distinguishability of a physical subsystem from its environment [1]. One way to create distinguishable states is by the *presence* or *absence* of material particles (information carrier) in a given location. For example, one can envision the representation of information as an arrangement of particles at specified physical locations as for instance, the depiction of the acronym "IBM" by atoms placed at discrete locations on the material surface (Figure 4.1).

Information of an arbitrary type and amount – such as letters, numbers, colors, or graphics specific sequences and patterns – can be represented by a combination of just two distinguishable states [1–3]. The two states – which are known as binary states – are usually marked as state 0 and state 1. The maximum amount of information, which can be conveyed by a system with just two states is used as a unit of information known as 1 bit (abbreviated from "binary digit"). A system with two distinguishable states forms a basis for *binary switch*.

The binary switch is a fundamental computational element in information-processing systems (Figure 4.2) which, in its most fundamental form, consists of:

- two states 0 and 1, which are equally attainable and distinguishable
- a means to control the change of the state (a gate)
- a means to read the state
- a means to communicate with other binary switches.

Nanotechnology. Volume 3: Information Technology I. Edited by Rainer Waser

ISBN: 978-3-527-31738-7

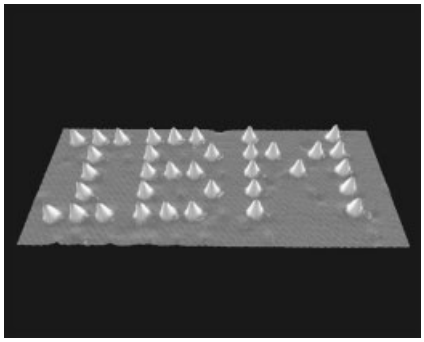


Figure 4.1 "IBM" representation via atoms. (Courtesy of IBM: http://www.almaden.ibm.com/vis/stm/atomo.html).

4.1.2 Properties of an Abstract Binary Information-Processing System

An elementary binary information-processing system consists of N binary switches connected in a certain fashion to implement a function. Each binary switch is characterized by a dimension L and a switching time t_{sw}. A related dimensional characteristic is the packing density (the number of binary switches per unit area). In order to increase the packing density, n_{bit}, the characteristic dimension, L, of the binary switch must decrease:

$$n_{bit} \sim \frac{1}{L^2} \tag{4.1}$$

Another fundamental characteristic of a binary switch is the *switching energy*, E_{sw}.

One indicator of the ultimate performance of an information processor, realized as an interconnected system of binary switches, is the *maximum binary throughput* (BIT); that is, the number of binary transitions per unit time per unit area.

$$\text{BIT} = \frac{n_{bit}}{t_{sw}} \tag{4.2}$$

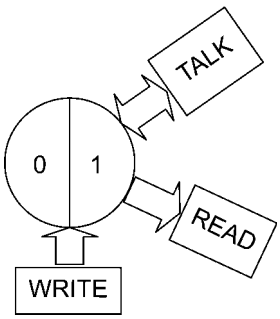


Figure 4.2 The constituents of an abstract binary switch.

One can increase the binary throughput by increasing the number of binary switches per unit area, n_{bit}, and/or decreasing the switching time – that is, the time to transition from one state to the other, $t_{sw.}$

It should be noted that, as each binary switching transition requires energy E_{sw}, the total power dissipation growth is in proportion to the information throughput:

$$P = \frac{n_{bit}}{t_{sw}} \cdot E_{sw} = \mathrm{BIT} \cdot E_{sw} \tag{4.3}$$

The above analysis does not make any assumptions on the material system or the physics of switch operation. In the following sections we investigate the fundamental relations for n_{bit}, t_{sw}, E_{sw} and the corresponding implications for the computing systems.

4.2
Binary State Variables

4.2.1
Essential Operations of an Abstract Binary Switch

The three essential properties of a binary switch are *Distinguishability*, *Controllability* and *Communicativity*. It is said that a binary switch is *Distinguishable* if – and only if – the binary state (0 or 1) can be determined with an acceptable degree of certainty by a measurement (READ operation). The binary switch is *Controllable* if an external stimulus can reliably change the state of the system from 0 to 1 or from 1 to 0 (WRITE operation). The binary switch is *communicative* if it is capable of transferring its state to other binary switches (TALK operation).

4.2.2
The Use of Particles to Represent Binary Information

Information-processing systems represent system states in terms of physical variables. One way to create physically distinguishable states is by the *presence* or *absence* of material particles or fields in a given location. Figure 4.3a illustrates an abstract model for a binary switch the state of which is represented by different positions of a material particle. In principle, the particle can possess arbitrary mass, charge, and so on. The only two requirements for the implementation of a particle-based binary switch are the ability to detect the presence/absence of the particle in, for example the location x_1, and the ability to move the particle from x_0 to x_1 and from x_1 to x_0.

$\Pi_{correct}$ is defined as the probability that the binary switch is in the correct state at an arbitrary time after the command to achieve that state is given. (Alternatively, one can use the probability of error $\Pi_{err} = 1 - \Pi_{correct}$). A necessary condition for the distinguishability of a binary switch is

$$\Pi_{correct} > \Pi_{err} \tag{4.4}$$

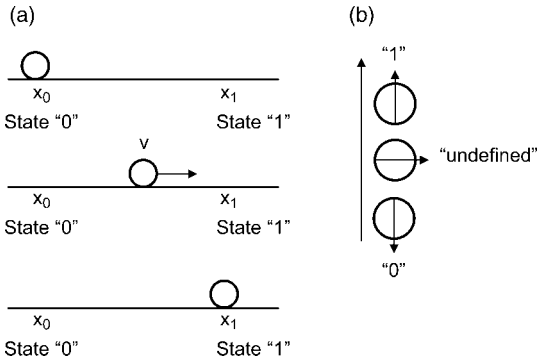


Figure 4.3 An abstract model for the operation of a binary switch formed (a) by different locations of material particles and (b) by opposite direction of the electron spin magnetic moment.

Or equivalently:

$$\Pi_{\text{err}} < 0.5 \tag{4.5}$$

As will be discussed below, in the physical realizations of binary switches, there always is some error probability ($\Pi_{\text{err}} > 0$) in the operation of the switch. As the error probability cannot exceed 0.5, in the following analysis we will use the condition in Equation 2.5 to estimate the parameters of a binary switch in the limiting case.

An elementary switching operation of a binary switch consists of three distinct steps. For example, consider the switch in Figure 4.3a switching from "0" to "1". The three steps are: (i) the initial STORE "0" mode; (ii) the transition CHANGE "0–1" mode; and (iii) the final STORE "1" mode. All three modes have characteristic times and can be described by the coordinate and velocity of the information carrier/material particle.

In STORE "0" the particle must be located in position $x = x_0$ and remain there for the time T_{state}. In CHANGE "0–1" mode, the state is *undefined*, as the particle is in the transition from x_0 to x_1 with a velocity $v_{01} > 0$ (for simplicity, velocity can be taken as the linear dimension of the switch divided by transition time). In STORE "1" ("0") the particle must be located in position $x = x_1$ (x_0) and have lifetime T_{state}. The switching time t_{sw} in this case is given by $t_{\text{sw}} = L/v_{01}$, where $L = x_1 - x_0$ is the linear size of the binary switch.

The question is, what are the requirements for T_{state} and t_{sw} in a binary switch for information processing? As binary logic operates with *two* logic states "0" and "1", while the binary switch has *three* physical states "0", "1" and UNDEFINED (i.e., CHANGE), if the READ operation of binary switch occurs in the UNDEFINED state, an error will result.

The conditions for maximum distinguishability of an ideal binary switch are:

- Unlimited lifetime of each state in the absence of control signal: $T_{\text{state}}^{\max} \to \infty$ in STORE mode. More specifically, for example, synchronous circuits with clock, $T_{\text{state}} \in [T_{\text{clock}}, \infty[$, where T_{clock} is the clock period.

- Fast transition between binary states at the presence of control signal: $t_{sw} \to 0$ in CHANGE mode (say a negligible fraction of the clock period).

As $T_{state}^{max} \to \infty$ in STORE mode, the particle velocity in both 0 and 1 states must be zero, as the particle must be at rest, that is, $v_0 = v_1 = 0$; that is, the kinetic energy $E = \frac{mv^2}{2}$ of the particle should ideally be zero in both STORE modes. In the CHANGE mode, the average particle velocity is $\langle v_{01} \rangle > 0$, and $E > 0$. The switching time can then be estimated as

$$t_{sw} = \frac{L}{\langle v_{01} \rangle} = L\sqrt{\frac{m}{2E}} \tag{4.6}$$

Equation 2.6 sets a limit for the switching speed in the non-zero distance case (non-relativistic approximation). Note that in binary switch operation, an amount of energy E must be supplied to the particle before the CHANGE operation begins, and taken out of the particle after the CHANGE.

If energy remains in the system, the information-bearing particle will oscillate between the two states with the period $2t_{sw.}$ In an oscillating system, if friction is neglected, then energy is preserved (no dissipation) but the information state is not: the lifetime of each binary state $T_{state} \to 0$. The conditions (i) and (ii) of maximum distinguishability will be violated and such a system cannot act as a binary switch.

Alternatively, if one wishes to preserve the binary state (i.e., $T_{state} > 0$), the energy must be rapidly taken out from the system. The time of energy removal t_{out} must be less then half of the switching time: $t_{out} < \frac{t_{sw}}{2}$, otherwise an unintended transition to another state may occur. One rapid way to remove energy is by *thermal dissipation to the environment*. If instead, the aim is to remove the energy in a controllable manner, for example for a possible re-use, a faster switch will be needed which, according to Equation 4.6 will require a greater energy for its operation. It is concluded that non-zero energy dissipation is a necessary attribute of binary switch operations.

The above analysis considers binary switches, with states represented by the *presence* or *absence* of material particles (the information-defining particle or information carrier) in a given locations, for example the utilization of electrons as information carriers. As mentioned above, the electromagnetic field also can, in principle, be used to represent information. For example, a popular candidate is a binary switch that uses the electron spin magnetic moment, when the two opposing directions of the magnetic field represent "0" and "1" (Figure 4.3b).

4.3
Energy Barriers in Binary Switches

4.3.1
Operation of Binary Switches in the Presence of Thermal Noise

Consider again, a binary switch where the binary state is represented by particle location (see Figure 4.3a). Until now, it has been assumed that the information-

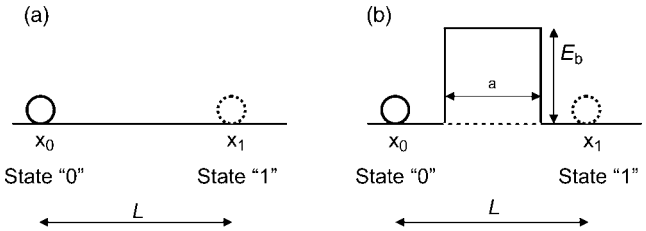


Figure 4.4 Illustration of an energy barrier to preserve the binary states in the presence of noise.

defining particle in the binary switch has zero velocity/kinetic energy, prior to a WRITE command. However, each material particle *at equilibrium* with the environment possesses kinetic energy of $1/2\ k_B T$ per degree of freedom due to thermal interactions, where k_B is Boltzmann's constant and T is temperature. The permanent supply of thermal energy to the system occurs via the mechanical vibrations of atoms (phonons), and via the thermal electromagnetic field of photons (background radiation).

The existence of random mechanical and electromagnetic stimuli means that the information carrier/material particle located in x_0 (Figure 4.4a) has a non-zero velocity in a non-zero-T environment, and that it will spontaneously move from its intended location. According to Equation 4.6, the state lifetime in this case will be

$$T_{\text{state}} \sim L\sqrt{\frac{m}{k_B T}} \tag{4.7}$$

For an electron-based switch ($m = m_e = 9.11 \times 10^{-31}$ kg) of length $L = 1\,\mu$m at $T = 300$ K. Equation 4.8 gives T_{state} as ~15 ps, and hence the time before the system would lose distinguishability would be very small.

In order to prevent the state from changing randomly, it is possible to construct energy barriers that limit particle movements. The energy barrier, separating the two states in a binary switch is characterized by its height E_b and width a (Figure 4.4b).

The barrier height, E_b, must be large enough to prevent spontaneous transitions (errors). Two types of unintended transition can occur: "classical" and "quantum". The "classical" error occurs when the particle jumps over barrier, and this can happen if the kinetic energy of the particle E is larger than E_b. The corresponding "classic" error probability, Π_C, is obtained from the Boltzmann distribution as:

$$\Pi_C = \exp\left(-\frac{E_b}{k_B T}\right) \tag{4.8}$$

The presence of energy barrier of width a sets the minimum device size to be $L_{\min} \geq a$.

4.3.2 Quantum Errors

Another class of errors, termed "quantum errors", occur due to quantum mechanical tunneling through the barrier of finite width a. If the barrier is too narrow, then

spontaneous tunneling through the barrier will destroy the binary information. The conditions for significant tunneling can be estimated using the Heisenberg uncertainty principle, as is often carried out in texts on the theory of tunneling [4]:

$$\Delta x \Delta p \geq \frac{\hbar}{2} \tag{4.9}$$

The uncertainty relationship of Equation 4.9 can be used to estimate the limits of distinguishability. Consider again a "two-well" bit in Figure 4.4b. As is known from quantum mechanics, a particle can pass (tunnel) through a barrier of finite width, even if the particle energy is less than the barrier height, E_b. An estimate of how thin the barrier must be to observe tunneling can be made from Equation 4.9; for a particle at the bottom of the well, the uncertainty in momentum is $\sqrt{2mE_b}$, which gives:

$$\sqrt{2mE_b}\Delta x \approx \frac{\hbar}{2} \tag{4.10}$$

Equation 4.10 states that by initially setting the particle on one side of the barrier, it is possible to locate the particle on either side, with high probability, if Δx is of the order of the barrier width a. That is, the condition for losing distinguishability is $\Delta x \lesssim a$, and the minimum barrier width is:

$$a_{\min} = a_H \approx \frac{\hbar}{2\sqrt{2mE_b}}, \tag{4.11}$$

where a_H is the *Heisenberg distinguishability length* for "classic to quantum transition".

For $a < a_H$, the tunneling probability is significant, and therefore particle localization is not possible. In order to estimate the probability of tunneling, Equation 4.10 can be re-written, taking into account the tunneling condition $a \leq \Delta x$:

$$\sqrt{2m}(a\sqrt{E_b}) \leq \frac{\hbar}{2} \tag{4.12}$$

From Equation 4.12, the "tunneling condition" can also be written in the form

$$1 - \frac{2\sqrt{2m}}{\hbar} a\sqrt{E_b} \geq 0, \tag{4.13}$$

Since for small x, $e^{-x} \sim 1 - x$, the tunneling condition then becomes

$$\exp\left(- \frac{2\sqrt{2m}}{\hbar} \cdot a \cdot \sqrt{E_b} \right) \geq 0 \tag{4.14}$$

The left-hand side of Equation 4.14 has the properties of probability. Indeed, it represents the tunneling probability through a rectangular barrier given by the Wentzel–Kramers–Brillouin (WKB) approximation [5]:

$$\Pi_{WKB} \sim \exp\left(- \frac{2\sqrt{2m}}{\hbar} \cdot a \cdot \sqrt{E_b} \right) \tag{4.15a}$$

This equation also emphasizes the parameters controlling the tunneling process, which include the barrier height E_b and barrier width a, as well as the mass m of the

information-bearing particle. If separation between two wells is less than a, the structure of Figure 4.4b would allow significant tunneling. In fact, it is instructive to examine the physical meaning of Equation 4.11, which we marked as the condition of significant tunneling or "classic to quantum transition". Substituting Equation 4.11 into Equation 4.15a provides an estimate for tunneling probability through a rectangular barrier of width a_H:

$$\Pi_{WKB} \sim \exp\left(- \frac{2\sqrt{2m}}{\hbar} \cdot a_H \cdot \sqrt{E_b} \right) = \exp(-1) \approx 0.37 \tag{4.15b}$$

Thus, the Heisenberg distinguishability length a_H from Equation 4.11 corresponds to a tunneling probability of approximately 37%.

4.3.3
A Combined Effect of Classical and Quantum Errors

As discussed in the previous sections, there are two mechanisms of spontaneous transitions (errors) in binary switching: the over-barrier transition ("classic" error); and through-barrier tunneling ("quantum" error). The probabilities of the classic and quantum errors are given by Equations 4.8 and 4.15a, respectively. The joint error probability of the two mechanisms is [3]:

$$\Pi_{err} = \Pi_C + \Pi_Q - \Pi_C \cdot \Pi_Q \tag{4.16}$$

Or, from Equations 4.8 and 4.15a, we obtain:

$$\Pi_{err} = \exp\left(- \frac{E_b}{k_B T} \right) + \exp\left(\frac{2\sqrt{2m}}{\hbar} \cdot a\sqrt{E_b} \right) - \exp\left(- \frac{\hbar E_b + 2ak_B T\sqrt{2mE_b}}{\hbar k_B T} \right) \tag{4.17}$$

4.4
Energy Barrier Framework for the Operating Limits of Binary Switches

4.4.1
Limits on Energy

The minimum energy of binary transition is determined by the energy barrier. The work required to suppress the barrier is equal or larger than E_b; thus, the minimum energy of binary transition is given by the minimum barrier height in binary switch. The minimum barrier height can be found from the distinguishability condition [Equation 4.5], which requires that the probability of errors $\Pi_{err} < 0.5$. First, the case is considered when only "classic" (i.e., thermal) errors can occur. In this case, according to Equation 4.8:

$$\Pi_{err} = \Pi_C = \exp\left(- \frac{E_b}{k_B T} \right) \tag{4.18}$$

These classic transitions represent the thermal (Nyquist–Johnson) noise. By solving Equation 4.18 for $\Pi_{\text{err}} = 0.5$, we obtain the Boltzmann's limit for the minimum barrier height, E_{bB}

$$E_{\text{bB}} = k_{\text{B}} T \ln 2 \approx 0.7 k_{\text{B}} T \tag{4.19}$$

Equation 4.19 corresponds to the minimum barrier height, the point at which distinguishability of states is completely lost due to thermal over-barrier transitions. In deriving Equation 4.19, tunneling was ignored – that is, the barrier width is assumed to be very large, $a \gg a_{\text{H}}$.

Next, we consider the case where only quantum (i.e., tunneling) errors can occur. In this case, according to Equation 4.15a

$$\Pi_{\text{err}} = \Pi_{\text{Q}} \sim \exp\left(- \frac{2\sqrt{2m}}{\hbar} \cdot a \cdot \sqrt{E_{\text{b}}} \right) \tag{4.20}$$

By solving Equation 4.20 for $\Pi_{\text{err}} = 0.5$, we obtain the Heisenberg's limit for the minimum barrier height, E_{bH}

$$E_{\text{bH}} = \frac{\hbar^2}{8ma^2} (\ln 2)^2 \tag{4.21}$$

Equation 4.21 corresponds to a narrow barrier, $a \sim a_{\text{H}}$, the point at which distinguishability of states is lost due to tunneling transitions. In deriving Equation 4.21, over-barrier thermal transitions were ignored – that is, the temperature was assumed to be close to absolute zero, $T \to 0$.

Now, we consider the case when both thermal and tunneling transitions contribute to the errors in a binary switch. In this case, the total error probability is given by Equation 4.17. An approximate solution of Equation 4.17 for $\Pi_{\text{err}} = 0.5$ is

$$E_{\text{b min}} = k_B T \ln 2 + \frac{\hbar^2}{8ma^2} (\ln 2)^2 \tag{4.22}$$

Equation 4.22 provides a generalized value for minimum energy per switch operation at the limits of distinguishability, that takes into account both classic and quantum transport phenomena. The graph in Figure 4.5 shows the numerical solution of Equation 4.17 and its approximate analytical solution given by Equation 4.22 for $\Pi_{\text{err}} = 0.5$. Is it clearly seen that for $a > 5$ nm, the Boltzmann's limit, $E_{\text{bB}} = k_{\text{B}} T \ln 2$, is a valid representation of minimum energy per switch operation, while for $a < 5$ nm, the minimum switching energy can be considerably larger.

4.4.2 Limits on Size

The minimum size of a binary switch L cannot be smaller then the distinguishability length a_{H}. From both Equations 4.11 and 4.19, one can estimate the Heisenberg's length for the binary switch operation at the Boltzmann's limit of energy:

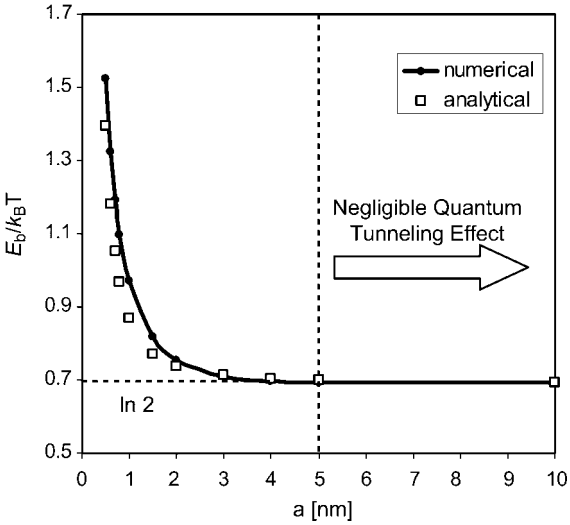


Figure 4.5 Minimum energy per switch operation as a function of minimum switch size.

$$a_{\mathrm{HB}} = \frac{\hbar}{2\sqrt{2mk_{\mathrm{B}}T\ln 2}} \tag{4.23}$$

For electrons ($m = m_{\mathrm{e}}$) at $T = 300\,\mathrm{K}$, we obtain $a_{\mathrm{HB}} \sim 1\,\mathrm{nm}$.

The distinguishability length a_{HB} defines both the minimum size of the switch and the separation between the two neighboring switches. Thus, the maximum density of binary switches is:

$$n_{\max} \leq \frac{1}{(2a_{\mathrm{HB}})^2} \tag{4.24}$$

For electron-based binary switches at 300 K ($a_{\mathrm{HB}} \sim 1\,\mathrm{nm}$) and $n_{\max} \sim 10^{13}\,\mathrm{cm}^{-2}$.

4.4.3 Limits on Speed

The next pertinent question is the minimum switching time, $\tau_{\min}$, which can be derived from the Heisenberg relationship for time and energy:

$$\Delta E \Delta t \geq \frac{\hbar}{2} \tag{4.25a}$$

or

$$\tau_{\min} \cong \frac{\hbar}{2\Delta E} \tag{4.25b}$$

Equation 4.25b is an estimate for the maximum speed of dynamic evolution [6] or the maximum passage time [7]. It represents the zero-length approximation for the

switching speed, in contrast with Equation 4.6, which is distance-dependent. If $L \sim a_H$, then Equation 4.6 converges to Equation 4.25b.

For the Boltzmann's limit, $E = E_{bB}$ [Equation 4.19], we obtain

$$\tau_{min\,B} \cong \frac{\hbar}{2k_B T \ln 2} \approx 2 \cdot 10^{-14}\ \mathrm{s} \tag{4.26}$$

It should be noted that Equation 4.26 is applicable to all types of device, and no specific assumptions were made about any physical device structure.

4.4.4
Energy Dissipation by Computation

Using Equations 4.1–4.3 allows one to estimate power dissipation by a chip containing the smallest binary switches, $L \sim a_{HB}$ [Equation 4.23], packed to maximum density [Equation 4.24] and operating at the lowest possible energy per bit [Equation 4.19].

The power dissipation per unit area of this limit technology is given by:

$$P = \frac{n_{max} E_{min\,B}}{\tau_{min\,B}} \sim \frac{10^{13}\ \mathrm{cm}^{-2} \cdot 3 \cdot 10^{-21}\ \mathrm{J}}{2 \cdot 10^{-14}\ \mathrm{s}} \sim 2 \cdot 10^{6}\ \frac{\mathrm{W}}{\mathrm{cm}^2} \tag{4.27}$$

The energy density bound in the range of $\mathrm{MW\,cm^{-2}}$ obtained by invoking $k_B T \ln 2$ as the lower bound for the device energy barrier height is an astronomic number. If known cooling methods are employed, it appears that that heat-removal capacity of several hundred $\mathrm{W\,cm^{-2}}$ represents a practically achievable limit. The practical usefulness of alternative electron-transport devices may be derived from lower fabrication costs or from specific functional behavior; however, the heat removal challenge will remain.

4.5
Physics of Energy Barriers

The first requirement for the physical realization of any arbitrary switch is the creation of distinguishable states within a system of such material particles. The second requirement is the capability for a conditional change of state. The properties of distinguishability and conditional change of state are two fundamental and essential properties of a material subsystem that represents binary information. These properties are obtained by creating and control energy barriers in a material system.

The physical implementation of an energy barrier depends on the choice of the state variable used by the information processing system. The energy barrier creates a local change of the potential energy of a particle from a value U_1 at the generalized coordinate q_1 to a larger value U_2 at the generalized coordinate q_2. The difference $\Delta U = U_2 - U_1$ is the barrier height. In a system with an energy barrier, the force exerted on a particle by the barrier is of the form $F = \frac{\partial U}{\partial q}$. A simple illustration of a one dimensional barrier in linear spatial coordinates, x, is shown in Figure 4.6a. It should

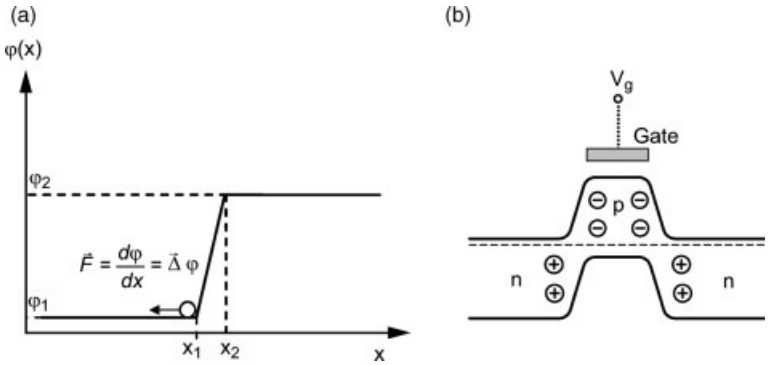


Figure 4.6 An illustration to the energy barrier in a material system. (a) Abstraction; (b) physical implementation by doping of a semiconductor.

be noted, that the spatial energy changes in potential energy require a finite spatial extension ($\Delta x = x_2 - x_1$) of the barrier (Figure 4.6a and b). This spatial extension defines a minimum dimension of energy barrier, $a_{\min}$: $a_{\min} > 2\Delta x$. In this section, we consider the physics of barriers for electron charge, electron spin, and optical binary switches.

4.5.1 Energy Barrier in Charge-Based Binary Switch

For electrons, the basic equation for potential energy is the Poisson equation

$$\nabla^2 \varphi = -\frac{\rho}{\varepsilon_0}, \tag{4.28}$$

where ρ is the charge density, $\varepsilon_0 = 8.85 \times 10^{-12}\,\mathrm{F\,m^{-1}}$ is the permittivity of free space, and φ is the potential: $\varphi = U/e$. According to Equation 4.28, the presence of an energy barrier is associated with changes in charge density in the barrier region. The barrier-forming charge is introduced in a material, for example by the doping of semiconductors. This is illustrated in Figure 4.6b, for a silicon n-p-n structure where the barrier is formed by ionized impurity atoms such as P^+ in the n-region and B^- in the p-region. The barrier height E_{b0} depends on the concentration of the ionized impurity atoms [8]:

$$E_{b0} \approx k_B T \ln \frac{N_A^- N_B^+}{n_i}, \tag{4.29}$$

where, N_A^-, N_B^+, and n_i are the concentration of negatively charged impurities (acceptors), positively charged impurities (donors), and the intrinsic carrier concentration in a semiconductor, respectively.

The minimum barrier extension is given by the Debye length [8]:

$$L_D \approx \sqrt{\frac{\varepsilon_0 \varepsilon k_B T}{e^2 N_D}}, \tag{4.30}$$

where ε is the relative dielectric permittivity of a semiconductor, and $N_D = N_A^- = N_B^+$ (abrupt p-n junction approximation). The maximum concentration of electrically active dopants N_{max} is close to the density of states in the conduction, N_c, and valence bands, N_v of the semiconductor. For silicon, $N_{max} \sim 10^{19}\,cm^{-3}$ [8], and it follows that $L_{Dmin} \sim 1.3\,nm$. The minimum barrier length therefore is $a_{min} = 2L_{Dmin} \sim 2.6\,nm$.

The energy diagram of Figure 4.6b is typical of many semiconductor devices, for example, field effect transistors (FET). The barrier region corresponds to the FET channel, while the wells correspond to the source and drain. In order to enable electron movement between the source and drain, the barrier height must be decreased (ideally suppressed to zero). To do this, the amount of charge in the barrier region needs to be changed, according to Equation 4.28. A well-known relationship connects the electrical potential difference $\Delta\varphi = V$ and charge, Δq, through capacitance

$$C = \frac{\Delta q}{\Delta\varphi} \tag{4.31}$$

In field-effect devices, in order to change the charge distribution in the barrier region – and hence lower the barrier – a voltage is applied to an external electrode (gate), which forms a capacitor with the barrier region (in bipolar devices, external charge is injected in the barrier region to control the barrier height). When voltage V_g is applied to the barrier region, the barrier will change from it initial height E_{b0} (determined by impurity concentration):

$$E_b = E_{b0} - eV_g \tag{4.32}$$

The voltage needed to suppress the barrier from E_{b0} to zero (the threshold voltage) is:

$$V_t = \frac{E_{b0}}{e} \tag{4.33}$$

Thus, the operation of all charge transport devices involves charging and discharging capacitances to change barrier height, thereby controlling charge transport in the device. When a capacitor C is charged from a constant voltage power supply V_g, the energy E_{dis} is dissipated, that is, it is converted into heat [9]:

$$E_{dis} = \frac{CV_g^2}{2} \tag{4.34}$$

The minimum energy needed to suppress the barrier (by charging the gate capacitor) is equal to the barrier height E_b. Restoration of the barrier (by discharging gate capacitance) also requires a minimum energy expenditure of E_b. Thus, the minimum energy required for a full switching cycle is at least $2E_b$.

It should be noted that in the solid-state implementation of binary switch, the number of electrons in both wells is large. This is different from an abstract system having only one electron (see above). In a multi-electron system, the electrons strike the barrier from both sides, and the binary transitions are determined by the net electron flow, as shown in Figure 4.7.

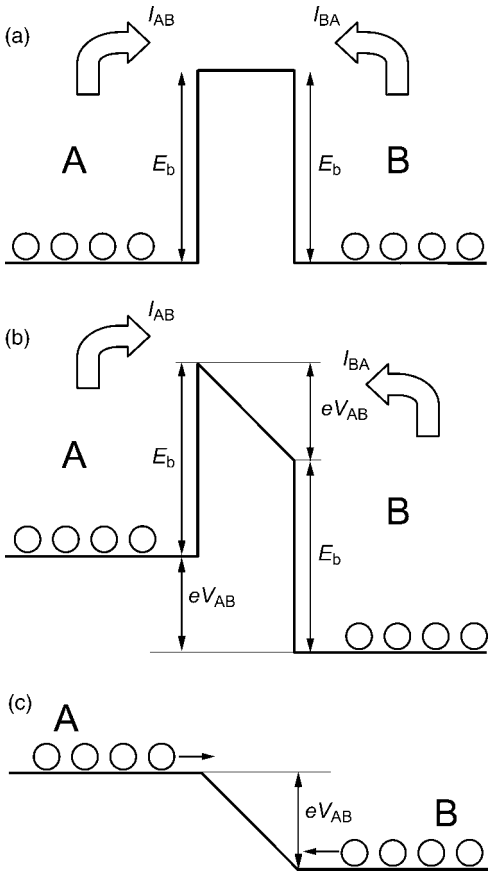


Figure 4.7 The fundamental operation of multi-electron binary switch. (a) No energy difference between wells **A** and **B**, resulting in a symmetric energy diagram. (b) Energy asymmetry is created due to energy difference eV_{AB} between the wells **A** and **B**. (c) State CHANGE operation: the barrier height E_b is suppressed by applying gate potential V_g, while energy difference eV_{AB} between the wells **A** and **B**.

Let N_0 be the number of electrons that strike the barrier per unit time. Thus, the number of electrons N_A that transition over the barrier from well **A** per unit time is

$$N_A = N_0 \exp\left(-\frac{E_b}{k_B T}\right) \tag{4.35}$$

The corresponding current I_{AB} is

$$I_{AB} = e \cdot N_A = eN_0 \exp\left(-\frac{E_b}{k_B T}\right) \tag{4.36}$$

Electrons in well **B** also can strike the barrier and therefore contribute to the over-barrier transitions with current I_{BA}. Thus, the net over-barrier current is

$$I = I_{AB} - I_{BA} \tag{4.37}$$

The energy diagram of Figure 4.7a is symmetric, hence $I_{AB} = I_{BA}$, and $I = 0$. Therefore, no binary transitions occur in the case of symmetric barrier. In order to enable the rapid and reliable transition of an electron from well **A** to well **B**, an energy asymmetry between two wells must be created. This is achieved by energy difference eV_{AB} between the wells **A** and **B** (Figure 4.7b).

For such an asymmetric diagram, the barrier height for electrons in the well **A** is E_b, and for electrons in the well **B** is $(E_b + eV_{AB})$. Correspondingly, from Equations 4.36 and 4.37 the net current is

$$\begin{aligned} I &= eN_0 \exp\left(-\frac{E_b}{k_B T}\right) - eN_0 \exp\left(-\frac{E_b - eV_{AB}}{k_B T}\right) \\ &= eN_0 \exp\left(-\frac{E_b}{k_B T}\right)\left[1 - \exp\left(-\frac{eV_{AB}}{k_B T}\right)\right] \end{aligned} \tag{4.38a}$$

By substituting Equation 4.32 for E_b, we obtain:

$$I = eN_0 \exp\left(-\frac{E_{b_0} - eV_g}{k_B T}\right)\left[1 - \exp\left(-\frac{eV_{AB}}{k_B T}\right)\right] \tag{4.38b}$$

By expressing E_{b0} as eV_t from Equation 4.33 and using the conventional notations $I = I_{ds}$ (source-drain current) and $V_{AB} = V_{ds}$ (source-drain voltage), we obtain the equation for the subthreshold I–V characteristics of FET [10]:

$$I_{ds} = I_0 \exp\left(\frac{e(V_g - V_t)}{k_B T}\right)\left[1 - \exp\left(-\frac{eV_{ds}}{k_B T}\right)\right] \tag{4.38c}$$

An example plot of Equation 4.38c is shown in Figure 4.8.

The minimum energy difference between the wells, eV_{ABmin}, can be estimated based on the distinguishability arguments for CHANGE operation, when is E_b is

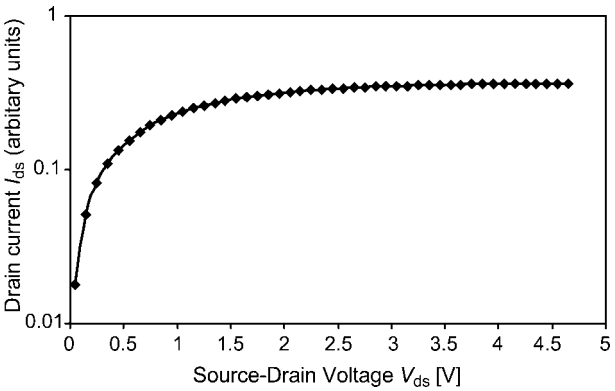


Figure 4.8 A source-drain *I-V* characteristic derived from the energy barrier model for a charge-based binary switch.

suppressed by applying the gate voltage, for example $E_b = 0$ (Figure 4.7c). For a successful change operation, the probability that each electron flowing from well **A** to well **B** is not counterbalanced by another electron moving from well **B** to well **A** should be less than 0.5 in the limiting case. The energy difference eV_{AB} forms a barrier for electrons in well **B**, but not for electrons in well **A**, and therefore, from Equation 4.18 we obtain

$$eV_{ABmin} = E_{b_{min}} = k_B T \ln 2 \tag{4.39}$$

If N is the number of electrons involved in the switching transition between two wells, the total minimum switching energy is

$$E_{SWmin} = 2E_b + NeV_{AB} = (N+2)k_B T n2 \tag{4.40a}$$

If $N = 1$, then

$$E_{SW_{min}} = 3k_B T \ln 2 \approx 10^{-20}\,\text{J} \tag{4.40b}$$

4.5.2 Energy Barrier in Spin-Based Binary Switch

In addition to charge, e, electrons possess intrinsic angular momentum (spin). As result, they also possess a permanent magnetic moment [11]:

$$\mu_s = \pm\frac{1}{2} g \cdot \mu_B \tag{4.41}$$

where μ_B is the Bohr magneton, $\mu_B = \frac{e\hbar}{2m_e}$, and g is the coupling constant known as the Landé gyromagnetic factor or g-factor. For free electrons and electrons in isolated atoms $g_0 = 2.00$. In solids, consisting of a large number of atoms, the effective g-factor can differ from g_0.

The energy of interaction, $E_{\mu-B}$, between a magnetic moment $\vec{\mu}$ and a magnetic field $\vec{B}$ is:

$$E_{\mu-B} = -\vec{\mu} \cdot \vec{B} \tag{4.42}$$

For the electron spin magnetic moment in a magnetic field applied in the z direction, the energy of interaction takes two values, depending on whether the electron spin magnetic moment is aligned or anti-aligned with the magnetic field. From Equations 4.41 and 4.42 one can write, assuming $g = 2$

$$\begin{aligned} E_{\uparrow\uparrow} &= -\frac{e\hbar}{2m_e} \cdot B_z \\ E_{\uparrow\downarrow} &= +\frac{e\hbar}{2m_e} \cdot B_z \end{aligned} \tag{4.43}$$

The energy difference between the aligned and anti-aligned states represents the energy barrier in the spin binary switch and is

$$E_b = E_{\uparrow\downarrow} - E_{\uparrow\uparrow} = 2\mu_B B_z \tag{4.44}$$

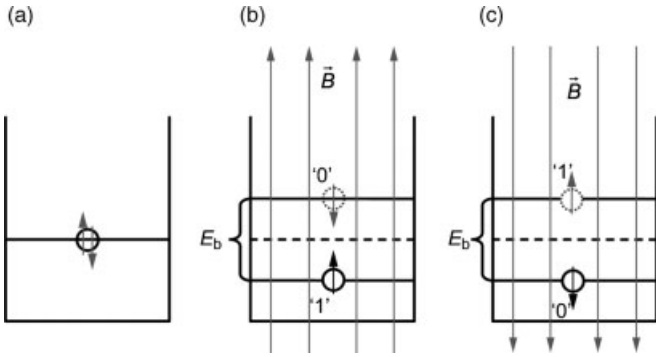


Figure 4.9 An abstract model of a single spin binary switch. (a) $B = 0$, the two states are indistinguishable. (b, c) $B \neq 0$, the two binary states are separated by the energy gap E_b.

Equations 4.43 and 4.44 represent a physical phenomenon known as Zeeman splitting [11]. The operation of a single spin binary switch is illustrated in Figure 4.9. In the absence of an external magnetic field, there is equal probability that the electron has magnetic moment $+\mu_B$ or $-\mu_B$; that is, the two states are indistinguishable (Figure 4.9a). When an external magnetic field is applied (Figure 4.9b and c), the two states are separated in energy. The lower energy state has higher probability of population and this represents the binary state "1" (Figure 4.9b) or "0" (Figure 4.9c) in this system. Binary switching occurs when the external magnetic field changes direction, as shown in Figure 4.9b and c.

This abstraction, while very simple, applies to all types of spin devices, at equilibrium with the thermal environment, including for example proposed spin transport devices [12–14] and coupled spin-polarized quantum dots [15].

The barrier-forming magnetic field B can be either a built-in field formed by a material layer with a permanent magnetization, or created by an external source (e.g., an electromagnet). In both cases, the change in the direction of the external magnetic field required for binary switching is produced by an electric current pulse in one of two opposite directions. The need for two opposite directions of electrical current flow requires additional electrical binary switches to control current flow in the electromagnet.

Some generic electrical circuits to manipulate the barrier height are a spintronic binary switch are shown in Figure 4.10. These require four electrical binary switches in the case of single power supply (Figure 4.10a), or two binary switches in the case of two power supplies (Figure 4.10b).

Thus, each spin-based binary switch requires two or four "servant" charge-based binary switches. This will result in larger area per device and also a larger energy consumption per operation, as compared to the charge-based switches. As the minimum switching energy of one charge-based device is $\sim 3k_BT$ [according to Equation 4.40b], the minimum switching energy of a spin device E_{spin} for single power supply scheme is

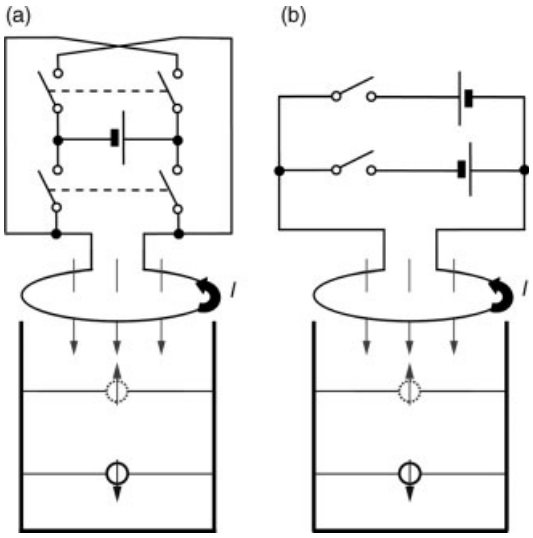


Figure 4.10 A generic electrical circuit to control spin binary switch. (a) Single power supply scheme; (b) double power supply scheme.

$$E_{\text{spin}} = E_{\text{M}} + 12k_{\text{B}}T \tag{4.45}$$

where E_{M} is the energy required to generate the magnetic field to change the spin state.

One possible way to address the "electrical" challenge for spin devices is to change the spin control paradigm. The paradigm described above uses a system of binary switches, each of which can be controlled independently by an external stimulus, and each switch can, in principle control any other switch in the system. However, it is not clear whether a spin state-based device can be used to control the state of subsequent spin devices without going through an electrical switching mechanism as discussed above. Although it is possible that local interactions may be used to advantage for this purpose [15, 16], feasibility assessments of these proposals in general information processing applications are clearly required.

Let us now consider a hypothetical single spin binary switch that, ideally, might have atom-scale dimensions. At thermal equilibrium there is a probability of spontaneous transition between spin states "1" and "0" in accordance with Equation 4.18. Correspondingly, for $\Pi_{\text{err}} < 0.5$, according to Equation 4.19 the energy separation between to state should be larger than $k_{\text{B}}T \ln 2$

$$2\mu_{\text{B}}B_{\text{min}} = k_{\text{B}}T \ln 2 \tag{4.46}$$

From Equations 4.19 and 4.44 we can obtain the minimum value of B for a switch operation

$$B_{\text{min}} = \frac{k_{\text{B}}T \ln 2}{2\mu_{\text{B}}} = \frac{m_{\text{e}}}{e\hbar} k_{\text{B}}T \ln 2 \tag{4.47}$$

Table 4.1 Maximum magnetic fields and limiting factors (from Ref. [17]).

Magnet	B_{max} [T]	Limiting factor
Conventional magnetic-core electromagnet	~2	Permeability saturation of the of magnetic core
Steady-field air-core NbTi and Nb_3Sn superconducting electromagnet	~20	Critical magnetic field
Steady-field air-core water-cooled electromagnet	~30	Joule heating
Pulsed-field hybrid electromagnet	~50	Maxwell stress

At $T = 300$ K, Equation 4.47 results in $B_{min} \approx 155$ Tesla (T), which is much larger than can be practically achieved (a summary of the technologies used to generate high magnetic fields is provided in Table 4.1).

One of the most difficult problems of very high magnetic field is the excessive power consumption and Joule heating in electromagnets. The relationship between power consumption, P, and magnetic field B, in an electromagnet is [17]:

$$P \sim B^2 \tag{4.48}$$

Moreover, as the production of magnetic fields with $B > 10$ T requires many megawatts of power, these magnetic field-production systems have large dimensions and a mass of many tons (see Table 4.2).

Thus, it is concluded that single electron spin devices operating at room temperature would require local magnetic fields greater than have been achieved to date with large-volume apparatus.

In multi-spin systems, it is possible to increase the magnetic moment μ and therefore, to decrease the magnitude of the external magnetic field B required for binary switch operation. The increase of μ may be due to an increase in number of co-aligned spins, which results in collective effects such as paramagnetism and ferromagnetism.

Table 4.2 Examples of practical implementations of the sources of magnetism.

Magnet	B [T]	P	Mass	Comments
Small bar magnet	~0.01	—	~g	—
Small neodymium–iron–boron magnet	~0.2	—	~g	—
Big Magnetic-core electromagnet	~2	~100 W	~kg	—
Steady-field superconducting electromagnet [19]	~16	~MW	~tons	Cryogenic temperatures
Current status of Pulse Magnet Program at National High Magnetic Field Laboratory [18]	60–65	~MW	~tons	Cryogenic temperatures; a 30-min cooling time between shots Lifetime ~400 cycles

4.5.3 Energy Barriers for Multiple-Spin Systems

In a system of N spins in an external magnetic field, there are $N_{\uparrow\uparrow}$ spin magnetic moments parallel to the external magnetic field, and the resulting magnetic moment is

$$\mu = \mu_B \cdot N_{\uparrow\uparrow} = \mu_B N(1 - \Pi_{\text{err}}) = \mu_B N\left(1 - \exp\left(-\frac{\mu_B B}{k_B T}\right)\right) \tag{4.49}$$

As seen above, for all practical cases $\mu_B B \ll k_B T$, and since $(1 - e^x) \approx x$, for $x \to 0$, there results

$$\mu \approx \frac{N\mu_B^2 B}{k_B T} \tag{4.50}$$

Equation 4.50 is known as the Curie law for paramagnetism [11]. From Equations 4.50 and 4.44 one can calculate the minimum number of electron spins required for spin binary switch operating at realistic magnitudes of the magnetic field:

$$N_{\min} \approx \frac{\ln 2}{2}\left(\frac{k_B T}{\mu_B B}\right) \tag{4.51}$$

For example, for $B = 0.1$ T (a small neodymium–iron–boron magnet; see Table 4.1), $N_{\min} \sim 7 \times 10^6$. If the number of electrons with unpaired spins per atom is f (f varies between 1 and 7 for different atoms), then the number of atoms needed is $N_{\min}/f$. Correspondingly, one can estimate the minimum critical dimension $a_{\min}$ of the binary switch as:

$$a_{\min} \sim \left(\frac{N_{\min}}{f \cdot n_V}\right)^{\frac{1}{3}}, \tag{4.52}$$

where n_V is the density of atoms in the material structure. Assuming an atomic density close to the largest known in solids, $n_V = 1.76 \times 10^{23}\ \text{cm}^{-3}$ (the atomic density of diamond) and $B \sim 0.1$ T, we obtain $a_{\min} \sim 41$ nm for $f = 1$ and $a_{\min} \sim 22$ nm for $f = 7$. Thus, it is concluded that for reliable operation at moderate magnetic fields, the physical size of a multi-spin-based binary switch is larger than the ultimate charge-based devices. The effect of collective spin behavior is currently used, for example, in magnetic random access memory (MRAM) [14] and in electron spin resonance (ESR) technologies.

As an example, it is estimated that the energy needed to operate a spin-based binary switch for the case when the magnetic field is produced by electrical current I in a circular loop of wire surrounding the switch, as shown in Figure 4.10. The magnetic field in the center of the loop is [20]:

$$B = \frac{\mu_0 I}{2r}, \tag{4.53}$$

where μ_0 is the magnetic permeability of free space.

In order to switch the external magnetic field, for example from zero to $+\vec{B}$ or from $+\vec{B}$ to $-\vec{B}$, work must be done. If the magnetic field is formed by electrical current I produced by an external voltage source, then the energy dissipated by one half of switching cycle (e.g., the rise from zero to $+\vec{B}$):

$$E_{\text{dis}} = \frac{LI^2}{2} \tag{4.54}$$

where L is electrical inductance. Equation 4.54 is analogous to the energy of $CV^2/2$ dissipated in charge-based devices [Equation 4.34].

If magnetic field needs to be sustained for the time period t, then additional energy will be dissipated due to resistance R, and thus the total energy dissipation is:

$$E_{\text{dis}} = \frac{LI^2}{2} + I^2 R \cdot t \tag{4.55}$$

If the magnetic field does not need to be sustained (e.g., in ferromagnetic devices), after switching the current must be reduced to zero, in which case another amount of energy of Equation 4.54 is dissipated. Thus, the energy expenditure needed to generate the magnetic field E_{M} [see Equation 4.55] in a full switching cycle is

$$E_{\text{M}} = LI^2 \tag{4.56}$$

By definition, electrical inductance L is a proportionality factor connecting magnetic flux Φ and the electric current I that produces the magnetic field [20]:

$$\Phi = L \cdot I \tag{4.57}$$

The magnetic flux in turn is defined as

$$\Phi = B \cdot A \tag{4.58}$$

where A is the area, $\text{A} = \pi r^2$, for the case of circular loop.

From Equations 4.53,4.57 and 4.58, the inductance of a circular loop of radius r is approximately:

$$L \approx \frac{\pi \mu_0 r}{2} \tag{4.59}$$

Equation 4.64 is an approximation assuming constant magnetic field inside the loop and ignoring the effects of wire thickness.[1)]

By combining Equations 4.53,4.56 and 4.59, we obtain:

$$E_M = \frac{2\pi}{\mu_0} r^3 B^2 \tag{4.60}$$

For the minimum device size given by Equation 4.52, and noting that $r \geq a_{\min}$ as obtained from Equations 4.19 and 4.60:

1) An accurate result for the inductance of the circular loop is $L = \mu_0 r[\ln(8r/b) - 7/4]$, where b is the radius of the wire [21]. This differs from Equation 4.59 by a factor of 1.6–3 for a realistic range of r/b ratios of 10–100.

$$E_{\mathrm{M\,min}} = \frac{\pi \ln 2}{\mu_0 n_V}\left(\frac{k_\mathrm{B}T}{\mu_\mathrm{B}}\right)^2 \tag{4.61}$$

For the largest atomic density of solids, $n_\mathrm{V} = 1.76\times10^{-23}\,\mathrm{cm}^{-3}$ (diamond), we obtain:

$$E_{\mathrm{M\,min}} \approx 2\cdot10^{-18}\,\mathrm{J} \approx 480 k_\mathrm{B}T \tag{4.62}$$

4.5.4
Energy Barriers for the Optical Binary Switch

Optical digital computing was – and still is – considered by some as a viable option for massive information processing [22]. Sometimes, it is referred to as "computing at speed of light" [23] and, indeed, photons do move at the speed of light. At the same time, a photon cannot have a speed other then the speed of light, c, and therefore it cannot be confined within a binary switch of finite spatial dimensions.

The minimum dimension of optical switch is given by the wavelength of light, λ. If $a_{\min} < \lambda$, there is high probability that the light will not "sense" the state – that is, the error probability will increase. For visible light, $a_{\min} \sim 400\,\mathrm{nm}$.

Binary state control in the optical switch is accomplished by local changes in optical properties of the medium, such as the refraction index, reflectivity or absorption, while photons are used to read the state. In many cases, the changes in optical properties are related to a rearrangement of atoms under the influence of electrical, optical, or thermal energy. The energy barrier in this case is therefore related to either inter-atomic or inter-molecular bonds. Examples of such optical switches are liquid crystal spatial light modulators [22] and non-linear interference filters [22]. Another example is a change in refractive index as the result of a crystalline-to amorphous phase change, which is used for example in the rewritable CD. The minimum switching energy of this class of optical switches is related to the number of atoms, N, and therefore to the size L. In the limiting case, $L \sim a_{\min} \sim \lambda$. In order for the atoms of an optical switch to have a distinguishable change of their position, the energy supply to each atom should be larger than $k_\mathrm{B}T$. The total switching energy is therefore:

$$E \sim N \cdot k_\mathrm{B}T \tag{4.63}$$

For a minimum energy estimate, we must consider the smallest possible N, which corresponds to an single-atom plane of size λ. If optical switch materials have an atomic density n, then one obtains:

$$E \sim n^{\frac{2}{3}} \cdot \lambda^2 \cdot k_\mathrm{B}T \tag{4.64}$$

For most solids, $n = 10^{22}$–$10^{23}\,\mathrm{cm}^{-3}$. Taking $n = 5\times10^{22}\,\mathrm{cm}^{-3}$, $T = 300\,\mathrm{K}$, and $a_{\min} \sim 400\,\mathrm{nm}$, we obtain $E \sim 10^{14}\,\mathrm{J}$, which is in agreement with estimates of the physical limit of switching energy of optical digital switches, as reported in the literature [22].

Optical switches may also be based on electroabsorption. In these devices, the absorption changes by the application of an external electric field that deforms the

energy band structure. One example that has attracted considerable interest for practical application is the Quantum-Confined Stark Effect [24]. If an electrical field is applied to a semiconductor quantum well, the shape of the well is changed, perhaps from rectangular to triangular. As result, the position of energy levels also changes, and this affects the optical absorption. As the formation of an electric field requires changes in charge distribution [Equation 4.28], the analysis of electroabsorption optical switch is analogous to a charge-based switch, where the energetics is determined by charging and discharging of a capacitor [Equations 4.31 and 4.32]. It should be noted that the capacitance of an optical switch is considerably larger then the capacitance of electron switch, because of larger capacitance area of the optical switch ($\sim\lambda^2$). By using the estimated minimum size of an electron switch $a_{e\min} \sim 1$ nm (as estimated in Section 4.4.2), and taking into account Equation 4.40b, we obtain an estimate for the switching energy of an electroabsorption device:

$$E \sim 3k_B T \frac{\lambda^2}{a_{e\,\min}^2} \approx 1.2 \cdot 10^{-20}\,\mathrm{J} \cdot \frac{(400\,\mathrm{nm})}{(1\,\mathrm{nm})} \approx 10^{-15}\,\mathrm{J} \tag{4.65}$$

The result from Equation 4.65 is in an agreement with estimates of physical limit of electro-absorption optical switch, as detailed in the literature [22, 25].

This energy barrier – and therefore the switching energy for an optical binary switch – is relatively high, with estimates for the theoretical limit for optical device switching energy varying between 10^{-14} and 10^{-15} J, for different types of optical switch [22]. It should be noted that the switching speed is the speed of re-arrangement for atoms or for charge in the material, and is not related to the speed of light.

4.6 Conclusions

Based on the idea that information is represented by the state of a physical system – for example, the location of a particle – we have shown that energy barriers play a fundamental role in evaluating the operating limits of information-processing systems. In order for the barrier to be useful in information-processing applications, it must prevent changes in the state of the processing element with high probability, and it also must support rapid changes of state when an external CHANGE command is given. If one examines the limit of tolerable operation – that is, the point at which the state of the information-processing element loses its ability to sustain a given state – it is possible to advance estimates of the limits of performance for various types of information-processing element. In these limit analyses, the Heisenberg uncertainty relationship can serve as a basis for estimating performance using algebraic manipulations only.

It was shown that charge-based devices in the limit could offer extraordinary performance and scaling into the range of a few nanometers, albeit at the cost of enormous and unsustainable power densities. Nonetheless, it appears that there is considerable room for technological advances in charge-based technologies. One could consider using electron spin as a basis for computation, as the binary system

state can be defined in terms of spin orientation. However, an energy barrier analysis, based on the equilibrium room-temperature operation of a digital spin-flipping switch, has revealed that extraordinarily large external magnetic fields are required to sustain the system state, and hence that a high energy consumption would result. (Although it has been proposed that if devices can be operated out-of-equilibrium with the thermal environment, then perhaps computational state variables can be chosen to improve on the switching energy characteristic of spin-based devices [26].) The need for very large magnetic fields can be eased by utilizing multiple electron spins to represent the state of the processing element. Unfortunately, the number of electrons that must be utilized is such that the size of the processing elements would be about an order of magnitude larger than that of a charge-based device. Finally, we briefly examined the different physical realizations for optical binary elements, and found that the inability to localize a photon, although an advantage for communication systems, works against the implementation of binary optical switches. As a general rule, optical binary switches are significantly larger than charge-based switches.

Although it appears that it will be difficult to supplant charge as a mainstream information-processing state variable, there may be important application areas where the use of spin or optics could be used to advantage. While the present chapter has focused on the properties of the processing elements themselves, information-processing systems are clearly comprised of interconnected systems of these elements, and it is the system consideration that must remain paramount in any application. Nonetheless charge-based systems, by using the movement of charge to effect element-to-element communication, avoid changing any state variable to communicate, and this is a decided advantage.

References

1 Ayres, R.U. (1994) *Information, Entropy and Progress*, AIP Press, New York.

2 Brillouin, L. (1962) *Science and Information Theory*, Academic Press, New York.

3 Yaglom, A.M. and Yaglom, I.M. (1983) *Probability and Information*, D. Reidel, Boston.

4 Gomer, R. (1961) *Field Emission and Field Ionization*, Harvard University Press.

5 French, A.P. and Taylor, E.F. (1978) *An Introduction to Quantum Physics*, W.W. Norton & Co, Inc.

6 Margolus, N. and Levitin, L.B. (1998) The maximum speed of dynamical evolution. *Physica D*, **120**, 1881.

7 Brody, D.C. (2003) Elementary derivation for passage times. *Journal of Physics A-Mathematical and General*, **36**, 5587.

8 Sze, S.M. (1981) *Physics of Semiconductor Devices*, John Wiley & Sons.

9 Cavin, R.K., Zhirnov, V.V., Hutchby, J.A. and Bourianoff, G.I. (2005) Energy barriers, demons, and minimum energy operation of electronic devices. *Fluctuation and Noise Letters*, **5**, C29.

10 Taur, Y. and Ning, T.H. (1998) *Fundamentals of Modern VLSI Devices*, Cambridge University Press.

11 Singh, J. (1997) *Quantum Mechanics – Fundamentals and Applications to Technology*, John Wiley & Sons.

12 Pearton, S.J., Norton, D.P., Frazier, R., Han, S.Y., Abernathy, C.R. and Zavada, J.M. (2005) Spintronics device concepts. *IEE Proceedings-Circuits Devices and Systems*, **152**, 312.

13 Jansen, R. (2003) The spin-valve transistor: a review and outlook. *Journal of Physics D-Applied Physics*, **36**, R289.

14 Daughton, J.M. (1997) Magnetic tunneling applied to memory. *Journal of Applied Physics*, **81**, 3758.

15 Bandyopadhay, S., Das, B. and Miller, A.E. (1994) Supercomputing with spin-polarized single electrons in a quantum coupled architecture. *Nanotechnology*, **5**, 113.

16 Cowburn, R.P. and Welland, M.E. (2000) Room temperature magnetic quantum cellular automata *Science*, **287**, 1466.

17 Motokawa, M. (2004) Physics in high magnetic fields. *Reports on Progress in Physics*, **67**, 1995.

18 (a) Marshall, W.S., Swenson, C.A., Gavrilin, A. and Schneider-Muntau, H.J. (2004) Development of "Fast Cool" pulse magnet coil technology at NHMFL. *Physica B*, **346**, 594. (b) National High Magnetic Field Laboratory website at: http://www.magnet.fsu.edu/magtech/core.

19 Lietzke, A.F., Bartlett, S.E., Bish, P., Caspi, S., Dietrich, D., Ferracin, P., Gourlay, S.A., Hafalia, A.R., Hannaford, C.R., Higley, H., Lau, W., Liggins, N., Mattafirri, S., Nyman, M., Sabbi, G., Scanlan, R. and Swanson, J. (2005) Test results of HD1b, and upgraded 16 Tesla Nb3Sn Dipole Magnet. *IEEE Transactions on Applied Superconductivity*, **15**, 1123.

20 Corson, D.R. and Lorrain, P. (1962) *Introduction to Electromagnetic Fields and Waves*, W.H. Freeman and Co., San Francisco and London.

21 Jackson, J.D. (1998) *Classical Electrodynamics*, 3rd edn., John Wiley & Sons, New York.

22 Wherrett, B.S. (1996) *Synthetic Metals*, **76**, 3.

23 Higgins, T.V. (1995) *Laser Focus World*, **31**, 72.

24 Miller, D.A.B., Chelma, D.S., Damen, T.C., Gossard, A.C., Wiegmann, W., Wood, T.H. and Burrus, C.A. (1984) *Physical Review Letters*, **53**, 2173.

25 Miller, D.A.B., Chelma, D.S., Damen, T.C., Gossard, A.C., Wiegmann, W., Wood, T.H. and Burrus, C.A. (1984) *Applied Physics Letters*, **45**, 13.

26 Nikonov, D.E., Bourianoff, G.I. and Gargini, P.A. (2006) Power dissipation in spintronic devices out of thermodynamic equilibrium. *The Journal of Superconductivity and Novel Magnetism*, **19**, 497.

II
Nanofabrication Methods

Nanotechnology. Volume 3: Information Technology I. Edited by Rainer Waser

ISBN: 978-3-527-31738-7

5
Charged-Particle Lithography

Lothar Berger, Johannes Kretz, Dirk Beyer, and Anatol Schwersenz

5.1
Survey

The extensive functional range of modern microelectronics is being driven by the ability to pack large numbers of transistors into a small piece of silicon as an integrated circuit. Today, the method used to pattern almost all integrated circuits is *photolithography* (also referred to as *optical lithography*), where circuit patterns from master images, the transmission photomasks, are transferred to silicon wafers by projection optics. In more detail, the wafer is coated with a photoresist, which is exposed with the desired circuit pattern (see Figure 5.1). The resulting resist pattern is transferred to the wafer by subsequent process steps. A detailed introduction to photolithography can be found in Ref. [1], while a comprehensive study is presented in Ref. [2].

The pivotal device of the integrated circuits of today is the metal oxide semiconductor field effect transistor (MOSFET), for which a sample photolithographic process, producing the electrical connections, is shown in Figure 5.2. Here, the bulk transistor has already been fabricated on the wafer, and consists of the doped areas of source, drain, and gate. The wafer is then coated with a positive photoresist, and exposed to form the contact areas for the transistor (Figure 5.2a). The exposed areas of the photoresist are removed by a developer chemical (b), after which the insulating layer, now open at source, drain, and gate, is etched away (c). After metallization, the remaining unexposed photoresist is removed by a stripping chemical (d).

In 2006, the gate length of the MOSFET in the most advanced integrated circuits is typically 65 nm. The gate length, being the smallest feature required, is also known as the *critical dimension* (CD) of the pattern. A CD of 65 nm is close to the resolution limit of the current photolithography. The smallest feature is determined by:

$$\mathrm{CD} = k_1 \frac{\lambda}{\mathrm{NA}} \tag{5.1}$$

where k_1 is a factor determined by the projection optics and process flow, λ is the wavelength of the photons, and NA is the numerical aperture between the objective optical lens and the resist plane.

Nanotechnology. Volume 3: Information Technology I. Edited by Rainer Waser

ISBN: 978-3-527-31738-7

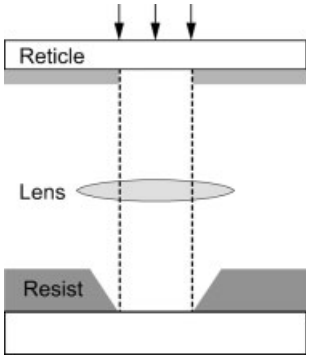


Figure 5.1 The principle of photolithography.

While a number of techniques have been developed to increase the resolution by reducing k_1 (referred to as *resolution enhancement techniques*; RET), the smallest feature that can be prepared by photolithography is ultimately dependent on the wavelength λ of the photons. Therefore, there is a history of reducing the wavelength, which is 193 nm for the current state-of-the-art *deep ultraviolet lithography* (DUVL). With $k_1 \approx 0.3$ for current projection optics and process flow, and NA ≈ 0.95 for a technically feasible projection in air, CD ≈ 61 nm represents the smallest feature.

Attempts to reduce the wavelength further have been investigated extensively, but encountered problems. For example, at $\lambda = 157$ nm no fully suitable material has been found to fabricate the transmission photomask and the lenses of the projection optics. Therefore, the current 193-nm photolithography is now scheduled to be extended into 193 nm immersion photolithography, which increases the NA by

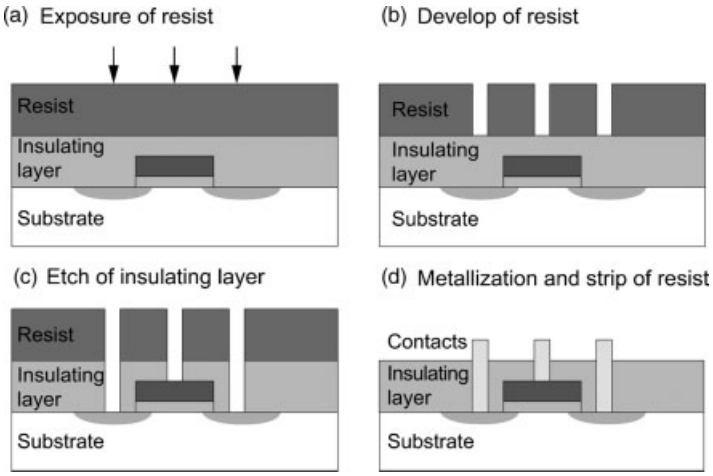


Figure 5.2 The photolithographic process for making electrical connections to a transistor.

conducting the exposure not in air, but in a liquid. With NA ≈ 1.4 for a technically feasible projection in water, CD ≈ 42 nm can be achieved. Much progress has been made with immersion photolithography, and the technique is currently at the stage of pilot production.

Another means of reducing the wavelength is to forego the use of transmission photomasks and projection optics with lenses, and to utilize reflective masks and mirror projection optics. Lithography involving reflection is no longer considered as classical photolithography. The wavelength of the photons where mirrors can be applied most effectively is 13.5 nm, and lithography of 13.5 nm involving masks with multilayer Bragg reflectors is referred to as *extended ultraviolet lithography* (EUVL), for which considerable research effort is currently being expended. At this point, it should be mentioned that EUVL differs greatly from DUVL in that it requires the redevelopment of almost all the exposure equipment and lithography processes currently in use. A comprehensive study of the process is presented in Ref. [3].

An alternative to any lithography involving photons is *charged-particle lithography*, where charged particles (electrons, ions) are used for patterning. While certain charged particle lithography techniques are already used for special applications, such as fabricating masks for photolithography, or prototyping, promising new charged-particle lithography techniques for preparing integrated circuits are currently under development, and may in time complement or even replace photolithography.

In order to illustrate the relationship between charged-particle lithography and photolithography, it is of help to examine the International Technology Roadmap for Semiconductors (ITRS), which demonstrates the status of lithographic techniques currently in use and under development within the microelectronics industry. The ITRS for the year 2006 is shown in Figure 5.3 [4]. According to the current ITRS, in 2007 the most likely successors of DUVL include EUVL, a multiple-electron-beam lithography technique called *maskless lithography* (ML2), and *nanoimprint lithography* (NIL). Charged-particle lithography techniques known as electron projection lithography (EPL) and ion projection lithography (IPL), both of which use a transmission mask and projection optics with electromagnetic lenses to direct electrons and ions, were removed from the ITRS in 2004. Proximity electron lithography (PEL), which uses a 1 : 1 transmission mask, was removed in 2005, although its re-emergence cannot be ruled out completely.

In this chapter we discuss the physical concepts and the principal advantages and limitations of charged-particle lithography techniques. A brief insight is also provided into the charged-particle lithography techniques currently in use and under development, while a strong focus is placed on ML2 techniques. The chapter comprises four sections: Section 5.1 includes a survey of the field, while Section 5.2 incorporates discussions of electron beam lithography and electron resists, and their major applications:

- the fabrication of transmission masks for DUVL and reflective masks for EUVL

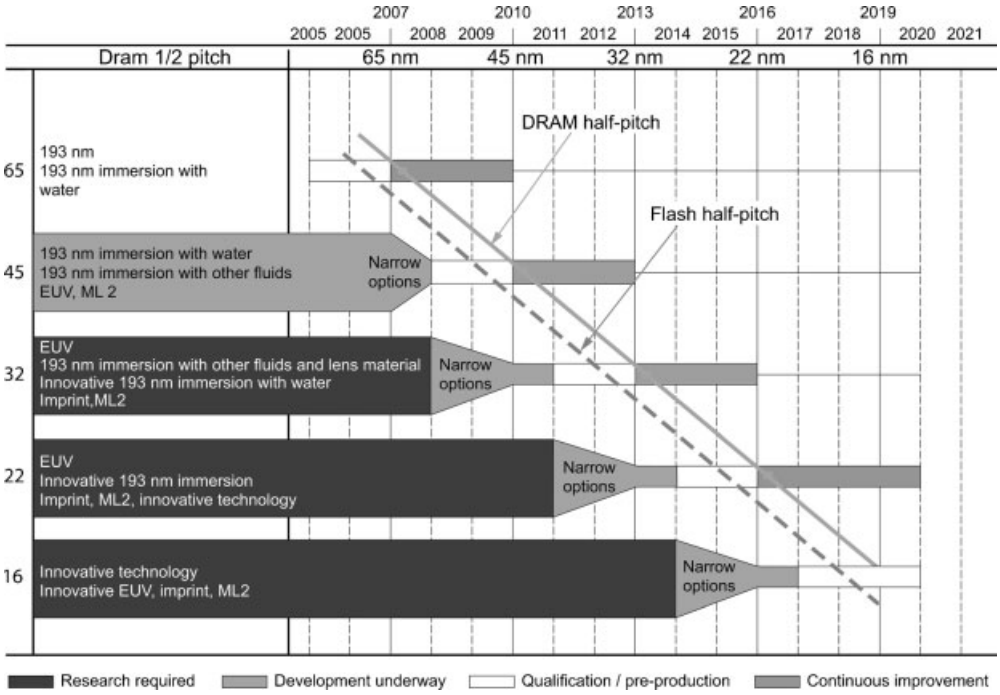


Figure 5.3 The lithography roadmap of 2006 [4].

- the direct-writing of patterns onto wafers with single beams for prototyping, low-volume production, and mix-and-match with photolithography
- the direct-writing of patterns on wafers with multiple beams for volume production: ML2
- the fabrication of imprint templates for NIL.

Section 5.2 concludes with a discussion of the special requirements for mix-and-match, namely the integration of electron beam lithography (EBL) with photolithography. Section 5.3 presents details of ion beam lithography (IBL), for which the major applications include:

- the direct-structuring of patterns on wafers without resist processing, for prototyping, low-volume production, and special applications
- the fabrication of imprint templates for NIL, with direct-structuring of patterns without resist processing.

A graphical overview of the charged-particle lithography techniques discussed in this chapter is provided in Figure 5.4. Finally, Section 5.4 provides a conclusion and outlook on charged-particle lithography techniques.

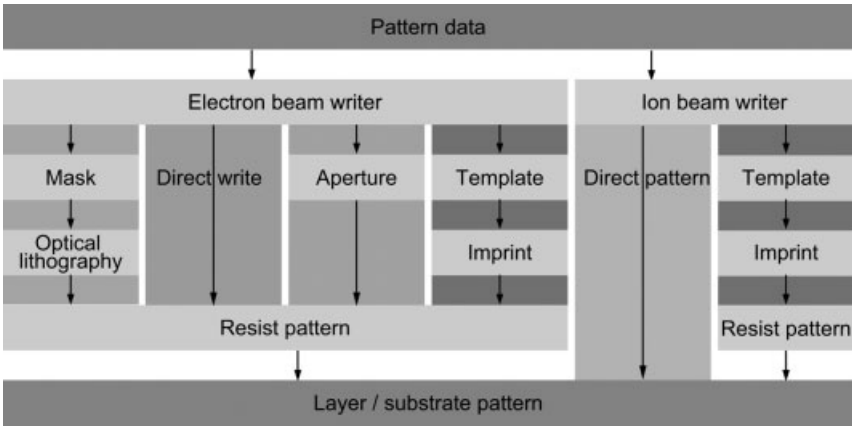


Figure 5.4 Charged-particle lithography techniques: an overview.

5.2 Electron Beam Lithography

5.2.1 Introduction

Electron beam lithography involves the use of electrons to induce a chemical reaction in an electron resist for pattern formation (the properties of electron resists are discussed in detail in Section 5.2.2). Because of the extremely short wavelength of accelerated electrons, EBL is capable of very high resolution, as $\lambda = h/(mv)$, $E_{kin} = (mv^2)/2$, and $E = eU$ gives:

$$\lambda = \frac{h}{\sqrt{2meU}} \tag{5.2}$$

Therefore, the typical acceleration voltage of $U = 100\,\text{kV}$ results in $\lambda = 0.004\,\text{nm}$, which is well below the atomic scale. Even with simple EBL systems, 10-nm patterns have been demonstrated [5], which is well beyond any other lithographic technique. Further, EBL is unaffected by the major issues of optical lithography – diffraction and reflection – and much less affected by the depth-of-focus (DOF) limit [1, 2].

The first EBL tools where based on the scanning electron microscope, and first developed during the 1960s [6]. A schematic representation of a simple EBL system is shown in Figure 5.5. The column consists of an electron source, and the electron optics. The substrate is mounted on a precision stage below the column, where its position is controlled by a laser interferometer. As the electron source and electron optics require a high vacuum, a load–unload system with an air-lock is also fitted.

5.2.1.1 Electron Sources

While early EBL systems based on scanning electron microscopy (SEM) used field emission sources, where the electrons are extracted out of a sharp-tip cathode, modern

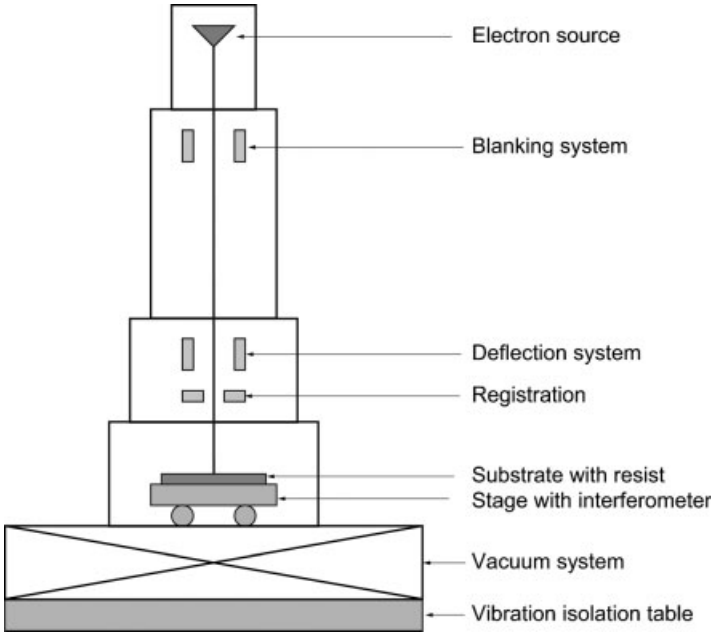


Figure 5.5 Schematic representation of a simple electron beam lithography system.

systems employ thermal sources. Such a thermal source consists of an emission region, where the cathode is heated, and an extraction region, where an electric field extracts the electrons and accelerates them to form a beam. The maximum current density j which can be obtained from this type of source for an acceleration voltage V and an acceleration distance d is limited by space-charge effects [6]:

$$j = \frac{1}{9\pi d^2}\sqrt{\frac{2q}{m}}\,V^{\frac{3}{2}} \tag{5.3}$$

The properties of common field emission and thermal sources are listed in Table 5.1 [6]. Because of its lower operating temperature and energy spread, LaB_6 is the source of choice for most current tools.

5.2.1.2 Electron Optics

Electron optics is based on the fact that electrons can be deflected by electromagnetic fields. The electric field between two grid electrodes, which causes bending of the trajectory of an electron, is shown in Figure 5.6.

Table 5.1 A comparison of the properties of common field emission and thermal sources [6].

Parameter	Field emission: tungsten	Thermal: tungsten	Thermal: LaB_6
Operating temperature T [K]	300	2700	1700
Energy spread [eV]	0.3	3	1.5

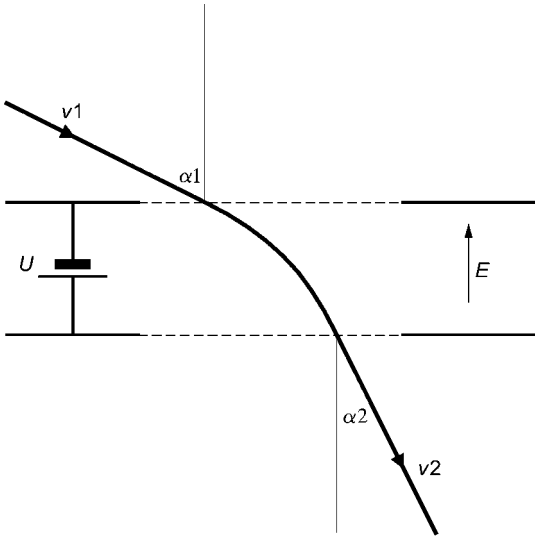


Figure 5.6 Electron–optical refraction.

For this configuration, the conservation of energy for an electron yields:

$$E_{kin,1} + eU = E_{kin,2} \tag{5.4}$$

where $E_{kin,1} = (mv_1^2)/2 = eU_0$ and $E_{kin,2} = (mv_2^2)/2$, and Equation 5.4 can be expressed as an electron–optical refraction law:

$$\frac{\sin\alpha_1}{\sin\alpha_2} = \frac{v_2}{v_1} = \sqrt{1 + \frac{U}{U_0}} \tag{5.5}$$

Electron Lenses A simple electrostatic lens (Einzel lens) consists of three rings, where the outlying rings have the same electrostatic potential (Figure 5.7).

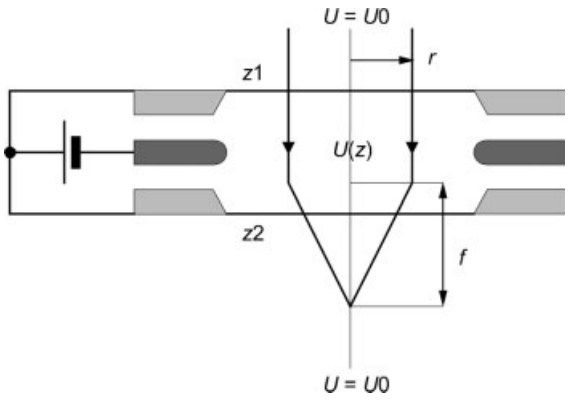


Figure 5.7 Electrostatic lens.

For this configuration, if the charge effects are negligible, then a simple equation for the paraxial trajectories of electrons can be obtained [6]:

$$\frac{d^2r}{dz^2} + \frac{1}{2U}\frac{dU}{dz}\frac{dr}{dz} + \frac{1}{4U}\frac{d^2U}{dz^2}r = 0 \tag{5.6}$$

It should be noted that as Equation 5.6 is invariant towards the scaling of the voltage U, voltage instabilities in general do not cause a jitter of trajectories through electrostatic lenses. With the geometrical relation of the trajectory r and the focal length f, $dr(z_1)/dz = -r_1/f$, and $r \approx r_1$, Equation 5.6 yields [6]:

$$\frac{1}{f} \approx \frac{1}{8\sqrt{U_0}} \int_{z_1}^{z_2} \left(\frac{dU}{dz}\right)^2 U^{-\frac{3}{2}}\, dz \tag{5.7}$$

Magnetic lenses utilize the force on an electron in a magnetic field:

$$\vec{F} = e\,\vec{v} \times \vec{B} \tag{5.8}$$

A simple magnetic lens is a solenoid ring (Figure 5.8).

For this configuration, if the radial field components are negligible, then a simple equation for the paraxial trajectories of electrons can be obtained [6]:

$$\frac{d^2r}{dz^2} = -\frac{e}{m}\frac{B_z^2}{8U_0}r \tag{5.9}$$

With the geometrical relation of the trajectory r and the focal length f, $dr(z_2)/dz = -r_1/f$, and $r \approx r_1$, Equation 5.9 yields [6]:

$$\frac{1}{f} \approx \frac{e}{8mU_0} \int_{z_1}^{z_2} B_z^2\, dz \tag{5.10}$$

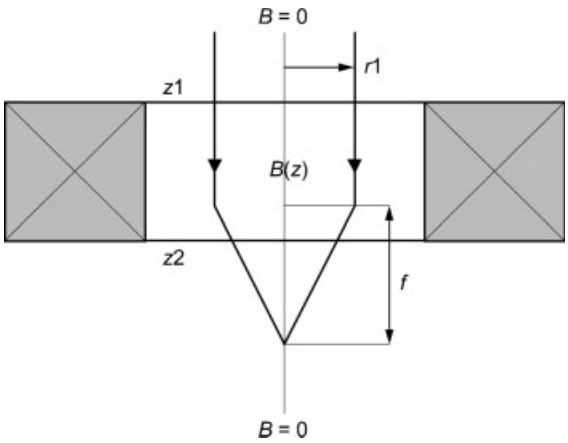


Figure 5.8 Magnetic lens.

Electron Optical Columns For the design of electron optical columns, the previous simple considerations are not adequate. It is required to derive the trajectories of electrons in a general form, which is valid also for non-rotational-symmetric lenses. A straightforward calculation can be based on the general equation of motion of electrons in an electron optical column:

$$\frac{d}{dt}(m\vec{v}) = e\left[\vec{E}(\vec{r},t) + \vec{v} \times \vec{B}(\vec{r},t)\right] \tag{5.11}$$

It is convenient to substitute with the arc element of the trajectory, $ds = |d\vec{r}| = vdt$:

$$m\frac{d\vec{v}}{ds} = e\left[\frac{\vec{E}}{v} + \frac{d\vec{v}}{ds} \times \vec{B}\right] \tag{5.12}$$

The trajectory equation can be derived in Cartesian coordinates, and with the introduction of the abbreviations $\eta = \sqrt{\frac{e}{2m_0}}$, $\varepsilon = \frac{e}{2m_0c^2}$, $\Phi_0 = \frac{E_0}{e}$, $\hat{\Phi} = (\Phi_0 + \Phi)[1 + \varepsilon(\Phi_0 + \Phi)]$, $\rho = |\vec{r}\,'| = \sqrt{1 + (x')^2 + (y')^2}$, where E_0 is the initial kinetic energy of an electron at the source, the trajectory equation results as follows [7]:

$$\begin{aligned} x'' &= \frac{\rho^2}{2\hat{\Phi}}\left(\frac{\partial\hat{\Phi}}{\partial x} - x'\frac{\partial\hat{\Phi}}{\partial z}\right) + \frac{\eta\rho^2}{\sqrt{\hat{\Phi}}}\left(\rho B_y - y' B_t\right) \\ y'' &= \frac{\rho^2}{2\hat{\Phi}}\left(\frac{\partial\hat{\Phi}}{\partial y} - y'\frac{\partial\hat{\Phi}}{\partial z}\right) + \frac{\eta\rho^2}{\sqrt{\hat{\Phi}}}\left(-\rho B_x + x' B_t\right) \end{aligned} \tag{5.13}$$

It should be noted that this trajectory equation is valid only if all the trajectories are continuous – that is, if $x'(z)$, $y'(z)$ are finite. This is not the case with electron mirrors, although these are not used in current electron optical columns.

Aside of the trajectory representation of Equation 5.13, more elaborate mathematical methods have been applied to the analytic investigation of electron optical columns concepts and designs, focusing on the prediction of projection imperfections, called *aberrations*. These methods include the classical mechanics approach of Lagrange or Hamilton formalism [7], with the latter leading towards the Hamilton–Jacobi theory of electron optics, which is capable of treating whole sets of trajectories, and therefore is a standard tool for the design of electron optical columns with minimal aberrations [8].

Recently, with the support of computer algebra tools, a Lie algebraic electron optical aberration theory has been derived, which makes accessible high-order canonical aberration formulas, and therefore may open up new possibilities in the design of high-performance electron beam projection systems [9].

Electron Optical Aberrations Electron optical columns suffer from projection deviations termed aberrations, which are caused either by non-ideal electron optical

elements, or by the physical limits of electron optics. An ideal electron optical lens should project an electron beam crossing the entrance plane at the point (x_0, y_0) onto the exit plane at the point $(x_1, y_1) = m(x_0, y_0)$, where m is a scalar called the *magnification*. Unfortunately, however, a real lens suffers from imperfections. For example, a real lens has the same focal length only for paraxial electron beams, while off-axis beams are slightly deflected. This is called *spherical aberration*, and is expressed by a coefficient C_s, relative to the beam position in the aperture plane (x_a, y_a):

$$(x_0, y_0) \rightarrow m\left(x_0 + C_s x_a\left(x_a^2 + y_a^2\right), y_0 + C_s y_a\left(x_a^2 + y_a^2\right)\right) \tag{5.14}$$

With $C_s > 0$, an electron beam is blurred into a finite disk. Electron optical aberration theory shows that C_s cannot be made to vanish completely for rotational symmetric lenses [7], which in turn led to the introduction of non-rotational symmetric elements such as quadrupoles or octopoles. An overview of lens aberrations occurring in an electron optical column is provided in Ref. [7]. Additionally, the non-ideality of the electron source must be considered: a real electron source has a small spread in electron energy and, as the focal length of electron lenses depends on the electron energy, a non-monochromatic electron beam is blurred into a finite disk. This is referred to as *chromatic aberration*. In addition to these lens aberrations, other electron optical elements, such as beam deflectors, also introduce aberrations.

For a high-resolution EBL system, optimal projection is crucial, and therefore the aberrations discussed above must either be minimized by the design of the electron optical column, or compensated by additional electron optical elements. For example, since it is not possible to prepare a column with perfect rotational symmetry, an electron beam always suffers from astigmatism and misalignment. In order to compensate for astigmatism, a non-rotational symmetric element is required, a *stigmator*. This may be an electrostatic quadrupole, which is a circular arrangement of four electrodes. However, stigmators are usually designed as electrostatic octopoles which, in addition to quadrupole fields, can also generate dipole fields at any angle to correct beam misalignment [10].

While major efforts have been made to apply aberration theory to the design of an electron column with minimal aberrations, only powerful numerical optimization tools can be utilized for any systematic approach [11]. Such a tool generally consists of the following packages:

- an electromagnetic field computation package
- an electron beam ray tracing package
- an electron exposure spot plotting package
- an optimizer for minimizing aberrations.

An aberrations optimizer typically implements the damped least squares (DLS) method for iteratively minimizing overall aberrations. The individual aberrations $f_i(x_1, \ldots, x_n)$, depending on the column parameters x_j, (lens positions, sizes,

strengths), are weighted and summed into a deviation function:

$$\Psi = \sum_i (w_i f_i)^2 \tag{5.15}$$

which is then minimized. With this method, existing electron column designs can be improved, and new designs optimized automatically.

In addition to the design of electron optical columns with minimal aberrations, it has been proposed to implement adaptive aberration correction by introducing novel electron optical elements: for example, an electrostatic dodecapole, with time-dependent voltage control for the poles [12]. The learning process could be based on exposure pattern images.

5.2.1.3 Gaussian Beam Lithography

Patterning in EBL can be accomplished by focusing the electrons into beams with very narrow diameters. Because the electrons are created by a thermal source, they have a spread in energy. The trajectories of the electrons therefore vary slightly, resulting in electron beams with near-Gaussian intensity distribution after traversing the electron optics [13].

The basic principle of Gaussian electron beam exposure is *raster scanning*. Similar to a television picture tube, the electron beam is moved in two dimensions across the scanning area on the electron resist, which typically is about 1 mm^2. Within that area, which is termed the *deflection field*, the electron beam can be moved very rapidly by the electron optics. In order to change the position of this area on the substrate, a mechanical movement of the precision stage is required. Patterns which stretch over more than 1 mm^2 must be stitched together from separate deflection fields; in order to avoid discontinuities in the patterns at the boundaries of the deflection fields (these are known as *butting errors*, and potentially are caused by mechanical movement of the stage), the positions of these boundaries are calibrated and corrected in sophisticated manner.

In order to create the pattern shown in Figure 5.9a, the electron beam is moved across the area where the desired pattern is located, and is blanked on spots not intended for exposure. Raster scanning certainly has the major problem that the electron beam must target the whole scanning area, which takes a considerable time. Therefore, a great improvement is to target only the area of the desired pattern; this is termed *vector scanning* (Figure 5.9b). However, another major limitation remains in that, if a pattern consists of large and small features, the diameter of the electron

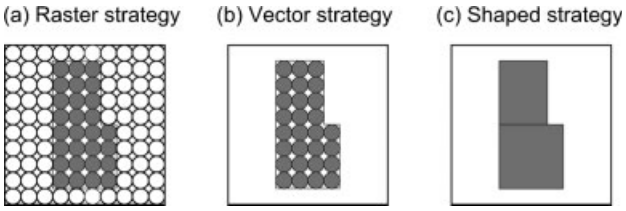


Figure 5.9 Writing a pattern with: (a) raster-, (b) vector-, and (c) vector shaped & beam strategies.

beam must be adapted to the smallest feature, thereby greatly increasing the exposure time of the large features.

5.2.1.4 Shaped Beam Lithography

In order to overcome the limitations of both raster and vector scanning, EBL tools have been developed which can shape electron beams. A shaped electron beam is created by special aperture plates and, in contrast to the near-Gaussian intensity distribution of standard EBL tools, shaped electron beam tools can apply rectangular, or even triangular, intensity distributions [14]. At present the fastest technique available is vector scanning using shaped electron beams; when using this technique the pattern shown in Figure 5.9c requires only two exposures.

The principal function of a variable-shaped beam (VSB) column is illustrated in Figure 5.10. The electron source illuminates a first shaping aperture, after which a first condenser lens projects the shaped beam onto a second shaping aperture. The beam position on the second aperture is controlled by an electrostatic deflector. A second condenser lens projects the shaped beam onto the demagnification system, consisting of two lenses, and the final aperture in between. After demagnification, the

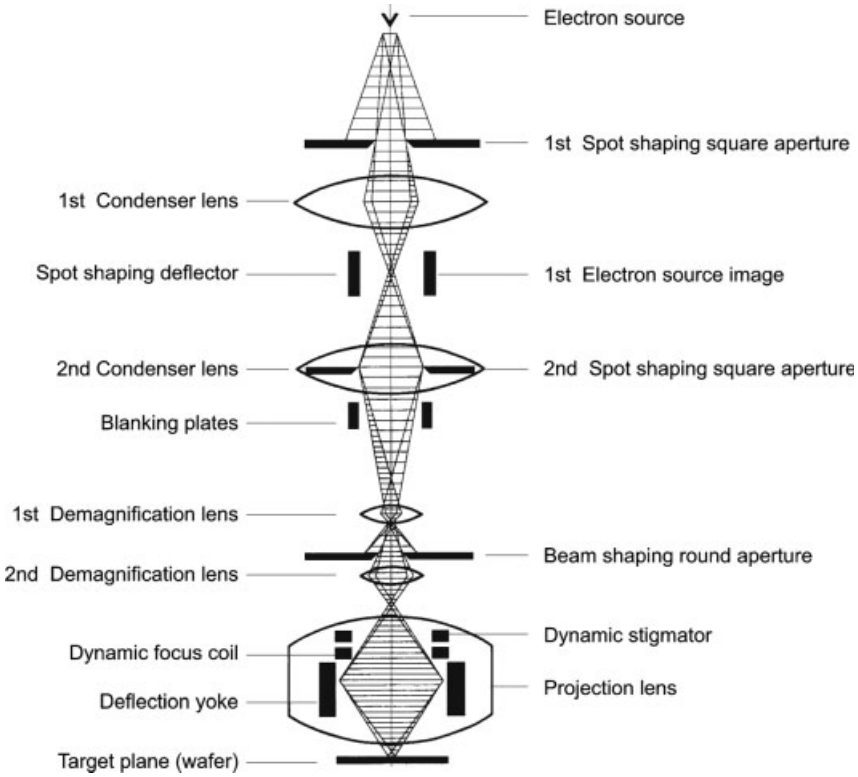


Figure 5.10 Variable beam-shaping column [15]. (© 1979, IEEE.)

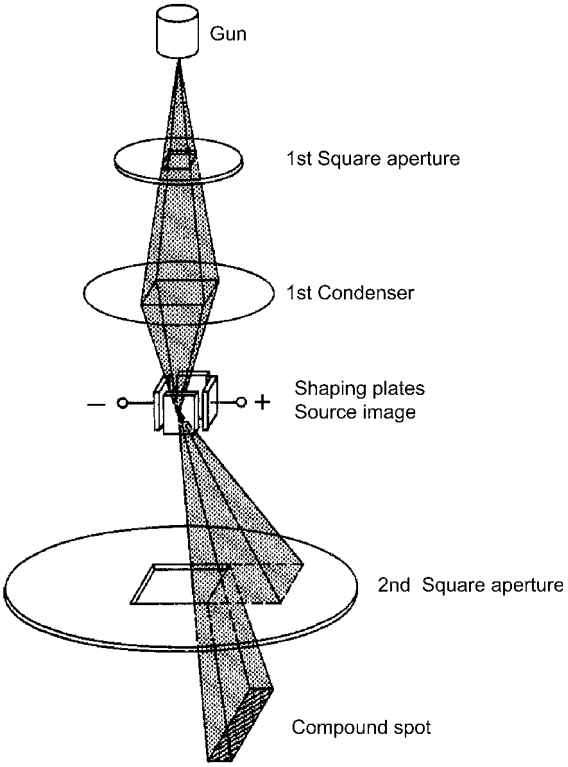


Figure 5.11 Variable beam-shaping method [15]. (© 1979, IEEE.)

shaped beam is projected onto the substrate by the final projection system, which consists of a stigmator and a projection lens with an integrated deflector.

The formation of shaped beams is illustrated for the example of rectangular spots in Figure 5.11. Two square apertures shape the spot; the image of a first square aperture, which appears in the plane of a second square aperture, can be shifted laterally with respect to the second aperture. This results in a rectangular spot which then is demagnified and projected onto the substrate.

Modern shaped electron beam tools can apply both rectangular and triangular spots. For example, the Vistec SB3050 [16] employs a LaB_6 thermal source, and utilizes a vector scan exposure strategy, a continuously moving stage, and the variable-shaped beam principle. The maximum shot area is $1.6 \times 1.6\,\mu m$, and rectangular shapes with 0° and 45° orientation, as well as triangles, can be exposed in a single shot. A detailed view of the two shaping aperture plates is shown in Figure 5.12.

The architecture and motion principle of the stage is decisive for pattern placement accuracy, so as to avoid butting errors. Position control by interferometer with a resolution $<1\,nm$, and the use of a beam tracking system, allow write-on-the-fly exposures with stage speeds of up to $75\,mm\,s^{-1}$. The driving range of the stage is

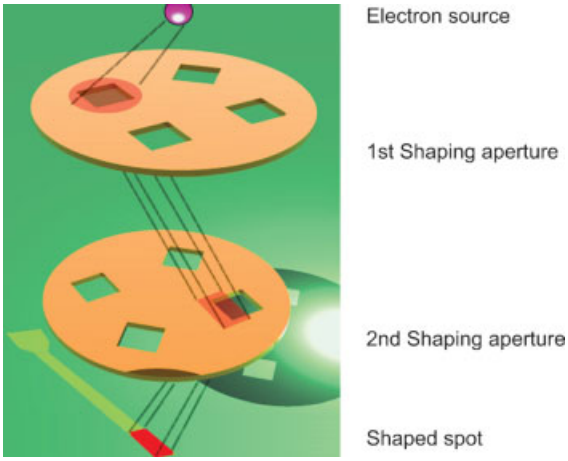


Figure 5.12 Schematic of electron beam shaping by double-aperture for rectangular and triangular spots [16].

310 × 310 mm, thus enabling the exposure of 6 in and 9 in masks, as well as 300 mm wafers (see Figure 5.13).

However, even such a precision stage cannot eliminate butting errors completely, and therefore the vector-shaped beam strategy involves overlapping exposure shapes, resulting in features being exposed at least twice, a process known as *multi-pass writing*.

A production-worthy EBL system is highly automated, with no human intervention required for operation, except for an operator issuing a command for the system to start loading the substrate and writing the pattern. The pattern is encoded in a digital data file, and stored in a computer memory or a mass storage device. Prior to writing, the original design data must be converted to a format which is usable by the writing tool. This data fracturing is accomplished using separate computer hardware, usually

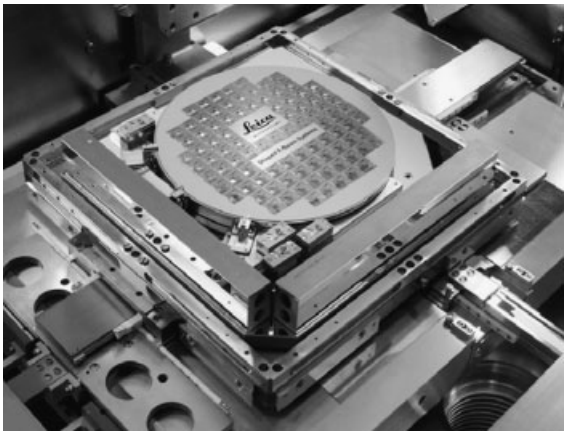


Figure 5.13 Electron beam lithography precision stage [16].

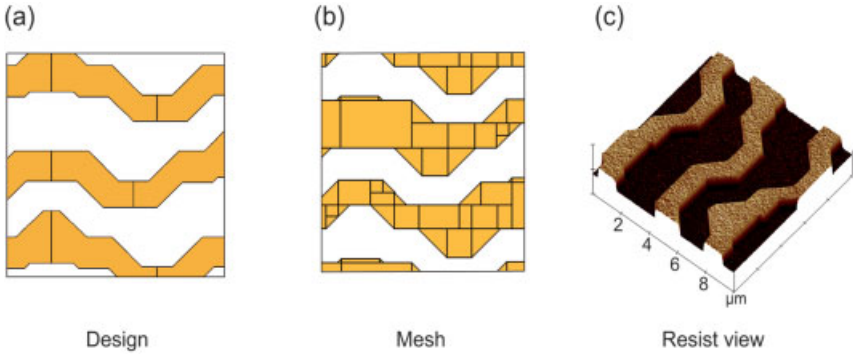


Figure 5.14 Writing a complex pattern with a vector-shaped beam strategy. (a) Design; (b) data fracturing into shapes; (c) the final atomic force microscopy resist picture.

a high-performance cluster, operating a real-time multi-threaded operating system, and data preparation software. An example of the writing of a complex pattern with vector shaped-beam strategy is shown in Figure 5.14.

EBL tools are usually tailored specifically for either photomask or wafer applications (this is also referred to as *direct-write lithography*), and the principal advantages and limitations of these processes are presented in more detail in Sections 5.2.3.1 and 5.2.3.2. Evaluations of the current leading-edge electron beam writing systems are available in Refs. [16–18].

5.2.1.5 **Patterning**

As highlighted at the start of Section 5.1, EBL – unlike optical lithography – is unaffected by major issues of optical lithography, such as diffraction and reflection, and much less affected by the DOF limit. However, it also encounters specific patterning issues, which must be addressed.

The most critical issue results from the scattering of the electrons passing through matter. While this scattering in part is indeed required for transferring energy to the electron resist for exposure, many electrons scatter into different directions from their original trajectories. In Figure 5.15, the shaded areas of the resist are intended for exposure. However, the electrons show two scattering modes, from resist intended for exposure into resist not intended for exposure – this is *forward scattering* (a); and through resist intended for exposure into the substrate and back into resist

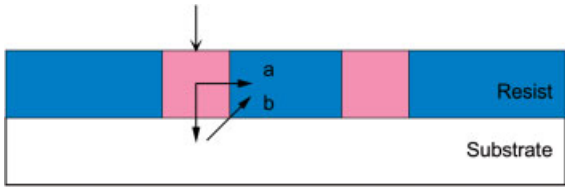


Figure 5.15 Electron-scattering modes. (a) Forward scattering; (b) back-scattering.

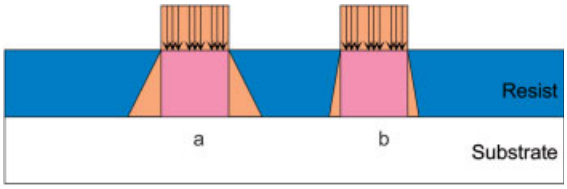


Figure 5.16 Forward scattering at: (a) low electron energy and (b) high electron energy.

not intended for exposure – this is *back-scattering* (b). Both scattering modes lead to unintended exposure, and therefore to degraded resolution and distortion of patterns. This issue is known as the *proximity effect* [19], and should be minimized as much as possible.

Forward scattering into the resist is addressed by increasing the acceleration voltage of the electrons, as shown in Figure 5.16. Although the first electron lithography tools used $U = 10\,\text{kV}$, and current tools employ 50 kV, more recently 100 kV tools have been introduced. This increase in acceleration voltage leads also to an improved projection performance, because the imperfections in the electron optics are less pronounced.

However, back-scattering intensifies with greater electron energy. With the currently required feature size being less than the range of back-scattered electrons, the features are broadened by back-scattering, thus offsetting the resolution improvement by reduced forward scattering.

For optimal resolution and minimal distortion, correction methods are therefore applied to overcome the proximity effect. A simple compensation technique, which accounts for back-scattering only, is to use a second exposure equaling the background exposure of the first one, with a reverse additional energy distribution [20]. However, as a second exposure is undesirable due to the additional writing time, a proximity correction by dose variation is introduced during data processing for converting the circuit designs [21]. Such a correction must be based on suitable models for proximity effect predictions. Here, a useful tool is the Monte-Carlo-based simulation of scattering from electron impact. Simulated trajectories for 100 electrons impacting into one point are shown in Figure 5.17, where both forward scattering and back-scattering appear distinctively.

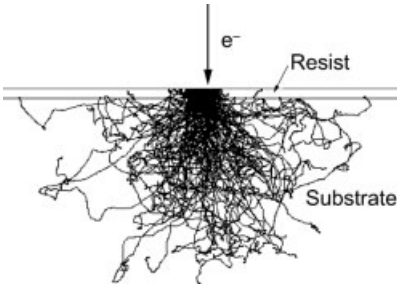


Figure 5.17 Simulated trajectories for 100 electrons impacting into one point.

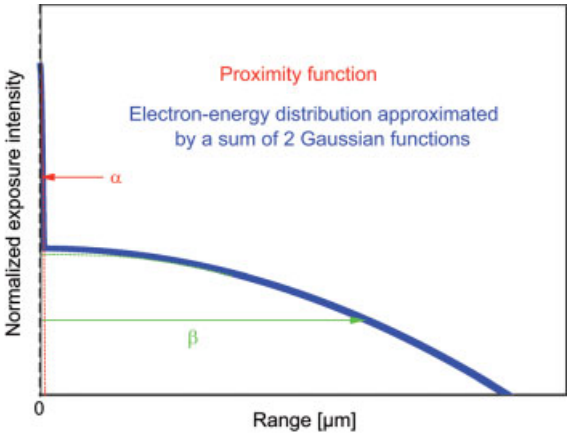


Figure 5.18 Proximity function (point exposure distribution fitted within two Gaussian functions).

In addition to model-based analysis, interpretations of generic pattern distortions of non-corrected patterns, as well as successive back-simulations, are utilized for the reconstruction of proximity effects [22], therefore enabling the best possible correction.

Current compensation techniques rely on either shot-by-shot modulation of the exposure dose, modification of the pattern geometry, or a combination of both methods. The proximity function $f(r)$ is usually described as a sum of two or more normalized two-dimensional Gaussian functions:

$$f(r) = \frac{1}{\pi(1+\eta)}\left[\frac{1}{\alpha^2}e^{-\frac{r^2}{\alpha^2}} + \frac{\eta}{\beta^2}e^{-\frac{r^2}{\beta^2}}\right] \tag{5.16}$$

The term α characterizes the forward scattering, and the term β the back-scattering, of the electrons. The parameter η is the deposited energy ratio of the back-scattering component towards the forward-scattering component, and r is the distance from the point of electron impact. The behavior of this function is shown by the diagram in Figure 5.18.

The concept of proximity correction by shot-to-shot dose modulation is illustrated in Figure 5.19: two features in close distance should be exposed. When applying a uniform exposure dose (Figure 5.19a) the intended resist exposure by forward scattering is compromised by back-scattering, leading to a distorted overall exposure, and to broadened features. When applying a suitably modulated exposure dose (Figure 5.19b), the resulting resist exposure by forward scattering complements the back-scattering, leading to the intended resist exposure.

5.2.2 Resists

Electron beam patterning requires specially designed electron-sensitive resists. As with photoresists, electron resists are available either as positive resists (which are insoluble in the developer chemical and become soluble when exposed), or as

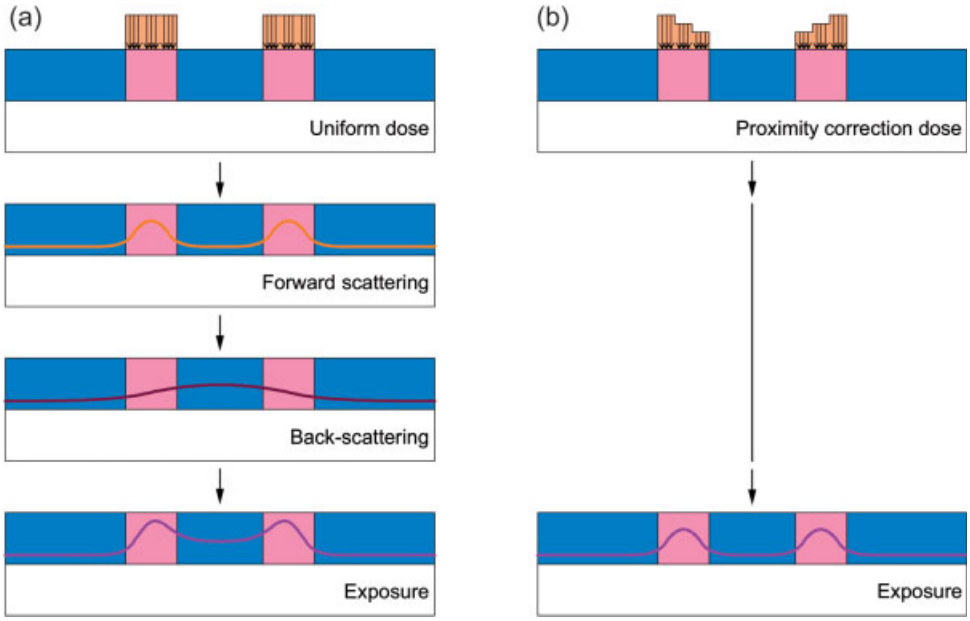


Figure 5.19 Proximity correction by shot-to-shot dose modulation.

negative resists (which are soluble from the start, and become insoluble when exposed). While the chemistry of electron resists is usually considerably different from that of photoresists, a parameter termed *contrast* can be defined to characterize the resolution of the resists.

In order to determine the contrast of an electron resist, the developing rate depending on the exposure dose is plotted; this is also called the characteristics of the electron resist. For a low-exposure dose, the resist still behaves like an unexposed resist, whereas for a high dose it is fully activated. The idealized characteristics of positive and negative electron resists are shown in Figure 5.20.

Linearization yields two parameters that define the characteristics: the resist sensitivity D_0, where the resist activation starts, and the resist activation D_v, after which the resist is fully activated. The contrast is then defined as:

$$\gamma = \frac{1}{\log \frac{D_V}{D_0}} \tag{5.17}$$

A high steepness of the transition in the characteristics therefore results in a high contrast.

Early electron resists employed just a single component for obtaining the latent image. In a positive resist, the image was created by electron-induced chain scission, with high-molecular-weight polymers with long chains being fragmented into smaller chains. The contrast of the resist was engineered by maximizing the difference in molecular weight before and after exposure. In a negative resist, the fragmentation into smaller chains generates radicals, which induce a crosslinking.

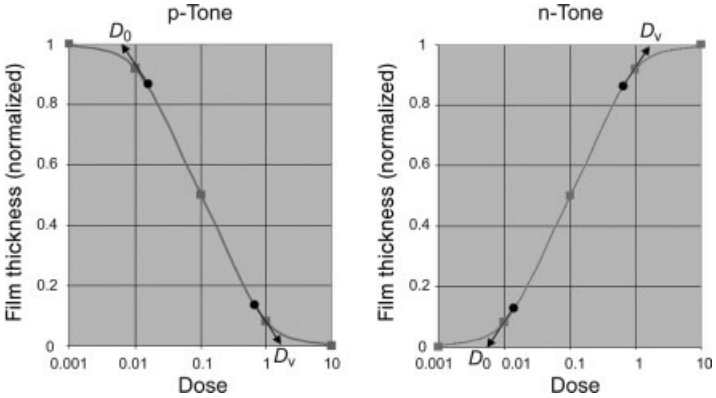


Figure 5.20 Idealized characteristics of positive and negative electron resists.

One of the first resists to be developed for EBL was polymethyl methacrylate (PMMA) [23]. Electron beam exposure breaks the polymer into fragments that are dissolved by a solvent-based developer (see Figure 5.21). Because of its very high resolution capability of <10 nm, PMMA is still used for certain R&D applications and electron beam writer resolution tests. However, it is not suitable for commercial lithography, mainly because of its poor resistance to dry etching.

Another group of early electron resists is based on a copolymer of chloromethacrylate and methylstyrene. One commercial resist for direct-write applications is ZEP-520 [24], which has considerable advantages compared to PMMA. While providing a comparable resolution of <10 nm, the sensitivity towards electrons is 10-fold higher, and the resistance to dry etching is 2.5-fold higher. Although these resists are still used for R&D applications, they have fallen out of favor for commercial lithography, because they require solvent-based developers that, because of their rapid evaporation rate in air, introduce temperature gradients on wafers, and therefore uniformity problems. A 45 nm 1 : 1 dense-lines pattern of a ZEP-520 resist is shown in Figure 5.22 [25].

For some time, UV photoresists have been used as electron resists. This group of resists is based on diazonaphthoquinone (DNQ), and has the advantage of using an aqueous developer. Their resistance to dry etching is also threefold higher. However, with resolutions <250 nm, and despite still being used extensively in optical lithography for micro-electromechanical systems (MEMS) fabrication, their time in

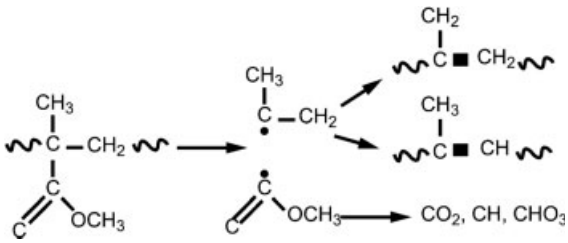


Figure 5.21 The reaction chemistry of polymethyl methacrylate (PMMA).

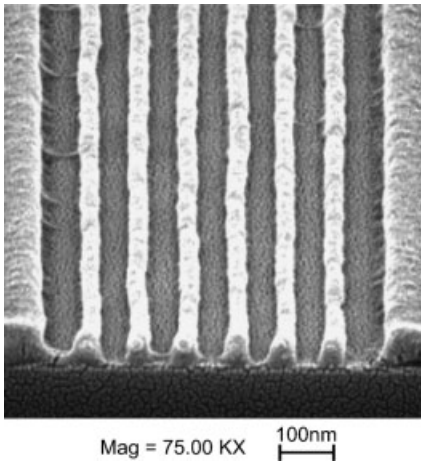


Figure 5.22 ZEP-520 : 45 nm 1 : 1 dense-lines test pattern [25].

EBL has passed. A detailed presentation of the chemistry of DNQ photoresists can be found in Ref. [26].

The impetus for the development of the electron resists used today was derived from the need for a new type of photoresist required for the introduction of DUV optical lithography. Because of the limitations of DNQ-based resists in resolution and sensitivity, *chemically amplified* (CA) resists have been introduced [27]. These are based on a multi-component scheme, where a sensitizer chemical causes dissolution modification within the exposed areas of the polymer matrix. The latent image is obtained from energy transfer to the sensitizer chemical molecules, causing a degradation into their ionic pairs or neutral species, which can catalyze the reaction events needed for solubility distinction. Commonly, photo acid generators (PAG) or photo base generators (PBG) are utilized as sensitizers in CA resists.

In a positive CA resist the PAG, upon exposure, releases an acid. During heating of the substrate after exposure (the post-exposure bake; PEB), this acid reacts with the resin, which in turn becomes soluble towards an aqueous developer. In addition, further acid is produced. With this multiplication reaction an exposed PAG molecule can trigger up to 1000 reactions. It is also acknowledged that a CA resist shows a high quantum yield compared to a DNQ-based resist.

Following the establishment of CA photoresists, specialized electron CA resists have now been developed. The reaction mechanism of a positive electron CA resist is shown in Figure 5.23. The CA electron resists that are used mainly in current commercial EBL are the positive-tone FEP-171 [28], and the negative-tone NEB-22 [29]. These resists both have resolutions $<$100 nm and show excellent process performance, especially in photomask fabrication [30, 31].

However, with the need for $<$50 nm resolution – especially for direct-write applications – the development and evaluation of more advanced positive and negative CA resists is currently the subject of intense investigation [32]. For example, the 50 nm dense lines and 70 nm dots with high contrast, obtained with a recently developed and evaluated positive CA resist [33], are shown in Figure 5.24.

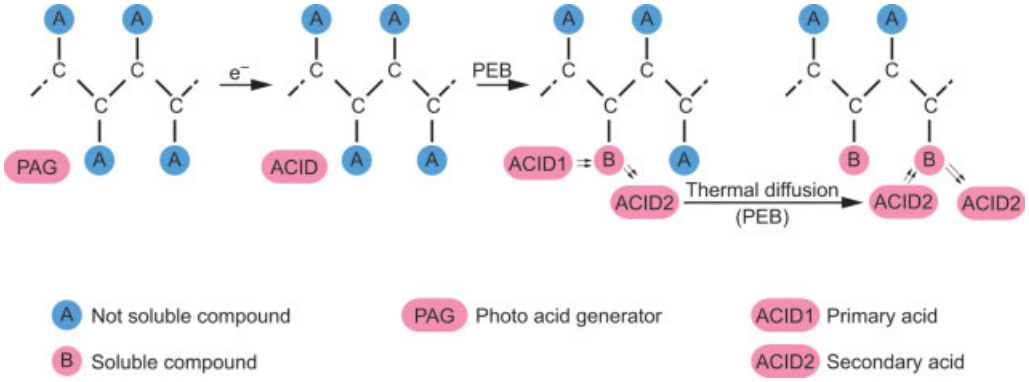


Figure 5.23 Reaction chemistry of a positive electron chemically amplified resist.

The major challenges in the development of CA resists for <50 nm resolution are to: (i) reduce the diffusion length during PEB; (ii) improve etch stability; and (iii) reduce the line edge roughness. For very small features, the molecular structure of the resist contributes to the roughness of the lines, which can be a significant fraction of the linewidth. A measure of line edge roughness is the standard deviation σ of the actual line edge relative to the average line edge. The reduction of line edge roughness is pursued by the application of resins with shorter molecules.

All of the electron resists discussed so far have been based on organic polymers. Although, in principle, it has been shown that such resists can achieve a resolution close to 10 nm, before applicable in manufacturing additional points must be taken into consideration, including the above-mentioned line edge roughness. As polymers are relatively large molecules, they cannot easily form smooth edges close to the atomic scale. Hence, in parallel to the improvement of CA resists, inorganic electron beam resists, such as hydrogen silsesquioxane (HSQ), are being pursued [34]. Initially, HSQ was used as a low-dielectric (low-*k*) material, with a *k*-factor of 2.5–3.0. In addition, HSQ demonstrates good spin-coating properties, such as good gap-fill, global planarization, and crack-free adhesion. It also shows excellent proces-

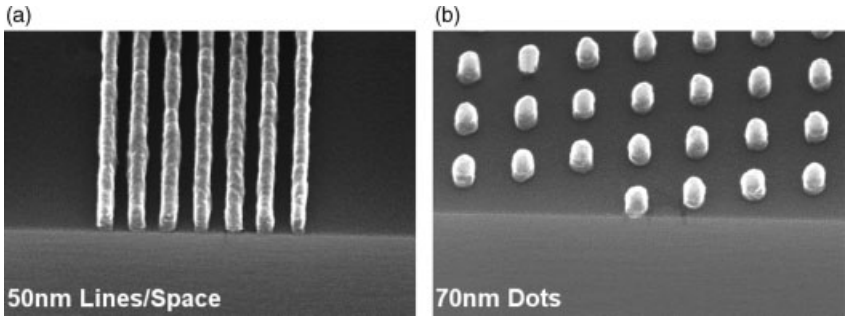


Figure 5.24 An advanced positive chemically amplified resist. (a) The 50 nm dense lines test pattern; (b) the 70 nm dots test pattern [33].

sing properties, notably a high thermal stability. HSQ is an oligomer composed of caged silsesquioxane within a linear Si−O network. A thermal curing is carried out to convert the caged species into a highly crosslinked network through the hydrolysis and condensation of the reactive Si−H functionalities.

Following the discovery that electron beam irradiation also initiates this curing reaction, HSQ was proposed as an inorganic electron resist. Our current understanding is that the Si−H bonds are broken during electron beam irradiation and are, in the presence of absorbed moisture, converted into silanol (Si−OH) groups. These silanol groups are unstable, and therefore condense, causing the caged molecule to break into a linear network. This transition drastically decreases the dissolution rate of the matrix within an aqueous base, thus enabling the use of HSQ as negative-tone electron resist. Furthermore, due to its high etch resistance HSQ can be utilized in a bi-layer resist process (BLR), where the patterns are transferred through a planarizing layer using reactive ion etching (RIE). As, with the rise in popularity of maskless lithography, BLR may become much more important, it is described in more detail in Section 5.2.3.3. With the implementation of a sensitizer chemical into the functional matrix (similar to the PAG in a polymeric CA resist), HSQ can, at least potentially, be made production-applicable.

5.2.3 Applications

Although EBL has a wide range of applications, the major focus is currently on the fabrication of transmission masks for DUVL and reflective masks for EUVL. Another important application is the direct-writing of patterns on wafers with single beams (which is also referred to as direct-write EBL; EBDWL). EBDWL is mainly used for fabricating device prototypes, and with recent improvements in shaped electron beam lithography tools, is also suitable for the low-volume production of application-specific integrated circuits (ASIC) and other specialized devices, for example hard disk heads. Multiple EBL such as maskless lithography (ML2), where a single electron beam is split into multiple beams to enable massively parallel EBDWL, shows the potential to complement or even replace optical lithography. The fabrication of imprint templates for NIL by using EBL techniques similar to photomask making is steadily gaining in importance. Finally, EBL is also combined with optical lithography in volume production, where it is used to fabricate critical structures such as gates. This approach – termed *mix-and-match lithography* production, or *hybrid lithography* if a single resist is utilized as both electron resist and photoresist – has special requirements.

5.2.3.1 Photolithography Masks

In optical lithography, the patterns on wafers are reproductions of those on a photomask. As the photomask is used for thousands of chip exposures, the quality of photomasks is critical for optical lithography. Photomasks are fabricated with techniques similar to those used in wafer processing (see Section 5.1). A photomask blank, a glass substrate with a deposited opaque film (usually chromium) is coated

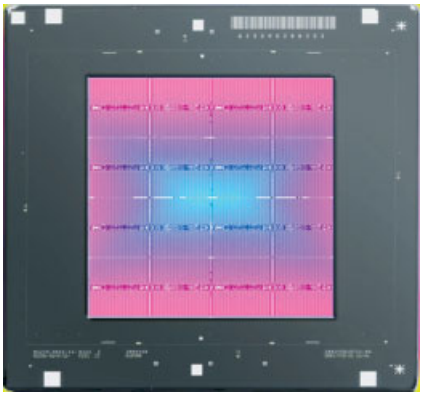


Figure 5.25 Photomask for deep ultraviolet optical lithography [35].

with a resist, and the latter is exposed with the pattern, developed, and the opaque film is etched. A processed photomask for DUV optical lithography is shown in Figure 5.25 [35].

Three types of photomask are currently in use. The simplest type is the binary mask (Figure 5.26a) [36], which employs only clear areas in the opaque chromium film to project the pattern to the wafer. Unfortunately, binary masks have the problem that, because of diffraction, the edges of the resist lines do not become straight. To rectify this problem, masks with molybdenum silicide films, which function as phase shifters, are used in phase-shift masks (PSM) (Figure 5.26b) [36]. Recently, so-called chromeless phase lithography (CPL) masks have been introduced for a resolution-enhancement technique (RET) called *off-axis illumination*, as shown in Figure 5.26c [36].

The pattern of a CPL mask featuring 125 nm lines applicable for 32 nm optical lithography is shown in Figure 5.27 [37]. An introduction to the functional principles of the different photomask types and their application in optical lithography is provided in Ref. [1].

Typically, photomask patterning is carried out with beam writers. Whilst for low- and medium-resolution masks, optical beam writers can be used, high-resolution masks are prepared using electron beam writers [38]. The general EBL techniques were described in Section 5.2.1; today, raster scanning has been replaced by vector scanning, and both Gaussian-beam and shaped-beam writers are currently employed for mask-making.

Gaussian-beam lithography is usually applied in a different way for photomasks than as described in Section 5.2.1.3. In addition to the IC patterns, photomasks also contain smaller features than the CD for optical proximity correction (OPC). When using Gaussian-beam writing, creation of the pattern in Figure 5.28 requires a beam size equivalent to the smallest feature (here, the right upper edge), but this leads to long writing times. However, it is possible to expose this pattern with a beam-size which is twice the size of the smallest feature. Because of the Gaussian intensity distribution, the 2σ circle touches the edges of the writing grid, whereupon the outline of the pattern can be made by multiple exposures of the spots; this is referred

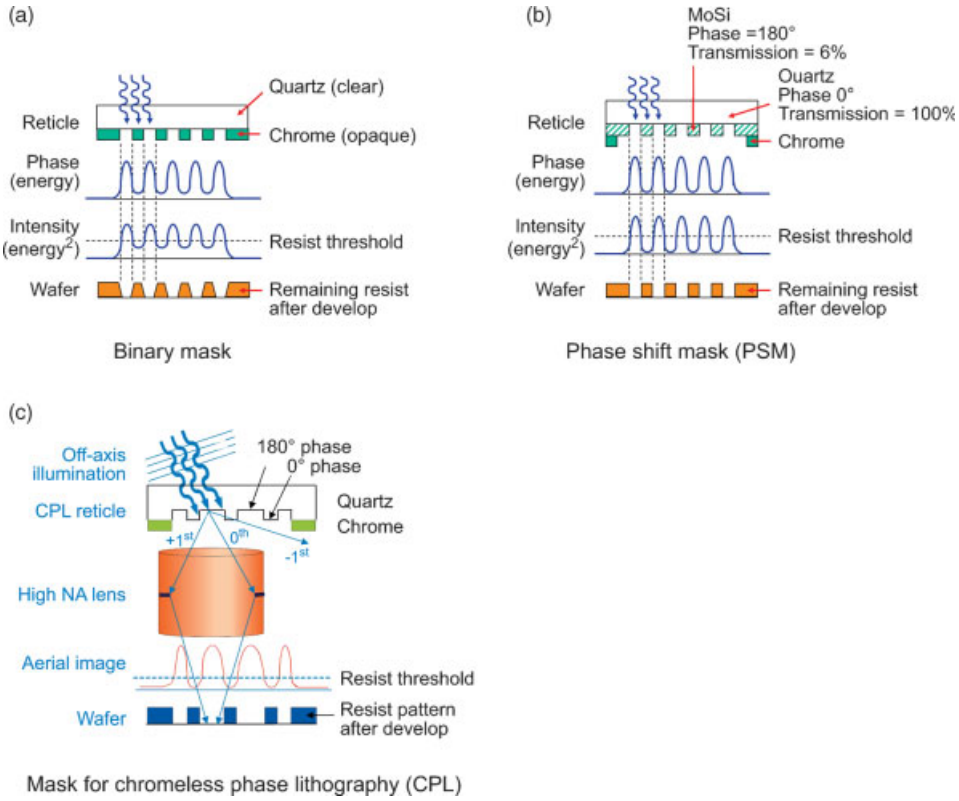


Figure 5.26 Mask types for deep ultraviolet optical lithography. (a) Binary; (b) phase-shift masks (PSM); (c) chromeless phase lithography (CPL) [36].

to as *multi-pass gray writing* [39]. The straight outlines are exposed four times, while the outlines of the pattern can be moved locally by exposing spots only three, two, or one times. Because of the overlap between spots and the multiple passes, this method has the additional effect of smoothing the exposure.

Although Gaussian-beam strategy is still used today, high-end mask-making has become a domain of 50 kV variable-shaped beam writers, which have a significantly higher throughput than Gaussian-beam writers. The shaped-beam strategy is applied as discussed in Section 5.2.1; however, mask-making encounters several specific issues, each of which must be addressed.

In addition to the proximity effect (see Section 5.2.1.5), which is a short-range phenomenon, long-range effects also appear which may stretch over large portions of the masks. The re-scattering of incident electrons at the objective electron lens can lead to a long-range background exposure, called a *fogging effect*. Further, an optimal exposure result may be compromised by succeeding processing steps, such as developing or etching. During developing, the concentration of the developer chemical decreases faster in areas with dense patterns, than in areas with sparse patterns. Similar effects appear in both wet and dry etching, and this may lead to long-

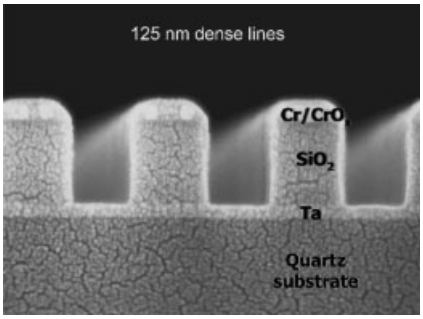


Figure 5.27 Photomask for deep ultraviolet optical lithography: 125 nm lines on mask result in 32 nm lines on the wafer [37].

range distortions of the final patterns, known as the *loading effect*. Whilst both fogging and loading effects can, in principle, be corrected within post-exposure process steps such as the PEB, the proximity effect correction methods of electron beam writers have been successfully augmented for handling both fogging and loading effects.

As shown in Section 5.2.1.5, forward scattering into the resist can be reduced by increasing the acceleration voltage of the electrons. However, this approach causes a significant increase in scattering within the substrate, and as a result the substrate is heated up. For silicon wafers, this problem is less pronounced, because silicon has a high thermal conductivity, and distortions of the wafer can be rectified by electrostatic chucks with wafer backside cooling. However, photomasks have a low thermal conductivity, and because of their large thermal capacity, the local temperature increases significantly during exposure. Therefore, mask writing tools currently do not exceed 50 kV acceleration voltage, and there is a reluctance to employ 100 kV tools.

The fabrication of high-end photomasks with EBL requires specialized post-exposure processing equipment. The PEB is a critical process step for CA resists, requiring a temperature uniformity of <0.1 K within the resist plane of the mask. Because of the large thermal capacity of photomasks, and their non-radial shape, such temperature uniformity is difficult to achieve. The PEB equipment is preferably connected directly to the electron beam writer, thus enabling the PEB and development to be conducted immediately after exposure. With such a direct connection, a

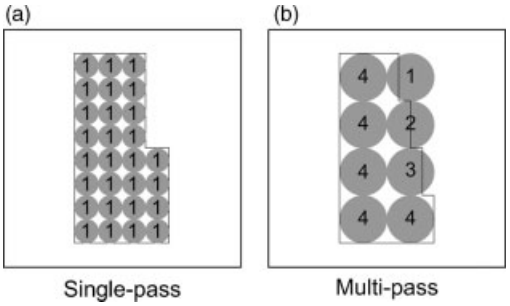


Figure 5.28 Writing a pattern with: (a) vector single-pass; and (b) with vector multi-pass strategies.

specially tailored post-exposure processing is applicable to compensate writing errors such as fogging and loading, and to improve overall pattern uniformity. The PEB is especially suitable for such compensation [40].

Besides photomask making, electron beam techniques have recently been introduced for the repair of photomasks [41] which, to date, has been the domain of ion beam structuring (see Section 5.3.2.1). If, due to a problem in the manufacturing process, a part of the opaque film is stuck when it should have been removed, it can be selectively removed by an electron beam-induced etching process; moreover, if part of the opaque film is damaged, it can be partially redeposited.

5.2.3.2 **Direct-Write Lithography**

Electron beam writers may also be used to create patterns on wafers directly, a process referred to as EBDWL. As shown in Section 5.2.1, EBDWL has the potential for fabricating features close to the atomic scale, and also provides very large depth-of-focus compared with optical lithography. As an example, a transistor demonstrator employing a gate with linewidth of 13 nm, and which has been fabricated by EBDWL utilizing vector Gaussian beam strategy [33], is shown in Figure 5.29.

EBDWL is especially suitable for fabricating device prototypes, as the device and the driving integrated circuit can be made directly from the computer-aided design (CAD) file. The production of device prototypes by optical lithography is not feasible, as a high-end photomask would first have to be made. Even if this effort were to be undertaken, it would not be possible to make any quick design changes. Notable applications of EBDWL in prototyping currently include research into nanoscale planar transistors (Figure 5.30) [42], and the next-generation transistor designs such as FinFET (Figure 5.31) [43, 44], or carbon nanotube FETs (CNTFET) [45].

In additional to the fabrication of device prototypes, EBDWL, with its recent improvements in repeatability and throughput of shaped EBL tools, has also been recognized as a feasible method for the low-volume production of ASICs and other special devices, for example hard disk heads.

As described in Section 5.2.1.5, forward scattering into the resist can be reduced by increasing the acceleration voltage of the electrons. While this approach causes a

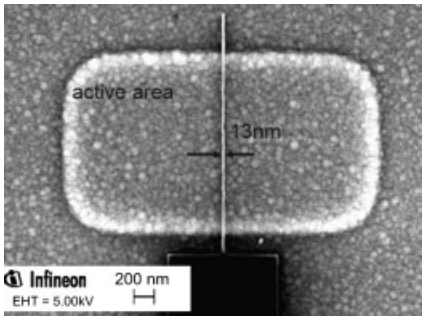


Figure 5.29 Nanoscale planar MOSFET demonstrator. A gate with a 13 nm linewidth prepared using direct-write electron beam lithography (EBDWL) [33].

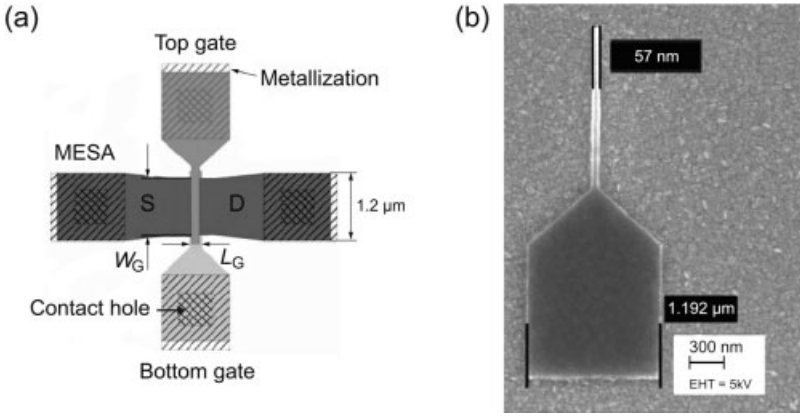


Figure 5.30 Nanoscale planar double gate transistor. (a) Top view of the design. (b) Scanning electron microscopy image of the first fabricated structure: bottom gate (produced by EBDWL) [42].

significant increase of the scattering within the substrate, and also heats up the substrate, in the case of silicon wafers this problem is less pronounced due to silicon's high thermal conductivity. As distortions of the wafer can be rectified by electrostatic chucks with wafer backside cooling, modern direct-writing tools employ an acceleration voltage of up to 100 kV.

Aside from the electron beam writer, specialized post-exposure processing equipment is required for reliable EBDWL. Similar to photomask fabrication, CA resists are employed, for which the PEB is a critical process step, requiring a temperature uniformity of <0.1 K within the resist plane of the wafer. However, because of the high thermal conductivity of silicon wafers, such uniformity is less difficult to achieve than with photomasks. As with photomask fabrication, the post-exposure processing equipment is preferably connected directly to the electron beam writer, enabling the PEB and development to be conducted immediately after exposure.

The fabrication of integrated circuits, either completely with EBL or with mix-and-match lithography, poses some significant challenges in integrating the required

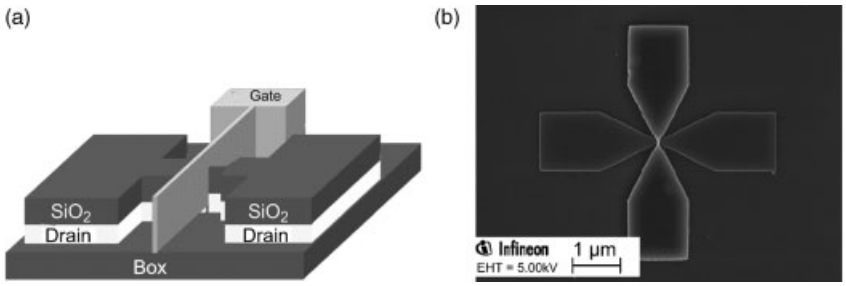


Figure 5.31 FinFET. (a) The design model [43]. (b) Scanning electron microscopy image of the actual device: with 50 nm gate (produced by EBDWL) [44].

process steps, and especially when aligning the different layers of patterned functional films. These challenges – and the methods used to overcome them – are discussed in Section 5.2.4.

5.2.3.3 Maskless Lithography

Considerable effort has been made towards making EBL available for volume production. The initial approach had been to implement either separate multiple electron optical columns [46] or separate multiple electron beams [47], to achieve massively parallel EBDWL. However, as the adequate calibration of either multiple columns or beams is a very challenging task, the suggestion was made to mimic optical lithography by introducing a transmission mask and projection optics with electromagnetic lenses to direct the electrons, which led in time to the development of electron projection lithography (EPL). In parallel, the use of a 1 : 1 transmission mask was also investigated, which led to the development of proximity electron lithography (PEL). As mentioned in Section 5.1, EPL was removed from the ITRS in 2004 due to significant difficulties with fabrication and application of the EPL transmission masks. PEL was subsequently removed in 2005, although its re-emergence cannot be ruled out completely. Whilst EPL and PEL are currently dormant, the advances made in electron optics have been significant, and hence the idea was conceived to devise electron projection and proximity techniques without the use of a mask – hence the term maskless lithography (ML2). As the current ML2 techniques are, loosely, derivatives of EPL and PEL, the details of both processes are explained in the following sections.

EPL: Principles and Limitations The transmission mask for EPL is a 4 : 1 stencil mask, a thin membrane through which holes are etched for the transmission of electrons. The stencil masks are themselves prepared using EBL, utilizing similar processes as for photomasks (see Section 5.2.3.1). Due to the instability of the membrane, fabrication has proven very challenging. In addition, a stencil mask absorbs electrons where there are no holes, thus causing the mask to undergo considerable heating, which then leads to distortions. Two concepts were devised to overcome this problem:

- SCALPEL (SCattering with Angular Limitation Projection Electron-beam Lithography) employed a scattering mask made from an extremely thin membrane (<150 nm) of low-atomic-number material (e.g., silicon nitride), through which the electrons can pass. The development of SCALPEL has been stopped, mainly because even the smallest deviation in mask membrane thickness resulted in intolerable intensity variations on the wafer. The main principles of SCALPEL are detailed in Ref. [48].
- PREVAIL (PRojection Exposure with Variable Axis Immersion Lenses) employed a stencil mask with a thick membrane (1–2 μm), thereby scattering the electrons to unexposed spots. This concept is quite similar to the vector-shaped beam strategy presented in Section 5.2.1.4, but instead of a simple shape a quadratic portion of the stencil mask was printed onto the wafer. The development of PREVAIL has also been stopped, mainly because even today it is not possible to make a 4 : 1 stencil

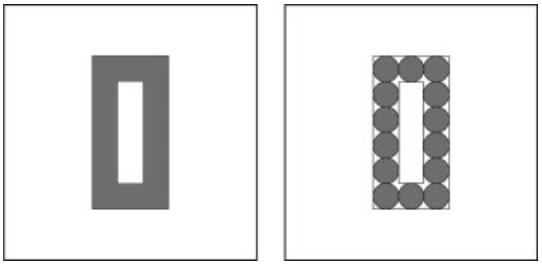


Figure 5.32 Electron projection lithography: the "donut problem".

mask covering an exposure field equal to that of an optical stepper with 26×26 mm, and therefore several masks must be stitched together to expose a single chip. Moreover, a stencil mask encounters the so-called "donut problem": the pattern shape shown in Figure 5.32 can only be made by two exposures with two complementary stencil masks, since on a stencil mask for a single exposure the center would fall out. For IPL, single stencil masks for fourfold exposure have been developed to circumvent this problem, as detailed in Section 5.3.2.1. The principle of PREVAIL is described in Ref. [49].

The electron optical columns for EPL experience a peculiar aberration in addition to the electron optical imperfections described in Section 5.2.1.2, namely the charge effects resulting from the electric charges of the electrons. Charge effects can be separated into a global charge effect, which influences any individual electron when traveling within an electron beam; this can be viewed as a continuous negative charge, with a charge density ρ. The single electron is within an electrostatic potential as of:

$$\nabla^2 \Phi = -\frac{\rho}{\varepsilon_0} \tag{5.18}$$

This electrostatic potential acts as an extended diverging lens. The global charge effect can, in principle, be compensated for by a suitable electron lens [50].

This is not the case for the stochastic charge effect, which is especially pronounced in a crossover, for example at demagnification, where all electrons interact within a small space. The stochastic charge effect leads not only to a beam blur but also to a beam energy spread, which in turn leads to chromatic aberrations. Global and stochastic charge effects are illustrated in Figure 5.33.

PEL: Principles and Limitations The transmission mask for PEL is a 1 : 1 stencil mask, and electrons are used for the proximity printing of this mask to the wafer. Initially, PEL employed 10 keV electrons, but currently low-energy electron-beam proximity lithography (LEEPL) [51], which utilizes low-energy electrons of 2 keV, is the PEL technique of choice.

The principle of LEEPL is shown in Figure 5.34. A single electron beam is generated in an electron beam column, and the mask is scanned. The low-energy electrons minimize the proximity effect, but forward scattering degrades the resolution.

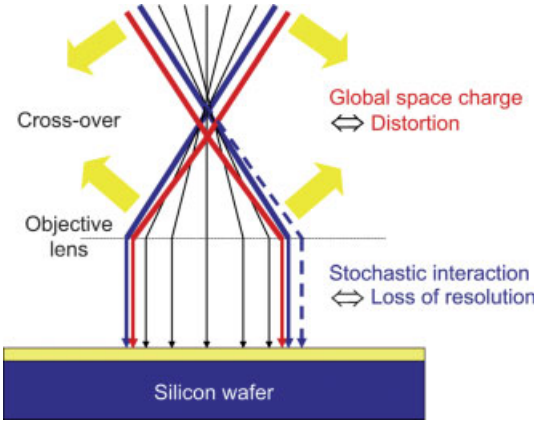


Figure 5.33 Electron projection space charge effects.

As the range of 2 keV electrons in the resist is <150 nm, the resist thickness for LEEPL is limited to 100 nm. In order to achieve the required aspect ratios, BLR [52] processes, which initially were developed for extending DUV lithography towards smaller features, must be applied (see Figure 5.35). Currently, LEEPL is not included in the ITRS as it has encountered several problems. For example, as this is a proximity technique, the distance between the stencil mask and the wafer is very small, usually <50 μm. Therefore, any distortion of the mask or wafer, or the presence of particles, would severely compromise the exposure. Furthermore, the implementation of bi-layer electron resist processes is not yet satisfactory. One positive development

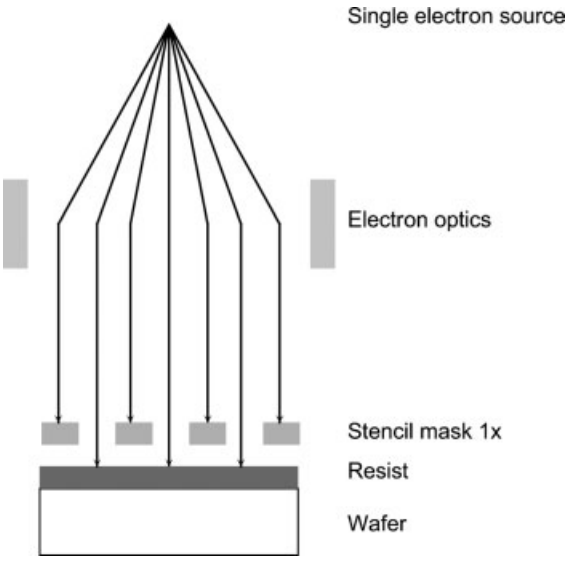


Figure 5.34 Proximity electron lithography techniques: LEEPL.

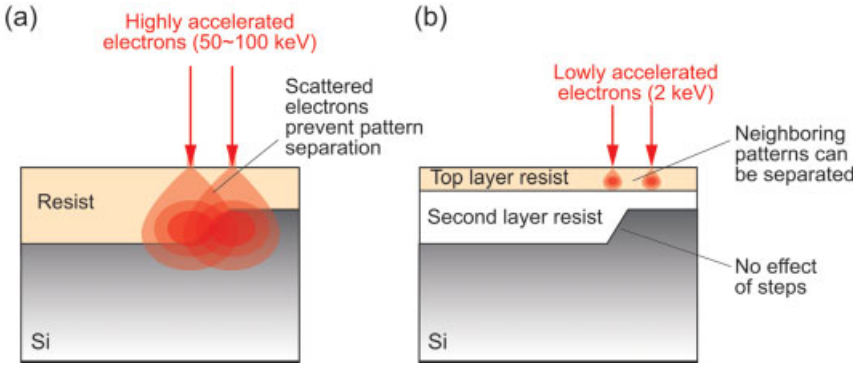


Figure 5.35 Comparison of single-layer (a) and bi-layer (b) resist processes [51].

however has been that, recently, stencil masks with a 26×26 mm exposure field could be prepared.

Projection Maskless Lithography Projection ML2 (PML2) [53] can be seen as the ML2 equivalent of EPL. Here, a single electron beam is split into multiple beams, with imaging being accomplished by a programmable aperture plate system (APS) [54, 55]. A range of innovative technologies was introduced to overcome the specific problems of both electron beam direct write and multiple beam application. The column of the demonstration system is shown in Figure 5.36 [53]. This employs a single electron source, therefore avoiding the control problems with multiple sources. The primary electron beam has a low acceleration voltage of $U = 5$ kV, and is widened by condenser optics to fully cover the APS. Because of the low energy of the electrons, the APS cover plate experiences no significant thermal expansion problems. Subsequently, hundreds of thousands of separate electron beams emerge from the APS,

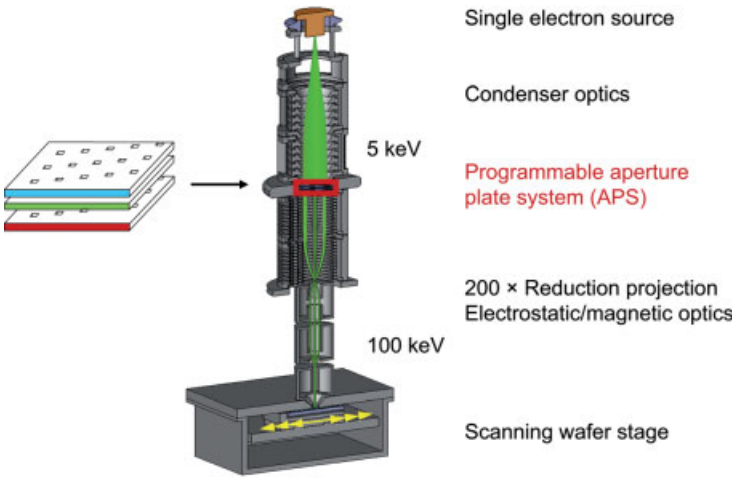


Figure 5.36 Schematic diagram of the PML2 multi-electron-beam column demonstrator [53, 54].

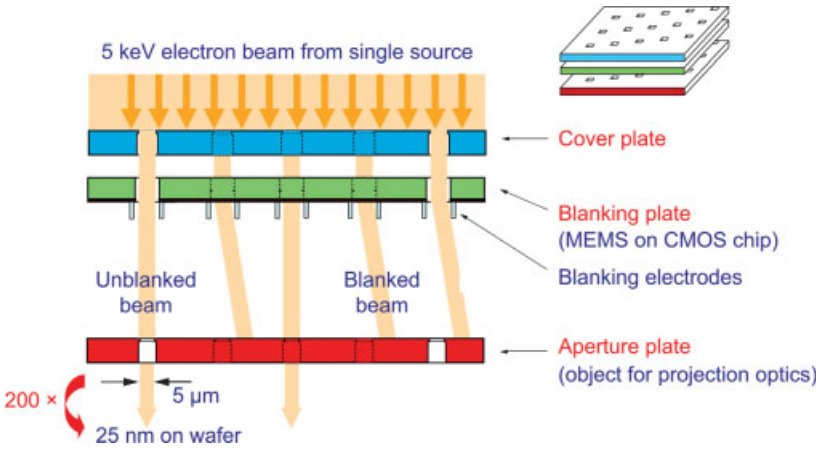


Figure 5.37 Schematic diagram of the multi-electron-beam modulator (APS) [53–55].

and are accelerated by $U = 100\,\text{kV}$, which results in a very high contrast when imaging on the wafer.

The APS, which is shown in detail in Figure 5.37 [53–55], consists of a cover, blanking, and aperture plate. The blanking plate employs MEMS-based structured electrodes for each of the transmission holes, which deflect the electron beam to strike the aperture plate and provide opaque features on the wafer. The diameter of each transmission hole is 5 µm. By using a two-stage electron optics, a reduction of ×200 is achieved, leading to a beam size of 25 nm on the wafer. Currently, the exposure area of the APS is $100 \times 100\,\mu\text{m}$.

With this exposure area, patterns on the wafer are written in stripes as shown in Figure 5.38 [53]. As discussed in Section 5.2.1, this may potentially lead to butting errors but, due to the small stripe size of $<300\,\mu\text{m}$ compared to current vector shaped-beam tools, no difficulties are expected.

The major challenge for PML2 is certainly to provide the required data transmission rate to the APS control electronics. A proof-of-concept (POC) tool was intended

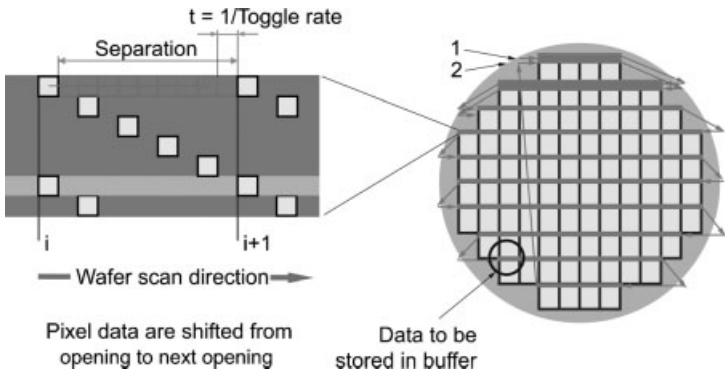


Figure 5.38 The writing strategy of a multi-electron beam system [53, 54].

within the MEDEA+ "CMOS logic 0.1 μm" project [56], for which a data transmission rate of 36 Gbit s^{-1} was demonstrated, scalable by channel count. A throughput of approximately 0.1 of a 300-mm wafer per hour was intended with this POC tool, although commercial tools should expose up to five 300-mm wafers per hour. However, this is still a small throughput compared to optical lithography steppers, with typical exposure throughputs of >100 wafers per hour. Nonetheless, this would be a great leap from EBDWL, which accomplishes much less than 0.1 wafer per hour in 65-nm patterning. Another issue is that the ×200 reduction requires all electrons to cross in a single region, thus leading to global and stochastic space charge effects, which potentially limit the throughput for the 22-nm node to less than one wafer per hour. To address this problem, an innovative PML2 scheme, with a throughput potential of up to 20 wafers per hour for the 32 and 22-nm nodes, is currently being investigated within the Radical Innovation MAskless NAnolithography (RIMANA) project [57].

Proximity Maskless Lithography Proximity ML2 (Mapper) [58] can be seen as the ML2 equivalent of LEEPL. Within Mapper, low-energy electrons of 5 keV are used, and the multiple electron beams are generated by splitting a single electron beam that originates from a single electron source. The multiple beams are then separately focused within an electrostatic lens array. The electron beams are arranged in such a way that they form a rectangular slit with a width of 26 mm, the same width as a field in current optical steppers. During exposure, the beams are deflected over 2 μm perpendicular to the wafer stage movement. With one scan of the wafer a full field of 26 × 33 mm can be exposed. During simultaneous scanning of the wafer, and deflection of the electron beams, these beams are switched on and off by light signals, one for each beam. The light signals are generated in a data system that contains the chip patterns in a bitmap format. The column of the proof-of-lithography (POL) tool, implementing 110 electron beams, is shown in Figure 5.39 [58]. A commercial tool would most likely implement 13 000 electron beams, so the bitmap would be divided over 13 000 data channels and streamed to the electron beams at up to 10 GHz, thus enabling a throughput of ten 300-mm wafers per hour.

As with LEEPL, the major challenges for Mapper are the problems arising with the proximity of the exposure, and the resist process. Additionally, as Mapper imple-

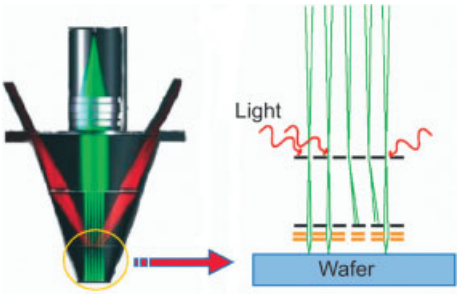


Figure 5.39 Schematic of the Mapper multi-electron-column demonstrator [58].

ments an electrostatic lens array within a <50 µm distance to the wafer, it is still unclear how the cleanliness of this array can be maintained during wafer throughput.

Overall, as ML2 techniques may provide the performance and throughput advantages of EPL and PEL, whilst avoiding the problems arising from mask fabrication and application, they should have the potential to rival optical lithography [59].

5.2.3.4 Imprint Templates

Nanoimprint lithography (NIL), considered to be a lithography method with the potential to rival optical lithography, is a technique where a patterned template is pressed onto a substrate coated with resist [60]. Currently, photoactivated NIL (PNIL), which uses a monomer resist with low viscosity, is considered to show the highest potential for volume production. The template, which must be constructed from a transparent material such as fused silica, is pressed onto the sample, after which a polymerization reaction is induced in the resist by applying UV light (thus, the technique is also called UV-NIL), and the template is removed.

Whilst NIL in itself does not involve exposure with photons or charged particles, the patterned template must be fabricated first, similar to a photomask for optical lithography. The templates are prepared by electron beam lithography, using similar processes as for photomasks (see Section 5.2.3.1). As the template is reproduced without demagnification, and therefore requires the same feature size as the pattern on the wafer, both fabrication and application still pose certain challenges. Currently, efforts are under way to utilize photomask fabrication methods for making PNIL templates [61]. As mentioned above, the templates must be transparent to UV light, and therefore fused silica photomask blanks may be used for template making. An overview of the template process flow is shown in Figure 5.40: in a first lithography step, the template is structured by EBL (first write). In a second lithography step, the pedestal required for imprint is made (second level write). Further details of the template process flow are presented in Refs. [62, 63].

When using photomask fabrication methods, currently four imprint templates are structured on a single photomask blank, as shown in Figure 5.41a and b. The photomask blank is then diced into separate templates (Figure 5.41c). As the dicing introduces contamination and mechanical strain, a modified fabrication approach must be developed before NIL can be employed in volume production. The size standard for templates resembles the exposure field of current optical lithography steppers (Figure 5.41d).

Complete details of the NIL imprint processes are described in Chapter 7 of this volume.

5.2.4 Integration

Although EBL is widely used in research because of its flexibility and high resolution, its low throughput and complex maintenance requirements of electron beam writing tools have limited the use of EBDWL in volume production. However, continuous improvements have led to the development of reliable tools with shaped beam writing

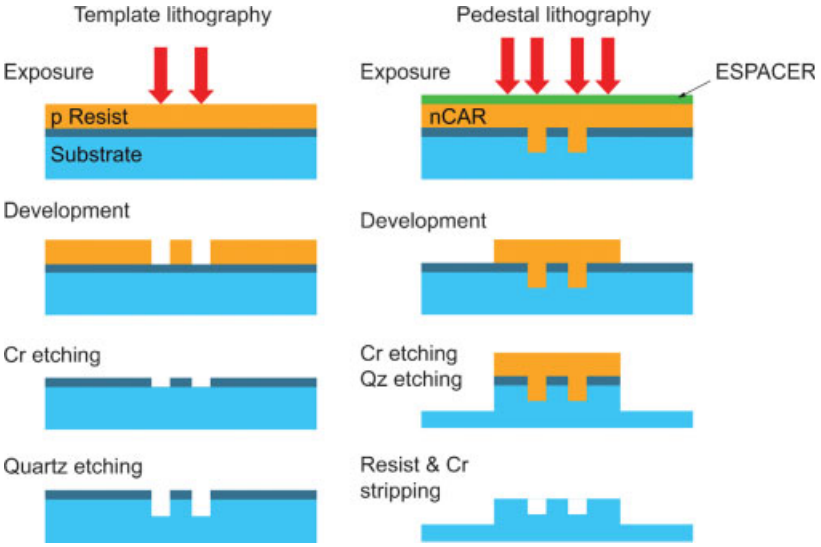


Figure 5.40 An overview of the fabrication process for two-dimensional templates [62, 63].

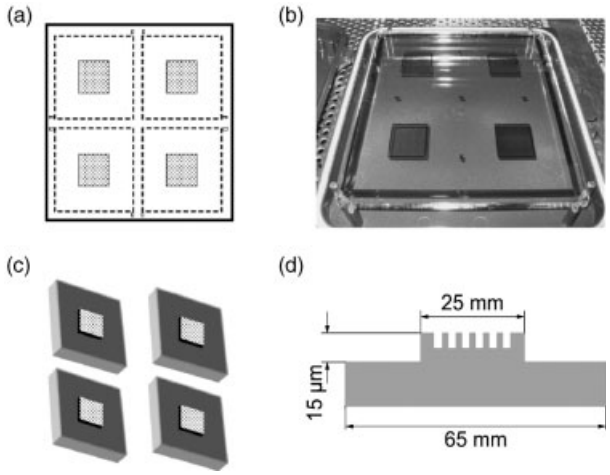


Figure 5.41 Application of photomask fabrication methods for template making [64].

strategy, which fulfill the requirements of the current 65 nm node fabrication which, according to the ITRS, are 65 nm dense lines, 45 nm isolated lines, 90 nm contact holes, and an overlay accuracy of <25 nm [16]. Yet, while the required resolution and repeatability has been achieved through developments in tools and CA resists, the overlay accuracy has long posed a major challenge.

Integrated circuits consist of several functional layers, for example metallizations and barriers, which must be fabricated sequentially. In optical lithography, each layer is patterned by a suitably made photomask. When using EBL, two scenarios can occur:

- Either all layers are made by EBL, which is the case for making device prototypes [65]
- Just one critical layer, for example the gates, are made by EBL, and all other layers are made by optical lithography.

The use of mixed EBL and optical lithography in volume production is referred to as "mix-and-match" lithography production [66], or hybrid lithography [67] if a single resist is utilized as both electron resist and photoresist.

In order to ensure that all subsequent layers are exactly matched, aligning techniques are used in optical lithography. One widely used alignment method in wafer steppers is through-the-lens alignment, where an alignment mark on the wafer is projected onto an alignment mark on the photomask, and a comparison is made. However, this approach is not possible with EBL tools; rather, two types of EBL alignment mark are currently used. The first option is to employ marks made from a film of high-atomic-weight material. This type of mark can be detected by secondary electron emission, but the method may lead to contamination issues and it is, therefore, mostly used only for back-end processing [68]. The second option is to create trenches as marks (see Figure 5.42), which are then scanned. This EBL alignment strategy has been used successfully in creating 65 nm node integrated circuits with hybrid lithography [69].

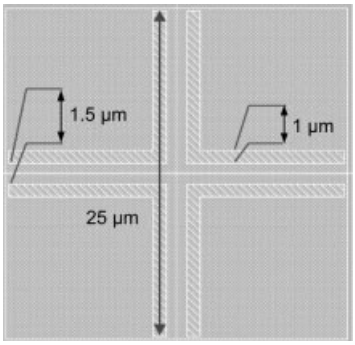


Figure 5.42 Trench alignment mark for EBDWL of a 25 × 25 µm integrated circuit [69].

5.3 Ion Beam Lithography

5.3.1 Introduction

Ion beam lithography (IBL) either utilizes ions to induce a chemical reaction in an ion resist for pattern formation, or can directly structure a functional film such as a metallization or barrier layer. When using ion resists, the wavelength of accelerated

ions is even smaller than that of electrons, because of their higher mass. For example, the mass of H^+ is 2000-times the mass of an electron, and therefore a calculation analogous to that in Section 5.2.1 yields:

$$\lambda = \frac{h}{\sqrt{2mQU}} \tag{5.19}$$

which, with an acceleration voltage of $U = 100\,\text{kV}$, gives $\lambda = 0.0001\,\text{nm}$.

Simple IBL tools use ion optics to focus ions from a source into a beam with a Gaussian energy distribution; therefore, they are referred to as focused ion beam (FIB) tools. The functional principle is analogous to the Gaussian EBL tools introduced in Section 5.2.1. A significant difference is that, because of the much higher mass of ions, a deflection is more difficult to achieve.

5.3.1.1 Ion Sources

IBL tools utilize volume ion sources [70], which consist of an ionizing region, where a plasma is formed, and an extraction region, where an electric field extracts the ions and accelerates them to form a beam. As with the thermal electron sources discussed in Section 5.2.1.1, the maximum current density j which can be obtained for an acceleration voltage V and an acceleration distance d is:

$$j = \frac{1}{9\pi d^2}\sqrt{\frac{2Q}{m}}V^{\frac{3}{2}} \tag{5.20}$$

5.3.1.2 Ion Optics

Ion optics function in a very similar manner to electron optics (see Section 5.2.1.2). The general trajectory representation as Equation 5.13 remains valid, when the electron charge e is replaced by the appropriate ion charge Q. However, there is an important difference with the design of ion optical columns: while magnetic lenses are used in electron optics, because of their lower aberrations compared to electrostatic lenses [70], in ion optics only electrostatic lenses can be used because, unlike magnetic fields, electric fields focus independently of the charge to mass ratio (see Section 5.2.1.2).

5.3.1.3 Patterning

The initial idea behind using IBL instead of EBL for fabricating integrated circuits was that ions scatter very little in solids. Unlike with EBL, there is no significant proximity effect, and therefore IBL can deliver very high resolution and contrast. Advances in EBL technology (especially proximity effect corrections), together with the fact that because of their high mass, ions are likely to damage functional films or doped areas on the substrate, EBL is the currently established technology for direct-write methods.

5.3.2 Applications

Although IBL is not used for integrated circuit fabrication, it is being applied and continuously improved for the direct-structuring of functional films in the

fabrication of special devices, such as nano-electromechanical systems (NEMS), nanophotonics, nanomagnetics, and molecular nanotechnology devices. Direct-structuring is also currently being investigated for the fabrication of imprint templates for NIL.

5.3.2.1 **Direct-Structuring Lithography**

Focused ion beam (FIB) tools can be used to create patterns in functional films, such as metallizations, or barriers on wafers directly, which is referred to as ion direct-structuring (IDS) lithography. The ions, when striking the functional film, cause the material to sputter, such that IDS is also known as *ion milling*. Another possibility is the local deposition of a functional film, with ions inducing the decomposition of a process gas at the surface of the wafer.

One major application of FIB tools is the repair of photomasks for optical lithography. As noted above (see Section 5.2.3.1), photomask quality is of utmost importance for yield. However, if due to a problem in the manufacturing process a part of the opaque film is sticking where it should have been removed, it can be sputtered away (see Figure 5.43) [71] or, if part of the opaque film is damaged, then it can be partially redeposited (see Figure 5.44) [71].

For some applications it is also reasonable to employ techniques initially developed for IPL, which introduced a transmission mask and projection optics with electromagnetic lenses to direct ions. This derivative of IPL is called ion projection direct structuring (IPDS). As mentioned in Section 5.1, IPL was removed from the ITRS in 2004 due to significant problems with fabrication and the application of the transmission masks.

IPL requires the use of stencil masks (see Section 5.2.3.3), because ions cannot pass through membrane masks, even with the thinnest imaginable membrane. However, a stencil mask absorbs the ions where there are no holes, and the resultant heating of the mask leads to its distortion. Further, with stencil masks it is not possible to make all required patterns with a single exposure, and so sets of complementary masks are required (see Section 5.2.3.3). Although initially these issues appear to make IPL impractical, studies to rectify the situation are ongoing. For example, thermal radiation cooling could be utilized to solve the heating problem,

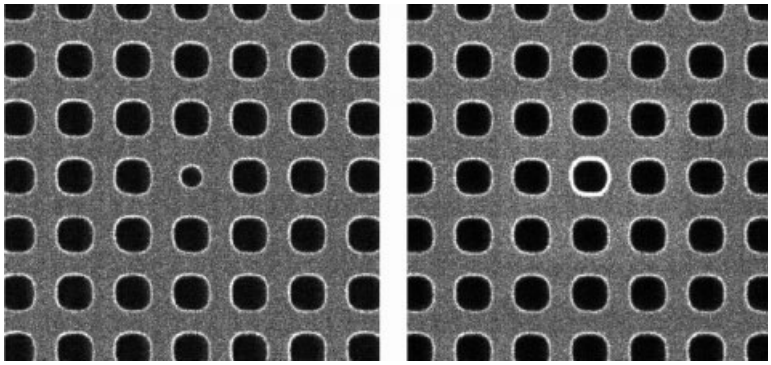


Figure 5.43 Opaque film defect: the repair of an undersized contact hole [71].

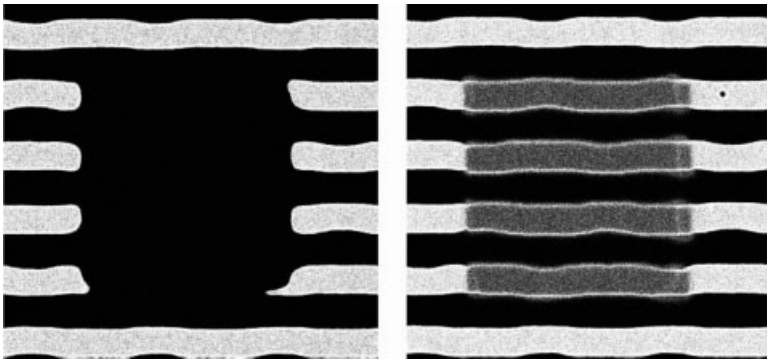


Figure 5.44 Pattern copy: the repair of missing lines [71].

while single stencil masks with fourfold exposure could enable complex patterns, but require half-sized features [54]. A comprehensive discussion of IPL, in addition to the details of a proposed IPL system, are presented in Refs. [54, 72].

IPDS can, for example, be applied to structure magnetic media for high-density data storage in a single exposure by inter-mixing the films of a multilayer structure [73]. For such an application a single stencil mask is sufficient [74].

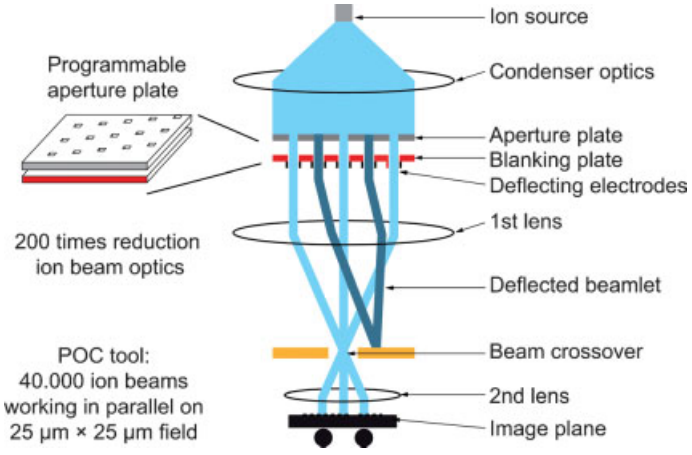


Figure 5.45 Schematic of multi-ion-beam column demonstrator [53, 75].

Considerable effort has also been made towards developing IPDS for volume production. The most-often investigated approach is to use the technique for multiple EBL (see Section 5.2.3.3), to split a broad ion beam into multiple beams, and to image with a programmable APS. This multiple ion beam projection maskless patterning (PMLP) technique is currently being developed within the project CHARPAN (CHARged PArticle Nanotech) [53, 75], and the column of the demonstration system used is shown schematically in Figure 5.45 [53, 75].

A possible multi-ion-beam tool resulting from CHARPAN would have a wide range of applications. In addition to IPDS, several other ion-beam-induced patterning

processes, such as beam-assisted etching, deposition, polishing, nanometer-resolved ion implantation, and ion-beam-induced mixing, are possible. All of these processes are considered to be fundamental for the fabrication of emerging nanoscale devices.

5.3.2.2 Imprint Templates

As discussed in Section 5.2.3, NIL requires the use of templates which are currently fabricated by EBL. However, this method requires a full process sequence similar to photomask making, such as resist coating, exposure, development, and etch to be applied to a blank. The use of a FIB tool enables direct-structuring of the chromium film on the blank, by sputtering.

Whilst FIB tools, compared to state-of-the-art EBDWL tools, lack the throughput required for sensible template fabrication, a reliable multi-ion-beam tool would clearly take over from EBL, and consequently the resistless fabrication of templates has been included within the CHARPAN project.

5.4 Conclusions

The continuous improvement of EBL has always placed it one step ahead of the most advanced optical lithography in integrated circuit fabrication. Although mask fabrication for optical lithography is still its principal application, EBL has become a time- and cost-effective technique for early device and technology development. Further, with mask costs currently showing huge increases, EBL represents a viable option for small-volume production, despite its comparatively low throughput. Additionally, even in medium-volume production, EBL is employed for writing critical layers within mix-and-match and hybrid lithography. Because of ever-increasing device complexity, applying EBDWL, using shaped-beam writing tools in combination with advanced CA resists, is mandatory. In parallel, efforts are being continued in the investigation of parallel electron beam writing systems (ML2), which show the potential almost to match the throughput of optical lithography, and thus may in time complement or even replace the latter process for high-volume production.

In contrast, in its current form, IBL is not applicable for integrated circuit fabrication, although further improvements in FIB tools towards IPDS, as well as the development of parallel ion beam writing systems (PMLP), may lead to its feasible application.

Integrated circuit fabrication aside, both EBL and IBL techniques are currently being used and continuously improved for the fabrication of special devices in low volume, such as nano-electromechanical systems, nanophotonics, nanomagnetics, and molecular nanotechnology devices.

Acknowledgments

The authors thank L. Markwort of Carl Zeiss AG, H. Loeschner of IMS Nanofabrication GmbH, and H.J. Doering and T. Elster of Vistec Electron Beam GmbH, for their

support. Special thanks are due to R. Waser of Forschungszentrum Jülich GmbH for carefully reviewing this chapter.

References

1 Levinson, H.J. (2001) *Principles of Lithography*, SPIE, The International Society for Optical Engineering, Bellingham, WA.

2 Rai-Choudhury, P. (1997) *Handbook of Microlithography, Micromachining and Microfabrication Vol. 1*, SPIE, The International Society for Optical Engineering, Bellingham, WA.

3 Attwood, D. (2000) *Soft X-rays and Extreme Ultraviolet Radiation, Principles and Applications*, Oxford University Press, Oxford.

4 International Technology Roadmap for Semiconductors, www.itrs.net.

5 Craighead, H.G. (1984) 10 nm resolution electron-beam lithography. *Journal of Applied Physics*, **55**, 4430.

6 Breton, B. (2004) *Fifty Years of Scanning Electron Microscopy*, Academic Press.

7 Hawkes, P.W. (1989) *Electron Optics*, Academic Press.

8 Ximen, J. (1990) Canonical aberration theory in electron optics. *Journal of Applied Physics*, **68**, 5963.

9 Hu, K. and Tang, T.T. (1998) Lie algebraic aberration theory and calculation method for combined electron beam focusing-deflection systems. *Journal of Vacuum Science & Technology B*, **16**, 3248.

10 Rose, H. and Wan, W. (2005) Aberration correction in electron microscopy. *IEEE Particle Accelerator Conference Proceedings*, p. 44.

11 Chu, H.C. and Munro, E. (1998) Computerized optimization of electron-beam lithography systems. *Journal of Vacuum Science & Technology B*, **19**, 1053.

12 Uno, S., Honda, K., Nakamura, N., Matsuya, M. and Zach, J. (2005) Aberration correction and its automatic control in scanning electron microscopes. *Journal for Light and Electron Optics*, **116**, 438.

13 Bas, E.B. and Cremosnik, G. (1965) Experimental Investigation of the Structure of High-Power-Density Electron Beams. First Electron and Ion Beam Science and Technical Conference Proceedings, p. 108.

14 Thomson, M.G.R., Collier, R.J. and Herriot, D.R. (1978) Double-aperture method of producing variably shaped writing spots for electron lithography. *Journal of Vacuum Science & Technology*, **15**, 891.

15 Pfeiffer, H. (1979) Recent advances in electron-beam lithography for the high-volume production of VLSI devices. *IEEE Transactions on Electron Devices*, **26**, 663.

16 Vistec Electron Beam, www.vistec-semi.com.

17 JEOL, www.jeol.com.

18 NuFlare, www.nuflare.co.jp.

19 Chang, T.H.P. (1975) Proximity effect in electron beam lithography. *Journal of Vacuum Science & Technology*, **12**, 1271.

20 Owen, G. and Rissman, P. (1983) Proximity effect correction for electron beam lithography by equalization of background dose. *Journal of Applied Physics*, **54**, 3573.

21 Murai, F., Yoda, H., Okazaki, S., Saitou, N. and Sakitani, Y. (1992) Fast proximity effect correction method using a pattern area density map. *Journal of Vacuum Science & Technology B*, **10**, 3072.

22 Hudek, P. and Beyer, D. (2006) Exposure optimization in high-resolution e-beam lithography. *Microelectronic Engineering Elsevier*, **83**, 780.

23 Haller, I., Hatzakis, M. and Srinivasan, R. (1968) High-resolution positive resists for electron-beam exposure. *IBM Journal of Research and Development*, **12**, 251.

24 Nippon Zeon Chemical Corp., www.zeon.co.jp.

25 Berger, L., Dieckmann, W., Krauss, C., Dress, P., Waldorf, J., Cheng, C.Y., Wei, S.L., Chen, W.S., Kao, M.J. and Tsai, M.J. (2005) E-beam direct-write lithography for the 45 nm node using a novel single substrate coat-bake-develop track. *SPIE Proceedings*, Volume 5751, p. 609.

26 Pacansky, J. and Waltman, R.J. (1988) Solid-state electron beam chemistry of mixtures of diazoketones in phenolic resins: AZ resists. *Journal of Physical Chemistry*, **92**, 4558.

27 Katoh, K., Kasuya, K., Sakamizu, T., Satoh, H., Saitoh, H. and Hoya, M. (1999) Chemically amplified positive resist for the next generation photomask fabrication. *SPIE Proceedings*, Volume 3873, p. 577.

28 Technical Bulletin, Fujifilm Arch Corp . (2001) EB Positive Resist for Mask Process FEP-171.

29 Technical Bulletin, Sumitomo Chemical Corp . (2001) Negative-type photoresist for electron beam lithography NEB22.

30 Irmscher, M., Beyer, D., Butschke, J., Constantine, Ch. Hoffmann, Th. Koepernik, C., Krauss, Ch. Leibold, B., Letzkus, F., Mueller, D., Springer, R. and Voehringer, P. (2002) Comparative evaluation of e-beam sensitive chemically amplified resists for mask making. *SPIE Proceedings*, Volume 4754, p. 175.

31 Irmscher, M., Butschke, J., Koepernik, C., Mueller, D., Springer, R., Voehringer, P., Beyer, D., Hudek, P., Tschinkel, M., Berger, L. and Dress, P. (2003) Investigation of e-beam sensitive negative-tone chemically amplified resists for binary mask making. *SPIE Proceedings*, Volume 5130, p. 168.

32 Schwersenz, A., Beyer, D., Boettcher, M., Choi, K.H., Denker, U., Hohle, C., Irmscher, M., Kamm, F.M., Kliem, K.H., Kretz, J., Sailer, H. and Thrum, F. (2006) Evaluation of most recently chemically amplified resists for high resolution direct write using a Leica SB350 variable shaped beam writer. *SPIE Proceedings*, 6153, 47.

33 Qimonda, www.qimonda.com.

34 Lutz, T., Kretz, J., Dreeskornfeld, L., Ilicali, G. and Weber, W. (2005) Comparative study of calixarene and HSQ resist systems for the fabrication of sub-20 nm MOSFET device demonstrators. *Microelectronic Engineering Elsevier*, **479**, 78–79.

35 Advanced Mask Technology Center, www.amtc-dresden.com.

36 ASML, www.asml.com.

37 Koepernik, C., Becker, H., Birkner, R., Buttgereit, U., Irmscher, M., Nedelmann, L. and Zibold, A. (2006) Extended process window using variable transmission PSM materials for 65 nm and 45 nm node. *SPIE Proceedings*, Vol. 6283, p. 1D.

38 Beyer, D., Löffelmacher, D., Goedel, G., Hudek, P., Schnabel, B. and Th. Elster, (2001) Tool and process optimization for 100 nm mask making using a 50 kV variable shaped e-beam system. *SPIE Proceedings*, Vol. 4562, p. 88.

39 Dameron, D.H., Fu, C.C. and Pease, R.F.W. (1988) A multiple exposure strategy for reducing butting errors in a raster-scanned electron beam exposure system. *Journal of Vacuum Science & Technology B*, **6**, 213.

40 Berger, L., Dress, P., Gairing, T., Chen, C.J., Hsieh, R.G., Lee, H.C. and Hsieh, H.C. (2004) Global CD uniformity improvement for mask fabrication with nCARs by zone-controlled post-exposure bake. *Journal of Microlithography, Microfabrication, and Microsystems*, **3**, 203.

41 Ehrlich, C., Edinger, K., Boegli, V. and Kuschnerus, P. (2005) Application data of the electron beam based photomask repair tool MeRiT MG. *SPIE Proceedings*, Vol. 5835, p. 145.

42 Weber, W., Ilicali, G., Kretz, J., Dreeskornfeld, L., Roesner, W., Haensch, W. and Risch, L. (2005) Electron beam lithography for nanometer-scale planar double-gate transistors. *Microelectronic Engineering Elsevier*, **206**, 78–79.

43 Kretz, J., Dreeskornfeld, L., Hartwich, J. and Roesner, W. (2003) 20 nm electron beam lithography and reactive ion etching for the fabrication of double gate FinFET

devices. *Microelectronic Engineering Elsevier*, **763**, 67–68.

44 Kretz, J., Dreeskornfeld, L., Schroeter, R., Landgraf, E., Hofmann, F. and Roesner, W. (2004) Realization and characterization of nano-scale FinFET devices. *Microelectronic Engineering Elsevier*, **803**, 73–74.

45 Seidel, R.V., Graham, A.P., Kretz, J., Rajasekharan, B., Duesberg, G.S., Liebau, M., Unger, E., Kreupl, F. and Hoenlein, W. (2005) Sub-20 nm short channel carbon nanotube transistors. *Nano Letters*, **5**, 147.

46 Parker, N.W., Brodie, A.D. and McCoy, J.H. (2000) High-throughput NGL electron-beam direct-write lithography system. *SPIE Proceedings*, Vol. 3997, p. 115.

47 Pickard, D.S. (2003) Distributed axis electron beam technology for maskless lithography and defect inspection. *Journal of Vacuum Science & Technology B*, **21**, 2834.

48 Berger, S. *et al.* (1994) The SCALPEL System. *SPIE Proceedings*, Vol. 2322, p. 434.

49 Okamoto, K. *et al.* (2000) High throughput e-beam stepper lithography. *Solid State Technology*, **5**, 118.

50 Harriot, L.R. *et al.* (1995) Space charge effects in projection charged particle lithography systems. *Journal of Vacuum Science & Technology B*, **13**, 2404.

51 Utsumi, T. (2006) Present status and future prospects of LEEPL. *Microelectronic Engineering*, **83**, 738.

52 Lin, Q. *et al.* (1998) Extension of 248 nm optical lithography: a thin film imaging approach. *SPIE Proceedings*, Vol. 333, p. 278.

53 IMS Nanofabrication, www.ims.co.at.

54 Loeschner, H., Platzgummer, E. and Stengl, G.(21–22 March 2002) *Projection-ML2 with programmable aperture plate, International Mask-Less Lithography Workshop, Erfurt, Germany,* (see also Ref. [55]).

55 Loeschner, H. *et al.* (2003) Large-field particle beam optics for projection and proximity printing and for maskless lithography. *Journal of Microlithography, Microfabrication, and Microsystems*, **2**, 34.

56 Doering, H.-J., Elster, T., Heinitz, J., Fortagne, O., Brandstaetter, C., Haugeneder, E., Eder-Kapl, S., Lammer, G., Loeschner, H., Reimer, K., Eichholz, J. and Saniter, J. (2005) Proof-of-concept tool development for projection mask-less lithography (PML2). *SPIE Proceedings*, Vol. 5751, p. 355.

57 RIMANA project, www.rimana.org.

58 Mapper Lithography, www.mapperlithography.com.

59 Lin, Burn J. (2006) The ending of optical lithography and the prospects of its successors. *Microelectronic Engineering*, **83**, 604.

60 Chou, S.Y. (1995) Imprint of sub-25 nm vias and trenches in polymers. *Applied Physics Letters*, **67**, 3114.

61 Sasaki, S., Itoh, K., Fujii, A., Toyama, N., Mohri, H. and Hayashi, N. (2005) Photomask process development for next generation lithography. *SPIE Proceedings*, Vol. 5853, p. 277.

62 Hudek, P., Beyer, D., Groves, T., Fortagne, O., Dauksher, W.J., Mancini, D., Nordquist, K. and Resnick, D.J. (2004) Shaped beam technology for nano-imprint mask lithography. *SPIE Proceedings*, Vol. 5504, p. 204.

63 Dauksher, J., Mancini, D., Nordquist, K., Resnick, D.J., Hudek, P., Beyer, D. and Fortagne, O. (2004) Fabrication of step and flash imprint lithography templates using a variable shaped-beam exposure tool. *Microelectronic Engineering Elsevier*, **75**, 345.

64 Institut für Mikroelektronik Stuttgart, www.ims-chips.de.

65 Pain, L. *et al.* (2006) Transitioning of direct e-beam write technology from research. and development into production flow. *Microelectronic Engineering Elsevier*, **83**, 749.

66 Narihiro, M., Wakabayashi, H., Ueki, M., Arai, K., Ogura, T., Ochiai, Y. and Mogami, T. (2000) Intra-level mix-and-match lithography process for fabricating sub-100-nm complementary metal-oxide-semiconductor devices using the JBX-9300FS point-electron-beam system. *Journal of Applied Physics*, **39**, 6843.

67 Steen, S.E. *et al.* (2006) Hybrid lithography: The marriage between optical and e-beam

lithography method to study process integration and device performance for advanced device nodes. *Microelectronic Engineering Elsevier*, **83**, 754.

68 Steen, S.E. *et al.* (2005) Looking into the crystal ball: future device learning using hybrid e-beam and optical lithography. *SPIE Proceedings*, Vol. 5751, p. 26.

69 Pain, L. *et al.* (2004) Manufacturing concerns for advanced CMOS circuit realization: EBDW alternative solution for cost and cycle time reductions. *SPIE Proceedings*, Vol. 5374, p. 590.

70 Melngailis, J. (1998) A review of ion projection lithography. *Journal of Vacuum Science & Technology B*, **16**, 927.

71 FEI Company, www.fei.com.

72 Kaesmaier, R., Wolter, A., Loeschner, H. and Schunk, S. (2000) Ion-projection Lithography Status and sub-70 nm Prospects. *SPIE Proceedings*, Vol. 4226, p. 52.

73 Loeschner, H. *et al.* (2002) Ion projection direct-structuring for nanotechnology applications. *MRS Proceedings*, Vol. 739.

74 Dietzel, A., Berger, R., Loeschner, H., Platzgummer, E., Stengl, G., Bruenger, W.H. and Letzkus, F. (2003) Nanopatterning of magnetic discs by single-step Ar^+ ion projection. *Advanced Materials*, **15**, 1152.

75 CHARPAN project, www.charpan.com.

6
Extreme Ultraviolet Lithography

Klaus Bergmann, Larissa Juschkin, and Reinhart Poprawe

6.1
Introduction

6.1.1
General Aspects

The ongoing reduction of structure sizes in semiconductor devices such as memory chips or microprocessors means that conventional optical lithography is reaching its physical and technological limits. This technology makes use of the demagnified imaging of structures on a mask onto a photo resist. Currently, optical lithography utilizes deep ultraviolet (DUV) light at 193 nm and a high numerical aperture (NA) optical system consisting of transmitting lenses. Generally, the achievable resolution at wafer level (RES) and the depth of focus (DOF) can be expressed by the Rayleigh formulas:

$$\mathrm{RES} = k_1 \frac{\lambda}{\mathrm{NA}}, \quad \mathrm{DOF} = k_2 \frac{\lambda}{(\mathrm{NA})^2} \tag{6.1}$$

Here, λ is the wavelength, NA is the numerical aperture ($=n \sin\alpha$, where n is the index of refraction of the medium between the wafer and the last optical element, and α is the half-opening angle of the beam), and k_1 and k_2 are process-dependent constants with values typically of approximately 0.5. Today, the best resolution is in the region of 60–65 nm, operating at a wavelength of 193 nm, a NA of 0.93, and a k_1-value of 0.31 [1]. Currently, the semiconductor industry is investigating all the possibilities for further reducing the structure size offered by Equation 6.1. That is to say, by reducing the wavelength, increasing the NA above unity, and operating with lower k_1-values. In this way, ASML – one of the leading stepper manufacturers – has successfully demonstrated a water-based immersion system with NA = 1.20 and $k_1 = 0.28$ for printing 45-nm structures. Another strategy is that of *double patterning*, where two masks are imaged successively onto the wafer, which permits a smaller k_1-value of 0.2. However, structure sizes below 32 nm are considered only to be achievable with a large reduction of the wavelength into the extreme ultraviolet

Nanotechnology. Volume 3: Information Technology I. Edited by Rainer Waser

ISBN: 978-3-527-31738-7

(EUV) range. A reduction from 193 to 13.5 nm in EUV lithography relieves the situation with the process constants and the numerical aperture. Thus, 16 nm is expected to be printable with a NA of 0.35 and k_1 equal to 0.41 [1].

According to the current roadmap for semiconductors [2], EUV lithography will be introduced into the production process at the 45 nm node during the year 2011, together with improved 193 nm technologies. In moving towards ever smaller features below 22 nm and approaching the physical limits of silicon-based chips, EUV seems to be the only photon-based solution from today's point of view [1].

The step from DUV to EUV, however, implies a variety of technological changes compared to the conventional technology. Operating in the EUV range requires that all components are held in a vacuum in order to avoid absorption in an ambient gas. EUV radiation has the strongest interaction with matter – that is, the highest cross-section for absorption. Typical penetration depths of EUV radiation into solids are in the range of a few hundreds of nanometers. The optical system and the mask to be imaged onto the wafer consist of reflecting multilayer mirrors, and the source is no longer a UV laser but rather an incoherent plasma source emitting isotropically with a wavelength around 13.5 nm. The technology still requires further development before reaching industrial maturity, which is expected to be achieved by about 2010.

A simple discussion of Equation 6.1 allows an estimation to be made of the required wavelength region aiming at a resolution of, for example, better than 70 nm and a depth of focus of more than several hundreds of nanometers. This consideration leads to a wavelength below 20 nm. Multilayer-based mirrors with a high normal incidence reflectivity at a wavelength of about 13 nm are currently available, and the semiconductor industry has fixed the wavelength to 13.5 nm as a standard to maintain lithium-based plasmas as an option, which have a strong line emission at this wavelength. With our current knowledge of all components, such as mirror reflectivity, sensitivity of the resist, parameters of the optical system and the desired wafer throughput, it is possible to specify the requirements for the source, which can be considered as the least known component of the system. Although a solution for the final concept has not yet been found, some early examples of source and system specifications have been identified, for example in Refs. [3, 4]. In addition, the actual requirements are updated continuously, taking into account new aspects and increasing knowledge [5].

The present chapter provides an overview of the system architecture of an EUV scanner, together with the demands made on each component, namely the light source, optical components for beam propagation and imaging, masks and resists. The current status of development and future challenges are also addressed.

6.1.2 System Architecture

A variety of EUV lithography systems have been installed during the past few years in order to test the whole chain, beginning at the source and ending at the wafer level or using simpler, high-NA systems with small imaging fields for printing fine structures [4, 6, 7]. The principle of an EUV scanner will be explained with the example of

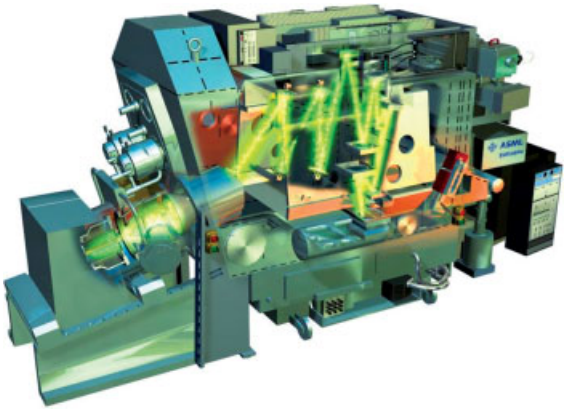


Figure 6.1 Schematic view of the ASML alpha demo tool, the first full-field scanner for extreme ultraviolet (EUV) lithography. The collector and source are not shown clearly; the beam propagation at the respective location of the source collector module is shown on the left.

the ASML alpha demo tool demonstrator, which can be regarded as the latest development and which is close to the future lithography tool with respect to design. A schematic of the alpha demo tool is shown in Figure 6.1 (taken from Ref. [8]). Using a collector – preferably a Wolter-type nested shell collector – the light of a plasma source is focused into the so-called *intermediate focus* as the second focal point of the collector (Figure 6.1 shows only the beam propagation from the source and the collector to the second focus, but not the hardware itself). The light is fed into the illuminator, which consists of a set of spherical multilayer mirrors and is used to produce a banana-shaped illumination of the mask (the top optical element in Figure 6.1). Another set of mirrors is used to image this field onto the wafer (the bottom optical element in Figure 6.1), with a typical magnification of 0.25. Wafer and mask are moved continuously to scan the whole mask and to transfer the structures onto a wafer of typically 300 mm diameter. In contrast to conventional DUV scanners, all of the components are contained inside a vacuum in order to achieve a high optical transmission for the EUV light. The etendue of the current optical system is around 3.3 mm^2 sr, which leads directly to a specification for the source size [9]. The optical system is able to use all the light from a spatially extended plasma source of around 1.6 mm in length along the optical axis and 1 mm in diameter in the radial direction, which is emitted into the solid angle of the collector. The plasma source is operated in a pulsed mode, ultimately requiring repetition rates of 7–10 kHz to guarantee a sufficiently homogeneous illumination of the resist. There are requirements on all components of the scanner – that is, the source, collector, optical system and components, masks, resist and on the system itself, concerning, for example, vacuum conditions and contamination issues. Today, the source is regarded as the most critical component, although this might also be due to the fact that other

components were developed prior to of the plasma source, and consequently less experience is available with this component.

The wafer throughput is furthermore dependent on the mechanical properties of the mask and the wafer handling system. According to Ref. [10], a simplified wafer throughput model can be formulated:

$$\begin{aligned} T &= T_{\text{scan}} N + T_{\text{oh}} \\ &= N(t_{\text{acc}} + t_{\text{settle}} + t_{\text{exp}} + t_{\text{settle}} + t_{\text{dec}}) + T_{\text{oh}} \\ &= N\left[\frac{2P}{a_{\text{w}} WR} + 2t_{\text{settle}} + \frac{(L+H)WR}{P}\right] + T_{\text{oh}} \end{aligned} \tag{6.2}$$

where T_{scan} is the scanning time per field, N is the number of fields per wafer, T_{oh} is the overhead time (wafer exchange, wafer alignment, ...), t_{acc} (t_{dec}) is the acceleration (deceleration) time, t_{exp} is the field exposure time, t_{settle} is the stage settling time after acceleration and before deceleration, P is the EUV intensity on wafer, a_{w} is the acceleration of the wafer stage, $W(H)$ is the field width (arc height + slit width) of the banana-shaped field, L is the field height, and R is the sensitivity of the resist. With this model the wafer throughput as a function of the reticle stage acceleration has been estimated for different illumination power levels ranging from 160 to 640 mW on the wafer [5]. The higher dose leads to a higher throughput only if the acceleration is increased; for example, an 80 wafers per hour throughput can only be achieved for the 640 mW, if the acceleration is more than 1.5 G. In the 160 mW case, a lower acceleration is required. With higher acceleration values and higher power levels at the wafer throughput is, of course, higher. These numbers are based on a $R = 5\ \text{mJ cm}^{-2}$ resist, a stage settling time of 25 ms, an overhead time of 11.5 s, a field size of 25×25 mm, a number of 89 fields, and an exposure slit of $2 \times 25\ \text{mm}^2$.

The first results have been obtained with the alpha demo tool, and the possibility of full-field imaging has been successfully proven. An example of different printed lines of 50 nm down to 35 nm and the corresponding line edge roughness (LER) using a resist of $18\ \text{mJ cm}^{-2}$ sensitivity, is shown in Figure 6.2. Full-field imaging

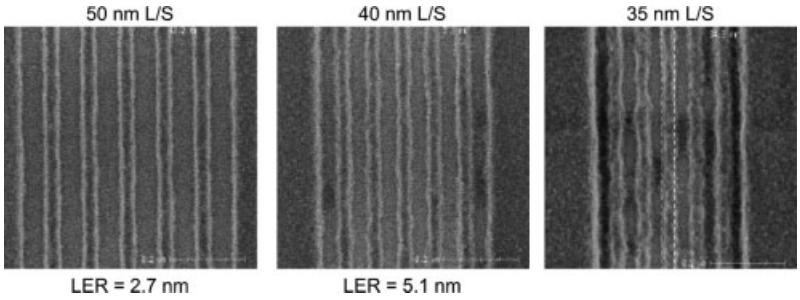


Figure 6.2 First results obtained with the ASML alpha demo tool of printed lines and spaces with resolution down to 35 nm and the respective LERs. The sensitivity of the resist was $\sim 18\ \text{mJ cm}^{-2}$.

over more than 20 mm slit height at the wafer level with a depth of focus of more than 240 nm has been demonstrated experimentally [6, 8].

6.2 The Components of EUV Lithography

6.2.1 Light Sources

According to Ref. [5], some of the requirements for the source and lifetime of the system for a production tool are as follows:

• Central wavelength (nm)	13.5
• Usable bandwidth (nm)	0.27
• Throughput (wafers h^{-1})	100
• EUV power at IF (W)	>115
• Repetition rate (kHz)	7–10
• Collector lifetime (months)	12
• Source electrode lifetime (months)	12
• Projection optics lifetime (h)	30 000
• Etendue of source output (mm^2 sr)	<3.3
• Spectral purity (% of EUV)	to be determined

The use of a multilayer mirror system (see Section 6.2.3) restricts the usable bandwidth to 2% around 13.5 or 0.27 nm, which is termed *inband radiation*. The throughput model is based on a 5 mJ cm^{-2} sensitivity of the resist, which has not yet been achieved (as discussed below). A less-sensitive resist would lead to higher source power specifications. The incoherent plasma source emits not only light but also debris in the form of particles, at least from the EUV-emitting plasma, irrespective of the source concept. Thus, some type of debris mitigation element is required between the source and collector appropriate to the actual source design. Typical collector half-opening angles range up to 70–80°. The total overall efficiency of the collector and the debris mitigation system can be estimated as around 20% of all the inband light emitted in the hemisphere of 2πsr [12], assuming a transmission of the debris mitigation system of 50%. This requires an inband emission of the source of at least 600 W/(2% b.w. 2πsr). Reasonable conversion efficiencies in the range of, at maximum, a few percent of the input energy for usable EUV radiation require a power input in the range of several tens of kilowatts. This imposes strict thermal demands, especially for the cooling of the debris mitigation system and the collector, which is the closest optical element to the source.

The multilayer mirrors have a finite transmission in the DUV range of 130 to 400 nm, which means that the emission from the source should not be too great within this wavelength region, as the resists are also sensitive in the DUV. The final

specification is still under consideration and is, for example, dependent on progress in spectral purity filters with minimum losses for EUV radiation.

6.2.1.1 Plasmas as EUV Radiators

Extreme ultraviolet radiation sources can be divided into thermal and non-thermal emitters. Non-thermal emitters are X-ray tubes or synchrotron radiation sources, where the radiation is generated by deflecting charged particles. Thermal emitters, based on the generation of hot plasmas, are a cost-effective and compact solution for EUV lithography. Generally, for thermal emitters matter is heated up to a high temperature, T, where the limit of the emission of light can be described by Planck's law of radiation:

$$B_\lambda(T) = \frac{2hc^2}{\lambda^2}\left(e^{\frac{hc}{k_B T_\lambda}} - 1\right)^{-1} \tag{6.3}$$

with speed of light, c, Planck's constant, h, and Boltzmann constant, k_B. For a black-body radiator, the temperature and the wavelength of maximum emission are related by Wien's law, which is derived from Equation 6.3:

$$\lambda_{max} T = 250\,\text{nm eV} \tag{6.4}$$

By aiming at a wavelength of $\lambda = 13.5$ nm, we obtain a temperature of around $T = 20$ eV (1 eV = 11 605 K). Such a high temperature is associated with matter in the plasma state. Usually, the emission of a plasma does not reach the Planck limit over the whole wavelength range, but only in individual strong emission lines of highly charged ions. Furthermore, for real plasmas the optimum emission is achieved at somewhat higher temperatures, depending on a variety of conditions, as discussed elsewhere [13]. The emission spectrum is characteristic of the respective element. Typical candidates discussed for EUV lithography are hydrogen-like lithium ions; that is, twofold ionized lithium with a strong single emission line at 13.5 nm, or tin and xenon as broadband emitters around 13.5 nm. In the case of xenon, the radiation around 13.5 nm arises from a transition of 10-fold ionized ions. For tin, the spectral efficiency is better compared to xenon (see Figure 6.3). With tin, more ionization levels exhibit transitions around 13.5 nm, leading to a more pronounced emission in the spectral range of interest.

Two concepts are pursued for generating such plasmas: laser-induced plasmas and discharge-produced plasmas. With laser-induced plasmas a pulsed laser beam is focused onto the target to be heated up. In the other case, the energy is taken from a pulsed electrical discharge. Many reports have been made concerning the different concepts, and discussing their special advantages and drawbacks [11]. In the next subsection, attention will be focused on the physical fundamentals of laser-induced and discharge-produced plasmas.

Irrespective of the individual concept, such plasmas are not only a source of light but also of debris consisting of fast ions and neutrals, clusters, droplets, and also heat. Sophisticated strategies are required to protect the optical system against this debris in order to avoid the deposition of matter onto the optics surface, or sputtering.

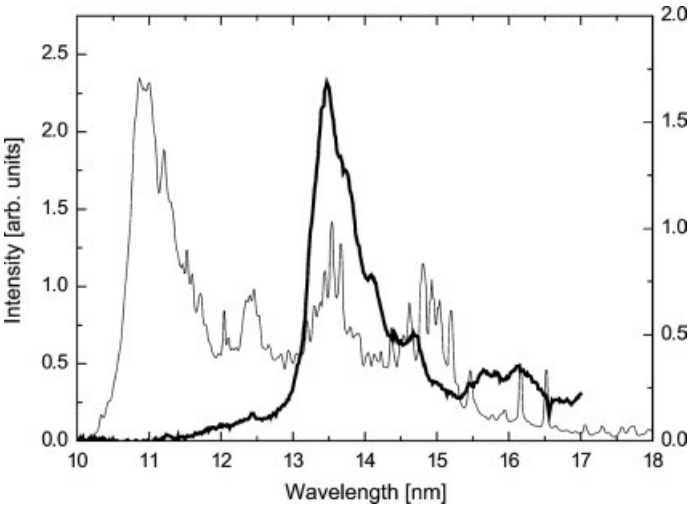


Figure 6.3 Typical emission spectra in the EUV for tin- and xenon-based gas discharge plasma sources. In the case of tin (bold line), the laser-induced plasma appears similar, whereas for xenon the laser-induced emission spectrum is smoother due to overlapping emission lines. The transitions of tin around 13.5 nm are iso-electronic to those of xenon around 11.0 nm.

6.2.1.2 **Laser-Induced Plasmas**

Hitting a target that is either solid, liquid or gaseous with a high-intensity laser beam leads to a plasma, where the laser energy is converted to thermal energy by inverse bremsstrahlung as the dominant process. Electrons are accelerated in the electrical field of the laser and transfer their energy to the ions. The laser energy is coupled to the plasma in a region where the plasma has the critical density, n_{crit}, which is dependent on the laser wavelength, λ_{laser} (ε_0 is the permittivity of free space, m_e the electron mass, e the electron charge, and ω_{laser} the laser frequency, that equals the plasma frequency at the critical density):

$$n_{crit} = \frac{\varepsilon_0 m_e \omega_{laser}^2}{e^2} = 1.11 \times 10^{21}\ \mathrm{cm}^{-3} \left(\frac{\mu m}{\lambda_{laser}}\right) \tag{6.5}$$

The temperature of the resulting plasma roughly scales with the laser intensity, I_{Laser}, according to $T_e \sim I_{Laser}^{4/9}$ [14, 15]. For a Nd:YAG laser with $\lambda_{laser} = 1.064\,\mu m$, the electron temperature, T_e, can be estimated as described in Ref. [16]:

$$T_e = 2.85 \times 10^{-4}\ \mathrm{eV} (I_{Laser}/(W/\mathrm{cm}^2))^{\frac{4}{9}} \tag{6.6}$$

Thus, in order to achieve an electron temperature of around 30 eV a laser intensity of $2 \times 10^{11}\ \mathrm{W\,cm^{-2}}$ is required, which is also observed experimentally as an optimum laser intensity [17, 18]. Typical pulse durations of a laser-induced plasma are in the range of nanoseconds or even less. The spatial extension of the EUV-emitting region is below 100 μm; thus, the etendue requirement of $<3.3\ \mathrm{mm^2\,sr}$ is easily fulfilled with

this type of plasma. Maximum conversion efficiencies of 5%/(2πsr 2% b.w.) for solid tin targets have been reported in the literature [18]. The conversion efficiency is defined as the ratio of usable inband EUV radiation into 2πsr to the incident laser light energy. In order to meet the source power requirement of 115 W in the intermediate focus, an average laser power of more than 5 kW is required, assuming an optimistic efficiency of 50% for the collector and the debris mitigation system. Obtaining high-power pulsed lasers at this level is an issue in current research and development activities. Different laser concepts are under discussion, such as pulsed CO_2 lasers or solid-state diode pumped laser, as reported elsewhere [19, 20, 22, 24]. However, it has not yet been shown that laser-induced plasmas can operate continuously on this power level. Further details on laser-induced plasma are available elsewhere [21, 23].

Besides the availability of the laser itself, the target is still an issue. Currently, different target concepts are under discussion, such as mass-limited targets to reduce debris production to a minimum level, and gaseous or droplets targets from frozen liquids or gases [18]. Most of the effort is currently being expended on tin-based targets, which have the highest expected conversion efficiencies.

6.2.1.3 Gas Discharge Plasmas

Producing the hot plasma by an electrical discharge is another well known method for the generation of light. For EUV-emitting plasmas, a pulsed electrical current is fed into an electrode system, which is filled with the working gas to be heated up at a neutral gas pressure of several tens of Pa. In a simplified concept, the current can be assumed to flow through a plasma cylinder, which is compressed by the self-magnetic field of the current to a high density of up to typically $n_e \sim 10^{19}\,\mathrm{cm}^{-3}$. The plasmas also experience ohmic heating, finally resulting in a dense and hot plasma column of several tens of eV electron temperature, and with a typical diameter of several hundreds of micrometers and a length in the range of few millimeters. The necessary current, I_o, can be approximated by assuming a Bennet equilibrium of the magnetic force and the plasma pressure [25]:

$$\frac{\mu_0}{8\pi^2}\frac{I_0^2}{r_p^2} = (n_i + n_e)k_B T_e \tag{6.7}$$

where μ_0 is the magnetic field constant, r_p is the radius of the compressed plasma column, and n_i and n_e are the electron and ion density, respectively. The term $r_p^2 \cdot (n_i + n_e)$ can be expressed by the starting radius, a, of the neutral gas column and the neutral gas pressure, p, by pa^2 [26]. As an example, for a xenon plasma with 10-fold ionized ions ($\langle Z\rangle = 10$, $n_e = \langle Z\rangle\ n_i$) and a desired electron temperature of 35 eV, a current of 8 kA results, which is also characteristic of the devices under investigation. This pulsed current is usually produced in a fast discharge of a charged capacity, C, which is connected in a low-inductive manner to the electrode system. Typical values for the inductance of the system are around 10 nH and few 100–1000 nF for the capacity. Stored pulse energies are in the range from 1 to 10 J.

A variety of different concepts exist for discharge-based plasmas, which differ mainly in the special geometry of the electrode system and the ignition of the plasma. For further information the reader is referred to numerous other reports [11, 27, 28].

6.2.1.4 Source Concepts and Current Status

During recent years, many industrial laboratories have increased their efforts into the development of sources for EUV lithography, and consequently a variety of concepts of either laser-induced plasmas or gas-discharge plasmas have been investigated, or are currently under investigation. An excellent overview of the current "players" and of the technological progress made can be found in Refs. [11, 27, 29]. An overview of the current status and progress, compared to the year 2000, for different concepts such as the hollow-cathode-triggered pinch plasma [30, 31], the plasma focus [32], different Z-pinch-like concepts [33–36] and laser-induced plasma [37], is shown in Figure 6.4 (from Ref. [28]). Of note, it is clear that rapid progress is being made, and that more powerful sources will be available in the near future. The overview in Figure 6.4 refers only to source power and repetition rate, which is of course not sufficient to assess a certain concept. For example, the source powers achieved do not generally refer to continuous operation but rather to short-term operations ranging down to a few seconds. Furthermore, other specifications such as the plasma source size or the lifetime must also be considered more closely. Often, only the best values are presented and the current status for the simultaneous achievement of specifications is difficult to identify.

One promising concept, which is also used in the ASML alpha demo tool described above, is the Philips laser-triggered vacuum arc [39] (Figure 6.5). This concept makes use of two electrodes, which rotate in a liquid tin bath to continuously re-cover the electrode surface. The system set-up is illustrated schematically in Figure 6.6. Both

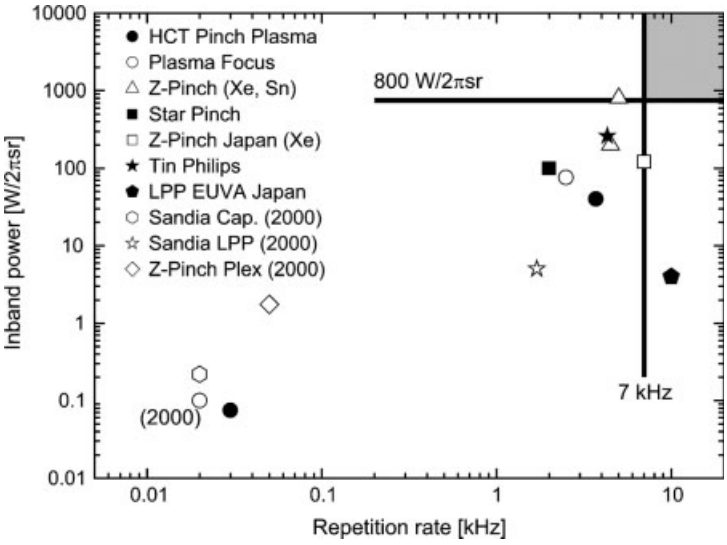


Figure 6.4 Currently achievable radiation power at 13.5 nm into 2% spectral bandwidth for different source concepts, and the corresponding repetition rates. Some data from 2000, taken from Ref. [38], are also shown to illustrate the rapid progress in source power.

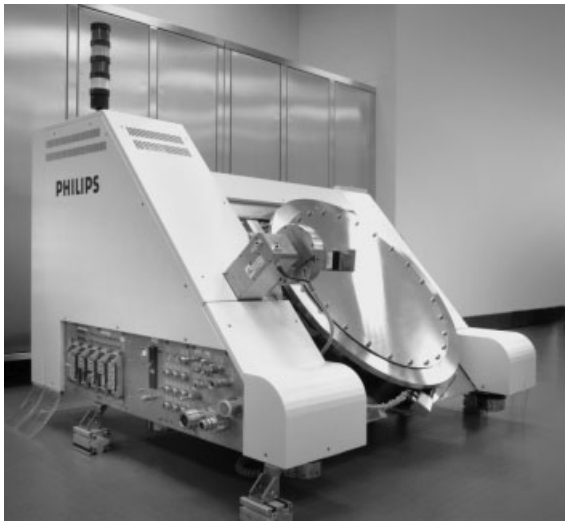


Figure 6.5 The Philips NovaTin EUV source based on the vacuum arc concept, which is used in ASMLs alpha demo tool.

electrodes are connected to a charged capacity, whereupon a laser pulse is used to evaporate a certain amount of tin, which also closes the electrical circuit in the gap between the two electrodes. The rapid discharge of the capacity heats up the tin plasma, which is used as an emitter of EUV radiation. Conversion efficiencies of up to 2.5%/(2πsr 2% b.w.) have been reported for this concept, with an average power of up to 300 W/(2πsr 2% b.w.) [40]. This is not too far away from the final specification for the source power. This concept has advantages with respect to cooling due to the rotating electrodes and electrode lifetime arising from covering the electrode surface with liquid tin and liquid metal cooling. However, a large amount of tin is also produced and emitted towards the collector. Thus, sophisticated means for debris

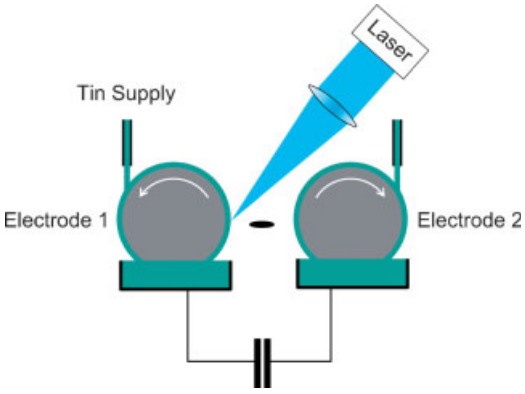


Figure 6.6 Scheme of the working principle of Philips NovaTin EUV source, based on the vacuum arc concept.

mitigation and cleaning strategies are required to guarantee a sufficient lifetime of the optical system. Recently, collector integration with a lifetime of more than 500 Gshots was successfully demonstrated using such a tin source [39]. Details of the collector design and the debris mitigation system are provided in the following section.

6.2.2 Collectors and Debris Mitigation

The currently preferred technical solution for collecting the light of the isotropically emitting plasma is the nested Wolter-type multi-shell collector [41]. A single collector shell consists of a hyperboloid and an ellipsoidal shell with identical focal points. The source is located within this common focal point, and is focused into the other focal point (intermediate focus), where the collector, although not an imaging element, leads to a "magnification" of the source by a factor of about 10 in the intermediate focus. The first optical element is about 50–100 cm behind this second focus. This type of collector allows light to be collected over a relatively large opening angle with a moderate gracing incidence angle at the reflecting surface, which is of particular importance for high reflectivity in the EUV. Such gracing incidence optics usually have a ruthenium coating with large reflectivity up to angles of $\sim 20^\circ$, which is typical of applications in EUV lithography. An example of a multi-shell collector which is used for modeling light distribution after the intermediate focus at the first optical element of the illuminator is shown in Figure 6.7. A collection angle of more than 80° half-opening angle with a total efficiency, including the finite reflectivity of the ruthenium coating, of more than $40\%/(2\pi\text{sr})$ is reported from the collector supplier [12].

Figure 6.8 shows the theoretical angular-dependent reflectivity of a ruthenium coating for different surface roughnesses, σ, and the typical range of operation for the collector. It should be noted that two reflections occur for the hyperboloid and the

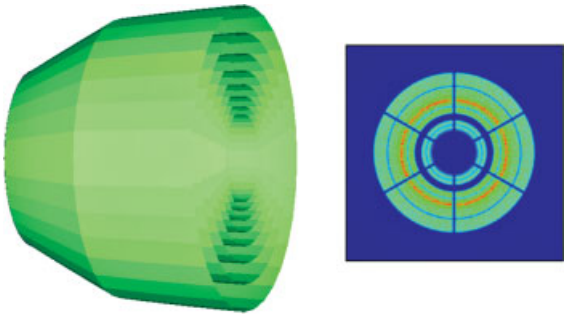


Figure 6.7 Schematic diagram of a Wolter-type nested shell collector with eight shells, as used in a ray-tracing calculation. This collector has opening angles between 11° and 45° corresponding to a collected solid angle of 1.7 sr or 27% of 2πsr. The right-hand diagram shows a simulated distribution in the far field for a spherical source of 50 μm FWHM.

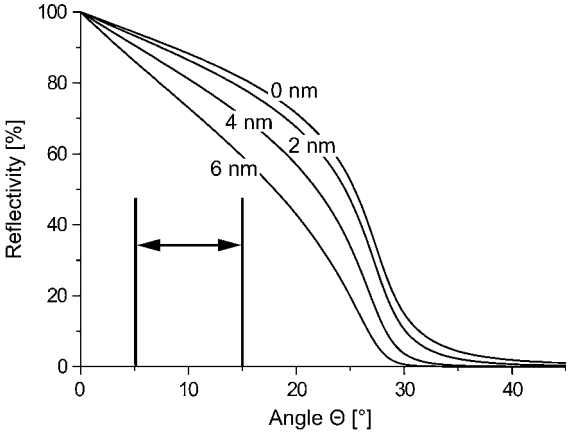


Figure 6.8 Calculated grazing incidence reflectivity of ruthenium for different roughness values, according to the CXRO database. The typical angle range for a multi-shell grazing incidence collector is also indicated. For collectors to be used in EUV lithography, the roughness is below 1 nm, leading to a reflectivity close to the theoretical limit.

ellipsoid. The data points are based on the atomic data for the refractive index published by CXRO [42]. The state of the art at the collector manufacturers – for example, Zeiss or Media Lario – involves surface roughness below 1 nm and chemically clean surfaces based on a physical vapor deposition (PVD) coating technology. Thus, the transmissions of the collectors are close to the theoretical limit. In addition, there is no longer any difficulty in fabricating substrates for the shells with diameters of several tens of centimeters.

Figure 6.7 also shows a typical light distribution for an eight-shell collector and a point source in a plane after the intermediate focus; this indicates the contributions of the different shells and the shadow of the mechanical support structure for the shells. Usually, the collector is located at a distance of only a few tens of centimeters from the plasma, which is also a thermal source in the 10 kW range. Cooling of the collector is achieved by using water-cooled lines around each collector shell. Results relating to the cooling capabilities are reported in Ref. [43], with a temperature increase of less than 1° when operated in the vicinity of a high-power source. Currently, research is going on in order to clarify whether this approach is sufficient, or whether more sophisticated cooling strategies must be applied, including for example a homogeneous temperature increase of the shell surfaces.

Another option is to have a normal-incidence collector based on a Schwarzschild design, with two spherical, multilayer coated mirrors. Some possible solutions to this problem are presented in Refs. [44–46].

As the closest optical element to the source, the collector experiences most of the heat load and debris emitted from the source. Consequently, overcoming the problems of a limited collector lifetime represents some of the major issues in current EUV lithography development activities. Today, such investigations are

under way not only in industry but also at various academic institutes, all of which have encountered this problem [47, 48]. Within the present chapter, the details of various debris mitigation schemes, and the results obtained, are restricted to the activities at Philips, whose clear aim is to integrate the above-mentioned tin-based gas discharge source.

The minimum number of particles necessary for the effective generation of an EUV-emitting pinch plasma is about 10^{15} atoms per pulse. These particles can be assumed to be emitted into 4πsr, partly redeposited onto the electrodes, and also emitted towards the optical system. For a 1-hour operation at 5 kHz this will require approximately 3.5 g of tin. Such an amount is clearly excessive, assuming that this will be deposited on the collector surface, where even a few nanometers' thickness is unacceptable due to the reduced reflectivity of a tin-coated surface. Hence, both the deposition of material and sputtering of the optical coating by the emitted particles must be avoided. The particles are emitted in the form of fast ions with energies exceeding at least 10 keV, as neutrals, and also as droplets from the wet electrode surfaces. One highly effective method of stopping and removing particles is the so-called "foil trap concept", which is described in more detail in Ref. [48]. The process, which is shown schematically in Figure 6.9, includes a system of lamellas located between the source and the collector. The foil trap is operated in combination with a buffer gas, usually argon with high transmission in the EUV. The emitted particles are deflected and finally stopped by the ambient argon atoms, and then stick to the walls of the foil trap. In the case of tin, the foil trap is heated above the tin melting point in order to avoid an accumulation of tin in the system of the lamellas. Using only this technique, a collector lifetime of more than 10^9 shots has been reported for an operation with a tin-based discharge source [39]. However, the foil trap concept does not permit complete suppression of the emission of particles towards the collector, and consequently for longer operating times a deposition of tin on the

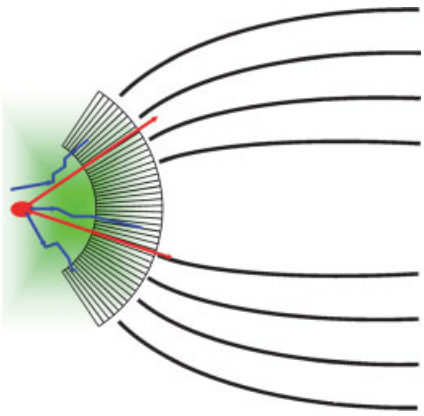


Figure 6.9 Schematic diagram of a foil trap system including buffer gas to protect the multi-shell collector against debris from the source. The foil trap has a high optical transmission, while the particles are efficiently stopped.

nanometer scale would be expected. In an attempt to overcome this problem several cleaning strategies for the collector have been proposed. One possibility would be to flood the collector chamber with a halogen gas, such as chlorine or iodine; the gas reacts with the tin to form volatile tin halides, which can be pumped away, while the ruthenium coating is unaffected. Using this technique permitted the complete recovery of a tin-coated ruthenium surface [49].

6.2.3 Multilayer Optics

For EUV radiation, the index of refraction is close to unity, and the absorption in matter is relatively high. A high reflectivity at surfaces is only achieved for incidence angles of the light of typically below 20°. This feature is, for example, exploited with gracing incidence optics as presented above. A high reflectivity for normal incidence is only achieved with multilayer systems, as shown schematically in Figure 6.10. Such multilayer systems consist of alternating layers of so-called "spacer material" and "absorber material", which have different indices of refraction and are thus reflective at their boundaries. Part of the incident light is reflected at each layer boundary, and the superimposed beam exhibits a high intensity. A well-known example, which is also used in EUV lithography, is a system consisting of silicon and molybdenum with high peak reflectivity around a central wavelength of 13–14 nm. A transmission electron microscopy (TEM) image of a real mirror is also shown in Figure 6.10. The

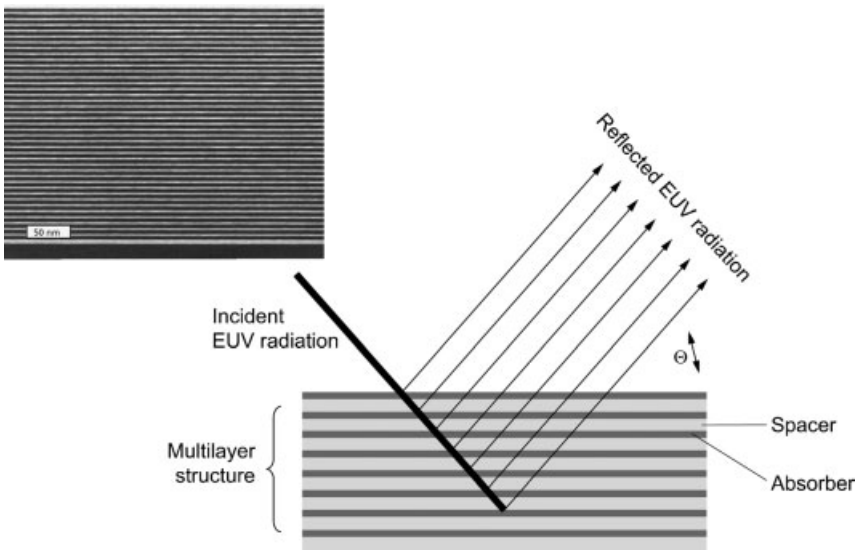


Figure 6.10 Schematic diagram of a multilayer mirror consisting of spacer (e.g., silicon) and absorber (e.g., molybdenum), where a high reflectivity is achieved by superimposing all rays reflected at the boundaries. A transmission electron microscopy image of a real Mo/Si system is shown in the top left section of the figure.

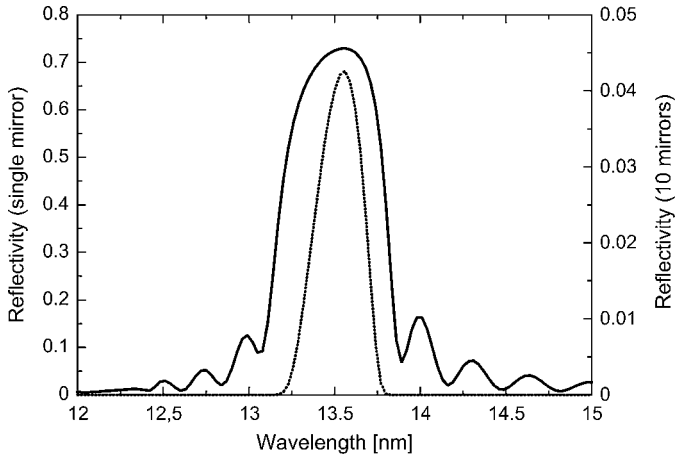


Figure 6.11 Wavelength-dependent reflectivity (solid line) of an ideal Mo/Si multilayer mirror according to the CXRO database, and the resulting reflectivity of a 10-mirror system (dotted line).

center wavelength of the maximum reflectivity of such a system can be expressed in terms of the total thickness, d, of a bi-layer, the incident angle, θ, ($\theta = 90^\circ$ corresponds to normal incidence) and a material constant δ' by the Bragg equation:

$$m\lambda = 2d \sin\theta \sqrt{1 - \delta'/\sin^2\theta} \tag{6.8}$$

where δ' is the weighted material constant for both elements with a complex index of refraction, $n = 1 - \delta + i\beta$. By using the atomic data of the CXRO database [43], Figure 6.11 shows the reflectivity of an ideal Mo/Si multilayer system with zero roughness and no intermixing of the layers with a peak reflectivity of more than 70%. Values of approximately 70% are also achieved for real mirrors. A multilayer reflectivity close to this maximum is achieved at different locations. In order for systems to be used in an EUVL scanner, some losses occur due to the capping layers necessary for avoiding contamination, or additional layers for improving the thermal stability of the multilayer system. Figure 6.11 also shows, graphically, the transmission of a corresponding system of ten multilayer mirrors, which are typically used in EUV lithography. The overall transmission is only 4%, as the theoretical limit and the bandwidth decrease to below 0.3 nm FWHM. This is also the main reason for the restriction to only 0.27 nm or 2% of the 13.5-nm bandwidth, as discussed for the source specifications.

A variety of activities have been developed to improve the multilayer coatings to be used in EUV lithography systems with respect, for example, to increasing the reflectivity or achieving better thermal stability. In order to achieve a higher reflectivity and better thermal stability, additional layers of boron carbide (B_4C) are introduced to reduce the interdiffusion of silicon and molybdenum at their boundaries [50]. This diffusion leads effectively to a higher surface roughness, and thus to a

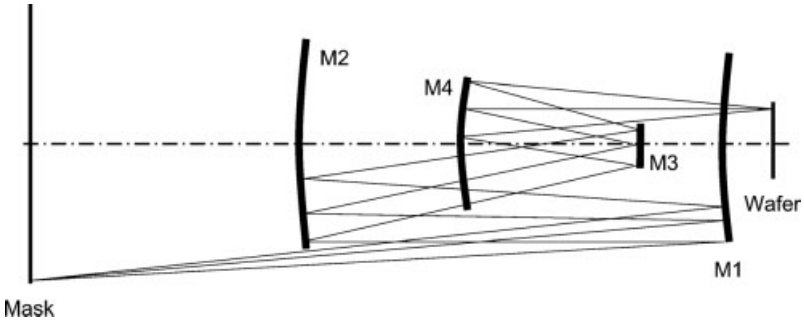


Figure 6.12 Schematic drawing of the projection optics of the ETS, consisting of four multilayer-coated reflective mirrors.

reduction in reflectivity. Furthermore, pure silicon molybdenum interfaces tend to be unstable and show even higher diffusion for temperatures above 100 °C. Here, ruthenium inter-layers are discussed as an alternative to boron carbide for increasing thermal stability [51]. Such protective layers are of special interest if multilayer coated components are to be used as collectors, as these must be heated for debris mitigation purposes [52]. Other activities are aimed at an improved coverage of the multilayer coating to protect against contamination or oxidation or at the suppression of the deep UV (100–200 nm) reflectivity in comparison to the EUV reflectivity. The latter strategy is one of several such approaches, including special thin filters for the deep UV, which are intended to fulfill the specification of spectral purity at the resist [53].

An imaging system for an EUV scanner consists of a number of multilayer-coated reflective mirrors. The design and specifications are discussed here as examples of the optical system of the Engineering Test Stand (ETS), which was the first full-field imaging system based on a plasma source. Details of the optical system can be found in Ref. [54]. A schematic of the projection optics consisting of four mirrors with NA = 0.1, a magnification of 0.25 and a resolution of 100 nm, is shown in Figures 6.12 and 6.13. Mirrors M1 and M3 are convex, while M2 and M4 are concave. The beam propagates off-axis, as indicated in Figure 6.13. In this special case, the mirror

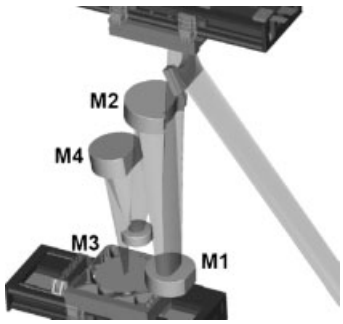


Figure 6.13 A three-dimensional view of the ETS projection optics system.

diameters are 165 mm for M1, 209 mm for M2, 104 mm for M3, and 170 mm for M4. The corresponding radii are −3055 mm for M1, +1088 mm for M2, −389 mm for M3, and +504 mm for M4 [55], where "+" indicates a concave and "−" a convex surface. Usually, a system of several mirrors is chosen in order to have sufficient degrees of freedom for the correction of aberrations and other imaging errors.

Typical diameters of the mirrors reach 200 mm for the ETS system, and even more for other optical systems. In order to meet the imaging specifications, the root mean square (RMS) figure error and the roughness must be below a certain level. The surface specifications will be discussed in more detail. Usually, the surface topology is described by a function $z(x, y)$. For simplicity, the following discussion and definitions are for the one-dimensional case $z(x)$. The extension to two dimensions is described elsewhere [56].

The average of the surface height is defined:

$$\bar{z} = \lim_{L\to\infty} \frac{1}{L} \int_{-L/2}^{L/2} z(x)dx \tag{6.9}$$

with L being the spatial extension under consideration of the surface. The surface roughness, σ, is given by:

$$\sigma^2 = \lim_{L\to\infty} \frac{1}{L} \int_{-L/2}^{L/2} (z(x) - \bar{z})^2 dx \tag{6.10}$$

It is useful to discuss the Fourier transform of the surface in terms of the spatial frequency, f_x:

$$Z(f_x, L) = \int_{-L/2}^{L/2} z(x) e^{-2\pi i f_x} dx \tag{6.11}$$

The power spectral density (PSD) function is often used for characterization of a surface, which can also be directly measured in scatterometry [56] and can be related to the Fourier transform $Z(f_x, L)$:

$$\mathrm{PSD}(f_x) = \lim_{L\to\infty} \frac{1}{L} |Z(f_x, L)|^2 \tag{6.12}$$

As defined in Equation 6.10, the roughness is the integral over all spatial frequencies of the PSD function. Often, different regions are defined in the specifications depending on the respective frequency interval $f_{\min}$ to $f_{\max}$

$$\sigma^2_{\Delta f} = 2 \int_{f_{\min}}^{f_{\max}} \mathrm{PSD}(f_x) df_x \tag{6.13}$$

The surface figure error corresponds to frequencies typically ranging from the inverse aperture to 1 mm^{-1}. This type of error is responsible for aberrations. The

roughness in the mid-spatial frequency range (MSFR), from $1\,mm^{-1}$ to $1\,\mu m^{-1}$, determines flare and contrast. The high spatial frequency range (HSFR) includes all frequencies above $1\,\mu m^{-1}$. The HSFR roughness influences the EUV reflectivity; for the ETS optical system, a surface figure roughness of <0.25 nm RMS, a MSFR roughness of <0.2 nm RMS, and a HSFR roughness of <0.2 nm RMS are specified. The specifications for the Zeiss projection optics system for the ASML alpha demo tool are similar [57]. Here, the figure error should be <0.2 nm RMS, and the MSFR and the HSFR roughnesses should be between 0.1 and 0.2 nm RMS. Different analysis methods for determining the PSD function show that these specifications are fulfilled [57], which is also confirmed by the successful printing of small structures with diffraction-limited resolution.

6.2.4 Masks

In contrast to conventional DUV lithography, masks are also based on multilayer-coated reflective mirrors. A cross-section of a mask is shown schematically in Figure 6.14. The mask blank is defined as that part including the substrate and a protective layer of, for example, SiO_2 necessary for the patterning process. The structures to be imaged onto the wafer are written onto the surface using an absorber layer of typically 100 nm thickness. The preferred absorber materials are Cr, TaN, Al or W. However, as the mask is imaged, there are additional specifications in comparison to the multilayer mirrors for the optical system [58, 59], and these will be addressed in the following.

The substrate must have a low thermal expansion coefficient (CTE) of typically less than 5 ppb K^{-1}, as approximately 40% of the incident EUV light is absorbed and heats up the mask. A low thermal expansion is required in order to avoid any magnification correction between the changing of a wafer, and also to minimize image placement distortion due to thermal expansion of the mask. With respect to multilayer optics, the roughness specification is divided into high spatial frequency roughness (HSFR) and mid-spatial frequency roughness (MSFR). HSFR ($\lambda_{spatial} < 1\,\mu m$) should be

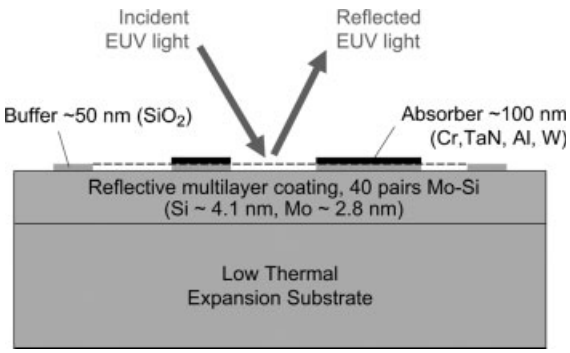


Figure 6.14 A cross-section of a mask to be used in EUV lithography.

below 0.10–0.15 nm (RMS) in order to reduce the losses due to scattering of light out of the entrance pupil of the optical system. In order to reduce the small angle scattering and image speckles, MSFR (1 μm $< \lambda_{spatial} <$ 10 μm) should also be below 0.1–0.2 nm (RMS). A peak reflectivity of more than 67% with a centroid wavelength uniformity across the mask of below 0.03 nm is required.

Another specification refers to the defect density of below 0.003 defects cm^{-2} for defects larger than 30 nm. This is the most challenging demand for the masks, and is one of the most critical issues in EUV lithography. As EUV light has a strong interaction with matter, and thus a short penetration depth of typically <100 nm, defects on the masks have a much higher probability of being printed, in contrast to other wavelengths as in the UV region. Thus, special care must be taken to reduce the defects on the masks to a level of 0.003 per cm^2 or, in other words, to less than a few defects per mask.

Many different categories of defect have been defined, and many activities are required simply to reduce the number on masks, mask blanks and the substrate by cleaning and repair techniques [60], detecting printable defects [61] and simulation of their influence on the picture at the wafer level [62]. Although defects on the substrate will be buried after the multilayer coating is applied, the various types of defect may lead to phase errors of the reflected light. Once such defect has been localized the absorber structure can be appropriately aligned to cover these defect and thus reduce its influence. The influence of defects (particles) on the mask depends on their size. If they are sufficiently small, they are not seen in the de-magnified image on the wafer; hence, only defects larger than 20–30 nm are of interest. The influence of defect size on image is discussed in Ref. [62].

In order to obtain an impression of the current status, the defect densities achieved on mask blanks are taken from Ref. [1]. For defects larger than 120 nm the density is 0.03 defects cm^{-2}, while for >60 nm a density of 0.3 defects cm^{-2} is achieved, this being more than two orders away from the final specification. As yet, no appropriate metrology is available for smaller defects.

6.2.5 Resist

The use of higher photon energies and the printing of ever-smaller features requires the development of a new generation of photo resists compared to those currently used in DUV lithography. According to the International Roadmap for Semiconductors (ITRS), a resist thickness of between 40 and 80 nm is required for EUV, while the line edge roughness (LER) and the critical dimension control (resolution) should be below 1 nm (3σ) [63, 64]. It should be noted that these specifications have near-atomic-scale resolution, which is not achievable with sufficiently high sensitivity when using the resists and concepts currently available. To ensure a certain wafer throughput, the resist sensitivity – that is, the number of photons or energy per unit area required to convert the resist molecules into solvable components – should be in the range of a few mJ cm^{-2}. The required thickness of less than 80 nm is lower than for DUV resists because of the short absorption length for EUV radiation in

polymers. To ensure an approximately homogeneous illumination as a function of the penetration depth, the permissible resist thickness is limited to these values below 100 nm. This also implies a loss of usable photons by having a rather large portion of transmitted light.

There is a trade-off between high sensitivity, low LER and high resolution, which means that improving one feature will lead to a worsening of the other features. This fact, and the special challenges involved in applications for EUV lithography, are discussed in more detail below.

Conventional lithography makes use of a chemically amplified resist (CAR). An incident photon releases a H^+ ion (acid), which serves as a catalyst to react with other molecules to form solvable components, volatile products and another acid to trigger this reaction again. This mechanism is used to increase the sensitivity of the resist, but has an impact on the achievable resolution and LER. The processes are shown schematically in Figure 6.15. The amplification of soluble production by the acids is accompanied by a diffusion process into the unexposed regions. This diffusion process takes place during the post-exposure baking process, as indicated in Figure 6.15, and is determined by the diffusion constant, D, and the duration of the process, t_f. The higher the diffusion, the more sensitive is the resist, but the achievable resolution decreases. In order to quantify this dependence, a one-dimensional exposure with a sinusoidal modulation of photo acids with pitch, p, is considered for simplicity. A measure of the achievable resolution is the modulation transfer function (MTF_{diff}) of this initial distribution altered by the diffusion process [65, 66]:

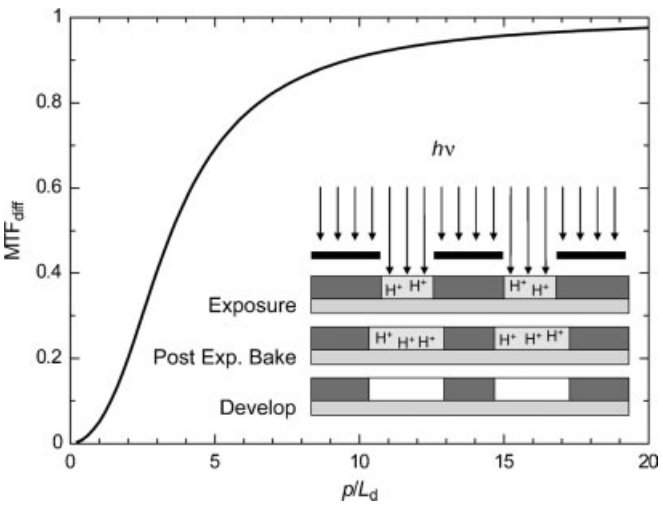


Figure 6.15 Modulation transfer function (MTF) for a sinusoidal exposure of lines and spaces with pitch, p, as a function of diffusion length and a schematic drawing of the exposure, post-exposure baking and developing process of a chemically amplified resist.

$$\mathrm{MTF}_{\mathrm{diff}} = \frac{p^2}{4\pi^2 D t_{\mathrm{f}}} \left\{ 1 - e^{\frac{4\pi^2 D t_{\mathrm{f}}}{p^2}} \right\} \tag{6.14}$$

The respective diffusion length, L_{d}, is defined by ${L_{\mathrm{d}}}^2 = 2Dt_{\mathrm{f}}$. The modulation transfer function is shown as a function of the ratio p/L_{d} in Figure 6.15, where $\mathrm{MTF}_{\mathrm{diff}} = 1$ implies no change of the initial distribution. For example, if a deterioration to 70% is accepted, the diffusion length should not exceed a value of $0.2p$. This limitation of the diffusion length also implies a limitation in resist sensitivity. It should be noted that, with decreasing pitch, the absolute value of the diffusion length must also decrease, which therefore means less sensitivity in the transition from DUV to EUV.

Another parameter describing the quality of a resist is the LER, which is distinct from the achievable resolution in terms of the above discussion. Generally, the LER is dependent on the spatial distribution of solvable components after exposure with a number density, A. The LER is given by the ratio of the standard deviation of this density and its gradient, leading to $\mathrm{LER} \propto \sigma_{\mathrm{A}}/\nabla A$. For a sufficiently low number of photons, both parameters are determined by the incident number of photons. Here, a variation due to the Poisson statistic (shot noise) of the absorbed photons and, in the case of CA resist, the number of produced acids also comes into play. In general, the standard deviation, σ_{N}, of the number of photons, N, in a certain volume is proportional to $\sqrt{N}$, while $A \propto N$. Consequently, the LER scales as $\mathrm{LER} \propto 1/\sqrt{N}$ or $1/\sqrt{E}$, where E is the incident dose. For a chemically amplified resist, the volume – which is relevant to the estimation of the number of photons – is the diffusion sphere. Thus, for a low diffusion length the LER is proportional to ${L_{\mathrm{d}}}^{-3/2}$ when the dose is kept constant. With increasing diffusion length and lower variation due to photon statistics, the diffusion process and the MTF become dominant. The scaling of LER with the diffusion length can be expressed [66]:

$$\mathrm{LER} \propto \left(\frac{1}{L_{\mathrm{d}}}\right)^{3/2} / \mathrm{MTF}_{\mathrm{diff}}\left(\frac{L_{\mathrm{d}}}{p}\right) \tag{6.15}$$

The LER scaling factor according to Equation 6.15 is shown in Figure 6.16 as a function of the diffusion length relative to the pitch. Two regions can be distinguished: for $L_{\mathrm{d}}/p < 0.33$, the scaling is dominated by the photon statistics, whereas for $L_{\mathrm{d}}/p > 0.33$ region the acid diffusion process is relevant for the LER. These two scaling regions are also observed experimentally [66].

The absolute value of LER is still dependent on the resist sensitivity or the necessary dose, which leads to the scaling with $1/\sqrt{E}$. This is illustrated in Figure 6.17, which shows the LER achieved for resists of different sensitivity [67]. The estimated shot noise limit is also indicated. The theoretical limit has clearly not yet been achieved, as the experimental data are slightly higher compared to this limit.

A number of other parameters, such as resist thickness, molecular size or outgassing, are also relevant to use in EUV lithography (see discussion in Ref. [65]). In summary, it is somewhat challenging to meet the specifications for a chemically amplified resist for use in EUV lithography, and in fact such a resist does not yet exist. In terms of LER and resolution, the specifications may be achieved with a

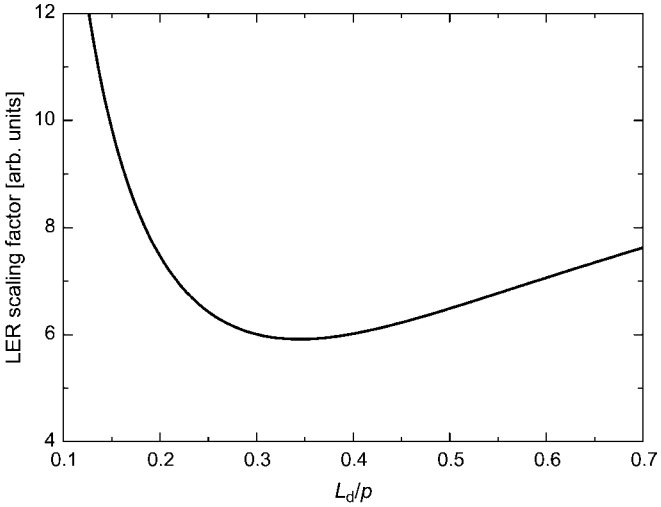


Figure 6.16 Line edge roughness (LER) of a resist with fixed sensitivity as a function of the diffusion length relative to the pitch.

non-chemically amplifying resist. One example is polymethyl methacrylate (PMMA), although this has a rather low sensitivity (in the range of 50–100 mJ cm^{-2}), and is therefore not acceptable for EUV lithography. Some printed lines down to 17.5 nm half pitch [68] are illustrated in Figure 6.18; the smallest currently achieved structure size is a half pitch of 12.5 nm [69].

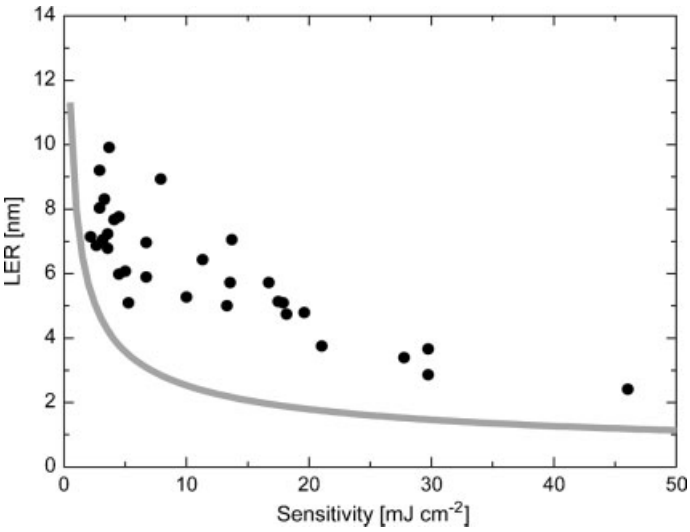


Figure 6.17 Experimentally determined line edge roughness (LER) as a function of sensitivity for different resists. The line gives an estimation of the shot noise limit, which has not yet been achieved.

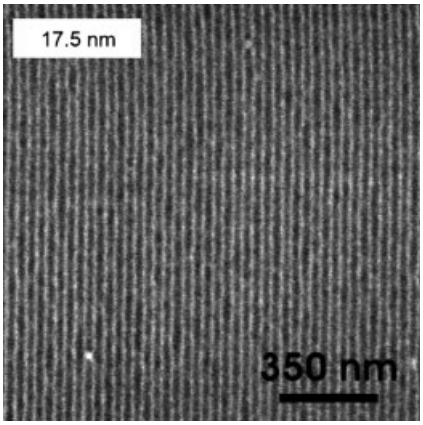


Figure 6.18 Printed lines and spaces with a non-chemically amplified resist (polymethyl methacrylate), with a half-pitch down of 17.5 nm.

Finally, it is illustrative to discuss the role of shot noise simply by estimating the number of photons involved in the exposure process. The incident number of photons per unit area, I, can be rewritten in terms of the necessary dose, E, and the wavelength, λ, of the photons:

$$I = 5.0 \times 10^{-2} \frac{N_{\mathrm{Ph}}}{\mathrm{nm}^2} \frac{\lambda}{\mathrm{nm}} E \frac{\mathrm{cm}^2}{\mathrm{mJ}} \tag{6.16}$$

In the transition from DUV with 193 nm to EUV at 13.5 nm, the reduced wavelength alone leads to a more serious influence of the photon statistics. For EUV radiation and an envisioned resist sensitivity of $E = 5\,\mathrm{mJ\,cm^{-2}}$, we obtain $3.4\,\mathrm{Ph\,nm^{-2}}$. With regards to the specifications of LER and CD control below 1 nm, it is clear that shot noise becomes a limiting factor in EUV lithography. This not only requires the development of new resist materials, but also implies that many research investigations will be necessary over the next few years.

6.3 Outlook

Intensive research and development activities conducted during the past decade have shown that EUV lithography has the potential to provide a solution for the high-volume manufacture of semiconductor devices. Moreover, the technique has the potential to decrease structure sizes to 11 nm, into the range of the physical limits of silicon-based semiconductor technology. Several machines have been installed to demonstrate the capability of printing small structures using EUV radiation; most notably, the ASML alpha demo tool exhibits the full architecture with respect to the optical system, wafer and mask handling for a scanning operation and full-field imaging. The diffraction-limited printing of small structures down to 29 nm was also

successfully demonstrated, though major efforts are still needed to meet the requirements of the components and to drive the technology to its theoretical limits in different areas. These challenges extend not only to the source power but also to the simultaneous high reliability and long lifetime of the source, and this is valid for both laser-induced and discharge-based plasmas. Further issues here include debris mitigation in order to increase the collector lifetime, collector thermal issues, and increase the opening angle. Additional studies are also required on the lifetime and contamination of the optical system by oxygen and hydrocarbons under EUV radiation, on defect-free masks, and on resists with a sufficiently high sensitivity at high resolution and low LER.

Despite the final specifications not having yet been met for several components, progress is nonetheless being made in all fields. For example, activities during the past few years have led to plasma sources which emit inband radiation into the hemisphere on a power level of several hundred watts – close to final specification that in the past was believed to be the most critical issue in EUVL. Major progress is also being achieved in improving the lifetime of both the source and the collector, using sophisticated debris mitigation techniques. In fact, a collector lifetime of more than 1 Gshot has recently been demonstrated, operating with a tin-emitting plasma source, and an ultimate lifetime of 100 Gshot seems feasible.

References

1 van den Brink, M. (2006) The only cost effective extendable lithography option: EUV, Third International Symposium on EUV Lithography, Barcelona.

2 *International Technology Roadmap for semiconductors (ITRS)*. The current version is available from www.sematech.org or www.itrs.net.

3 Ceglio, N.M., Hawryluk, A.M. and Sommargren, G.E. (1993) Front-end design issues in soft-X-ray projection lithography. *Applied Optics*, **32** (34), 7050–7056.

4 Gwyn, C.W., Stulen, R., Sweeney, D. and Attwood, D. (1998) Extreme ultraviolet lithography. *Journal of Vacuum Science & Technology B*, **16** (6), 3142–3149.

5 Ota, K., Watanabe, Y., Banine, V. and Franken, H. (2006) EUV source requirements for EUV lithography, in *EUV Sources for Lithography* (ed. V. Bakshi), SPIE Press, Bellingham, Washington, pp. 27–43.

6 Meiling, H. Meijer, H., Banine, V., Moors, R., Groeneveld, R., Voorma, H.-J., Mickan, U., Wolschrijn, B., Mertens, R., van Baars, G., Kürz, P., Harned, N., (2006) First performance results of the ASML alpha demo tool, in Emerging Lithographic Technologies X, Proceedings of SPIE, Vol. 6151, San Jose, USA (ed. Lercel, M.J.), pp. 615108.

7 Booth, M., Brioso, O., Brunton, A., Cashmore, J., Elbourn, P., Elliner, G., Gower, M., Greuters, J., Grünewald, P., Gutierrez, R., Hill, T., Hirsch, J., Kling, L., McEntee, N., Mundair, S., Richards, P., Truffert, V., Wallhead, I., Whitfield, M. and Hudyma, R. (2005) High-resolution EUV imaging tools for resist exposure and aerial image monitoring. Proc. SPIE 5751, 78–89.

8 Groeneveld, R., Harned, N., Zimmermann, J., Meijer, H., Meiling, H., Mickan, U., Voorma, H.J. and Kuerz, P. (2006) Full Field Imaging by the ASML

Alpha Demo Tool, International EUVL Symposium, Barcelona, Spain, Proceedings available at www.sematech.org.

9 Derra, G., and Singer, W. (2003) Collection efficiency of EUV sources. Proc. SPIE 5037, 728–741.

10 Ota, K., Tanaka, K. and Kondo, H. (2003) Throughput model considerations and impact of throughput improvement request on exposure tool, Second International EUVL Symposium, Antwerp, Belgium. Proceedings available from www.sematech.org.

11 Bakshi, V. (2006) *EUV Sources for Lithography*, SPIE Press, Bellingham, Washington.

12 Rigato, V. (2006) Evolution from current demonstrated α-hardware collector to full HVM, EUV Source Workshop, Barcelona, Spain. Proceedings available from www.sematech.org.

13 For example: Krücken, T., Bergmann, K., Juschkin, L. and Lebert, R. (2004) Fundamentals and limits for the EUV emission of pinch plasma sources for EUV lithography. *Journal of Physics D-Applied Physics*, **37** (23), 3213–3224.

14 Puell, H. (1970) Heating of laser produced plasmas generated at plane solid targets. *Zeitschrift fur Naturforschung*, **A25**, 1807–1815.

15 Wood, O.R., II, Silvfast, W., Macklin, J. and Maloney, P. (1986) Comparison of extreme-ultraviolet flux from 1.06- and 10.6-μm laser-produced plasma sources for pumping photoionization lasers. *Optics Letters*, **11**, 198–200.

16 Schriever, G., Mager, S., Naweed, A., Engel, A., Bergmann, K. and Lebert, R. (1998) Laser-produced lithium plasma as a narrow-band extended ultraviolet radiation source for photoelectron spectroscopy. *Applied Optics*, **37** (7), 1243–1248.

17 Spitzer, R., Orzechowski, T., Phillion, D., Kauffman, R. and Cerjan, C. (1996) Conversion efficiencies from laser produced plasmas in the extreme ultraviolet regime. *Journal of Applied Physiology*, **79**, 2251–2258.

18 Richardson, M., Koay, C.-S., Takenoshita, K., Keyser, Ch., George, S., Al-Rabban, M. and Bakshi, V. (2006) Laser plasma EUV sources based on droplet target technology, in *EUV Sources for Lithography* (ed. Bakshi, V.), SPIE Press, Bellingham, Washington, pp. 687–718.

19 Takahashi, A., Tanaka, H., Akinaga, K., Matsumoto, A., Uchino, K. and Okada, T. (2005) Laser-wavelength dependence of laser produced plasma EUV emission, Third International Symposium on EUV Lithography, November 2004, Miyazaki, Japan. Proceedings available at: www.sematech.org.

20 Stamm, U., and Gäbel, K. (2006) Technology for LPP sources, in *EUV Sources for Lithography* (ed. V. Bakshi), SPIE Press, Bellingham, Washington, pp. 537–561.

21 (a) Eidmann, K., and Schwanda, W. (1991) *Laser Particle Beams*, **9**, 551.(b) Sigel, R. (1989) *Proceedings of SPIE*, **1140**, 6. (c) Sigel, R., Eidmann, K., Lavarenne, F. and Schmalz, R.F. (1990) *Physics of Fluids B*, **2**, 199.(d) Eidmann, K., Kühne, M., Müller, P. and Tsakiris, G.D. (1990) *Physics of Fluids B*, **2**, 208.

22 Hertz, H.M., Rymell, L., Berglund, M. and Malmqvist, L. (1996) Debris-free liquid-target laser-plasma soft X-ray source for microscopy and lithography, in *X-ray Microscopy and Spectromicroscopy* (eds J. Thieme, G. Schmahl, E. Umbach and D. Rudolph), Springer, Heidelberg.

23 Bakshi, V.(ed.) (2006). Section IV Laser-Produced Plasma (LPP) Sources, in *EUV Sources for Lithography*, SPIE Press, Bellingham, Washington, pp. 535–718.

24 Hansson, B.A.M., and Hertz, H.M. (2006) Liquid-xenon-jet LPP source, in *EUV Sources for Lithography* (ed. V. Bakshi), SPIE Press, Bellingham, Washington, pp. 619–648.

25 For example: Krall, N.A., and Trivelpiece, A.W. (1986) *Principles of Plasma Physics*, San Francisco Press, New York.

26 Bergmann, K., Lebert, R. and Neff, W. (1997) Scaling of the K-shell line

emission in transient pinch plasmas, *Journal of Physics D-Applied Physics*, **30** (6), 990.

27 Lercel, M.J. (ed.), Emerging Lithographic Technologies X, in Proceedings of SPIE, Vol. 6151, San Jose, USA, February 2006.

28 Juschkin, L., Derra, G. and Bergmann, K. (2007) EUV light sources, in *Low-Temperature Plasma Physics* (ed. Hippler, R.), Springer Verlag, pp. 619–654.

29 (2004) Special cluster on extreme ultraviolet light sources for semiconductor manufacturing. *Journal of Physics D-Applied Physics*, **37** (23), 3207–3284.

30 Bergmann, K., Schriever, G., Rosier, O., Müller, M., Neff, W. and Lebert, R. (1999) Highly repetitive, extreme-ultraviolet radiation source based on a gas-discharge plasma. *Applied Optics*, **38**, 5413–5417.

31 Pankert, J. Bergmann, K.,Klein, J., Neff, W., Rosier, O., Seiwert, S., Smith, C., Apetz, R., Jonkers, J., Loeken, M., Derra, G., (2002) Physical properties of the HCT EUV source, SPIE 27th International Symposium on Microlithography, Santa Clara, USA, 3–8 March.

32 Fomenkov, I.V., Partlo, W.N., Böwering, N.R., Khodykin, O.V., Rettig, C.L., Ness, R.N., Hoffman, J.R., Oliver, I.R. and Melnychuk, S.T. (2006) Dense plasma focus source, in *EUV Sources for Lithography* (ed. V. Bakshi) SPIE Press, Bellingham, Washington, pp. 373–394.

33 Stamm, U., Kleinschmidt, J., Bolshukin, D., Brudermann, J., Hergenhan, G., Korobotchko, V., Nikolaus, B., Schürmann, M.C., Schriever, G., Ziener, C. and Borisov, V.M. (2006) Development status of EUV sources for use in beta-tools and high volume chip manufacturing, in Proceedings of SPIE, Vol. 6151, San Jose, USA. (ed. M.J. Lercel) p. 61510O.

34 Teramoto, Y., Sato, H. and Yoshioka, M. (2006) Capillary Z-pinch source, in *EUV Sources for Lithography* (ed. Bakshi, V.) SPIE Press, Bellingham, Washington, pp. 505–522.

35 McGeoch, M. (1998) Radio-frequency preionized xenon z-pinch source for extreme ultraviolet lithography. *Applied Optics*, **37**, 1651.

36 McGeoch, W. (2006) Star pinch EUV source, in *EUV Sources for Lithography* (ed. V. Bakshi) SPIE Press, Bellingham, Washington, pp. 453–476.

37 Endo, A. (2006) Driver Laser, Xenon Target, and System Development for LPP Sources, in *EUV Sources for Lithography* (ed. V. Bakshi), SPIE Press, Bellingham, Washington, pp. 607–618.

38 Stuik, R., Fledderus, H., Bijkerk, F., Hegeman, P., Jonkers, J., Visser, M., Banine, V., Flying Circus EUV Source Comparison, Second International EUVL Workshop, San Francisco, 19–20 October 2000, Presentation available from www.sematech.org.

39 Pankert, J., Apetz, R., Bergmann, K., Damen, M., Derra, G., Franken, O., Janssen, M., Jonkers, J., Klein, J., Kraus, H., Krücken, T., List, A., Loeken, M., Mader, A., Metzmacher, C., Neff, W., Probst, S., Prümmer, R., Rosier, O., Schwabe, S., Seiwert, S., Siemons, G., Vaudrevange, D., Wagemann, D., Weber, A., Zink, P. and Zitzen, O. (2006) EUV sources for the alpha-tools, in Proceedings of SPIE, Vol. 6151, San Jose, USA. (ed. M.J. Lercel) p. 61510Q.

40 Corthout, M., Bergmann, K., Derra, G., Jonkers, J., Pankert, J. and Zink, P. (2006) The Philips Extreme UV Sn Source: Recent progress in power, lifetime and collector lifetime, International EUVL Symposium, Barcelona, Spain. Proceedings available at www.sematech.org.

41 Thompson, P.L. and Harvey, J.E. (2000) Systems engineering analysis of aplanatic Wolter type X-ray telescopes. *Optical Engineering*, **39** (6), 1677–1691.

42 Center for X-Ray Optics, Berkeley, USA, www-cxro.lbl.gov.

43 Zocchi, F.E., Bianucci, G., Rigato, V., Pirovano, G., Cassol, G.L., Salmaso, G.,

Bind, P., Zink, P., Bergmann, K., Nikolaus, B. and Schürmann, M.C. (2006) Experimental validation of collector's thermo-optical design, EUV Source Workshop, Barcelona, Spain. Proceedings available from www.sematech.org.

44 Kortright, B. and Underwood, J.H. (1991) Design considerations for multilayer coated Schwarzschild objectives for the XUV, Proceedings of SPIE, Vol. 1343, X-ray/EUV optics for astronomy, microscopy, polarimetry, and projection lithography, pp. 95–103.

45 Artioukov, I.A., and Krymski, K.M. (2000) Schwarzschild objective for soft X-rays. *Optical Engineering*, **39** (8), 2163–2170.

46 Geyl, R. (2006) Near normal incidence collectors for easier debris mitigation, EUV Source Workshop, Barcelona, Spain. Proceedings available from www.sematech.org.

47 Klebanoff, L.E., Anderson, R.A., Buchenauer, D.A., Fornaciari, N.R. and Kimori, H. (2006) Erosion of condensor optics exposed to EUV sources, in *EUV Sources for Lithography* (ed. V. Bakshi) SPIE Press, Bellingham, Washington, pp. 995–1031.

48 Ruzic, D.N. (2006) Origin of debris in EUV sources and its mitigation, in *EUV Sources for Lithography* (ed. V. Bakshi), SPIE Press, Bellingham, Washington, pp. 957–993.

49 Pankert, J., Apetz, R., Bergmann, K., Derra, G., Janssen, M., Jonkers, J., Klein, J., Krücken, T., List, A., Loeken, M., Metzmacher, C., Neff, W., Probst, S., Prümmer, R., Rosier, O., Seiwert, S., Siemons, G., Vaudrevange, D., Wagemann, D., Weber, A., Zink, P. and Zitzen, O. (2005) Integrating Philips' extreme UV source in the alpha-tools. Proceedings of SPIE 5751, 260–271.

50 Böttger, T., Meyer, D.C., Paufler, P., Braun, S., Moss, M., Mai, H. and Beyer, E. (2003) Thermal stability of Mo/Si multilayers with boron carbide interlayers. *Thin Solid Films*, **444**, 165–173.

51 Rigato, V., Mattarello, V., Nannarone, S. and Borgatti, F. (2005) Thermal stability of Mo/Si multilayers with ruthenium interlayers, International EUVL Symposium, San Diego, CA USA. Proceedings available from www.sematech.org.

52 Bajt, S., Dai, Z.R., Nelson, E.J., Wall, M.A., Alamenda, J.B., Nguyen, N.Q., Baker, S.L., Robinson, J.C. and Taylor, J.S. (2006) Oxidation resistance and microstructure of ruthenium-capped extreme ultraviolet multilayers. *Journal of Microlithography, Microfabrication, and Microsystems*, **5** (2), 023004.

53 Van de Kruijs, R.W.E., Yakshin, A.E., van Herpen, M.M.J.W., Klunder, D.J.W., Louis, E., Alonso van der Westen, S., Enkisch, H., Müllender, S., Bakker, L., Banine, V. and Bijkerk, F. (2006) Multilayer optics with spectral purity layers for the EUV wavelength range, Conference on Physics of X-Ray Multilayer Structures, Sapporo, Japan.

54 Sweeney, D.W., Hudyma, R., Chapman, H.N. and Shafer, D. (1998) EZV optical design for a 100 nm CD imaging system, in Proceedings of the SPIE 23rd Annual International Symposium on Microlithography, Santa Clara, CA, USA, February 22–27.

55 Montcalm, C., Grabner, R.F., Hudyma, R.M., Schmidt, M.A., Spiller, E., Walton, C.C., Wedowski, M. and Folta, J.A. (1999) Multilayer coated optics for an alpha-class extreme ultraviolet lithography system, in Proceedings of the 44th Annual Meeting of the International Symposium on Optical Science, Engineering and Instrumentation, Denver, CO, USA, July 18–23.

56 Stover, J.C. (1995) *Optical Scattering: Measurement and Analysis*, SPIE.

57 Kuerz, P., Böhm, T., Müllender, S., Bollinger, W., Dahl, M., Lowisch, M., Münster, C., Rohmund, F., Stein, T., Louis, E. and Bijkerk, F. (2005) Optics for EUV lithography, International EUVL Symposium, San Diego, CA, USA. Proceedings available from www.sematech.org.

58 Vernon, S.P., Kerney, P.A., Tong, W., Prisbrey, S., Larson, C., Moore, C.E.,

Weber, F., Cardinale, G., Yan, P.Y. and Hector, S. (1998) Masks for extreme ultraviolet lithography, Proceedings, 18th Annual BACUS Symposium on Photomask Technology and Management, Redwood City, CA, USA, September 16–18.

59 Tong, W. (1999) EUVL Mask Substrate Specifications (wafer type), UCRL-ID-135579, Report of U.S. Department of Energy, Lawrence Livermore National Laboratory.

60 Yu, Y.S., Kim, T.G., Lee, S.H., Park, J.G., Kim, T.H., Busnaina, A. and Lee, J.M. (2006) Removal of nano particles on EUV mask buffer and absorber layers by laser shockwave cleaning, International EUVL Symposium, Barcelona, Spain. Proceedings available from www.sematech.org.

61 Barty, A., Liu, Y., Gullikson, E., Taylor, J.S. and Wood, O. (2005) Actinic inspection of multilayer defects on EUV masks, Proceedings of the SPIE Microlithography, San Jose, CA, USA, April 2–3.

62 Lin, Y., and Bokor, J. (1997) Minimum critical defects in extreme-ultraviolet lithography masks. *Journal of Vacuum Science & Technology B*, **15** (6), 2467–2470.

63 Leeson, M., Cao, H., Yueh, W., Meagley, R., Sharma, G. and Sharma, S. (2006) EUV Resist Materials, Properties and Performance, International EUVL Symposium, Barcelona, Spain. Proceedings available from www.sematech.org.

64 Oizumi, H., Tanak, Y., Kumise, T., Nishiyama, I., Shiono, D., Hirayama, T., Hada, H., Onodera, J. and Yamaguchi, A. (2006) Performance of new molecular resist in EUV lithography, International EUVL Symposium, Barcelona, Spain. Proceedings available from www.sematech.org.

65 Okoroanyanwu, U. and Lammers, J.H. (2004) Resist Road to the 22 nm Technology Node. *Future Fab International*, **17**, Available at www.future-fab.com.

66 Zandbergen, P., Domke, W.D., Cantu, P., Thony, P., Postnikov, S. and Robic, J.Y. (2005) EXCITE, the MEDEA + Extreme UV Consortium for Imaging Technology, International EUVL Symposium, San Diego, CA, USA. Proceedings available from www.sematech.org.

67 Nalleau, P., Rammeloo, C., Cain, J.P., Dean, K., Denham, P., Goldberg, K.A., Hoef, B., La Fontaine, B., Pawloski, A., Larson, C. and Wallraf, G. (2005) Investigation of the current resolution limits of advanced EUV resists, International EUVL Symposium, San Diego, CA, USA. Proceedings available from www.sematech.org.

68 Solak, H.H., He, D., Li, W. and Cerrina, F. (1999) Nanolithography using extreme ultraviolet lithography interferometry: 19 nm lines and spaces. *Journal of Vacuum Science & Technology B*, **17** (6), 3052–3057.

69 Solak, H.H., Ekinci, Y., Käser, P. and Park, S. (2007) Photon-beam lithography reaches 12.5 nm half pitch resolution. *Journal of Vacuum Science & Technology B*, **25** (1), 91–95.

7
Non-Optical Lithography

Clivia M. Sotomayor Torres and Jouni Ahopelto

7.1
Introduction

In the quest to use nanofabrication methods to exploit the know-how and potentials of nanotechnology, one major roadblock is the high cost factor which characterizes high-resolution fabrication technologies such as electron beam lithography (EBL) and extreme ultraviolet (EUV) lithography. The need to circumvent these problems of cost has inspired research and development in alternative nanofabrication, also referred to as "emerging" or "bottom-up" approaches. Hence, it is within this context that the status and prospects of nanoimprint lithography (NIL) are presented in this chapter.

Nanofabrication needs are highly diverse, not only in the materials used but also in the range of applications. Within the physical sciences, the drive is to realize nanostructures in order to produce artificial electronic, photonic, plasmonic or phononic crystals. This, in turn, depends on an ability to realize periodic or quasi-periodic arrays of nanostructures, on the one hand to meet the stringent demands of periodicity, order and critical dimensions to obtain the desired dispersion relation and, on the other hand, to identify a reproducible, cost-effective and reliable way in which such materials may be fabricated, using a suitable form of nanopatterning.

Nanopatterning covers a wide range of methods from top-down approaches, as well as bottom-up approaches (for discussions, see Chapters 5, 6, 8 and 9 of this volume). In fact, the needs for lithography are found in several fields:

- In nano-CMOS (complementary metal-oxide semiconductor) for example, to produce pattern gates of lengths down to a few nanometers in order to reach the technology nodes of the semiconductor industry roadmap [1], whilst at the same time complying with the most strict lithography demands.
- In (nano)photonics, a field in which – in addition to packaging – the cost of fabrication of III-V semiconductor optoelectronic devices containing nanostructures in the form of photonic crystals, is prohibitive.

Nanotechnology. Volume 3: Information Technology I. Edited by Rainer Waser

ISBN: 978-3-527-31738-7

- In nanobiotechnology, to fabricate a variety of sensors and lab-on-a-chip platforms based on micro- and nano-fluidics.
- In organic opto- and nano-electronics, where the lifetime issue of the organic materials is compounded with that of a cost-effective volume production with lateral resolution down to a few hundreds of nanometers for electrodes and pixels.
- In micro electro-mechanical systems (MEMS) and nano electro-mechanical systems (NEMS), where the fabrication of resonators, cantilevers and many other structures with and without direct interface to Si-based electronics, requires the control of 3-D nanofabrication with minimum damage to the underlying electronic platform.

Moreover, nanopatterning methods act as enabling technologies to facilitate the progress of research in chemistry, such as the realization of nanoelectrodes to monitor electric activity; in biology, to connect electrically to cells; in physics, to realize nanostructures commensurate with the De Broglie wavelength of a given excitation; in material sciences, through research on novel nanostructured artificial materials; and also in several other engineering disciplines.

In 2003, the state of the art covering most bottom-up emerging nanopatterning methods was collected in Ref. [2] under the title *Alternative Lithography: unleashing the potentials of Nanotechnology*. Of these methods, probably the most advanced is polymer molding or nanoimprint lithography [3]. Other emerging bottom-up methods are those based on scanning probes [4, 5], self assembly (see Chapters 9 and 10 in this volume), micro-contact printing or soft-lithography [6, 7] and stenciling [8], as well as atom lithography [9] and bio-inspired lithography [10].

In this chapter, attention is focused primarily on NIL as an example of non-optical lithographies, as it covers the 1 μm to few nanometer lateral resolution range. Here, the basic principles of this method are described, the state of the art is reviewed, and the main scientific and engineering issues are addressed.

7.2 Nanoimprint Lithography

7.2.1 The Nanoimprint Process

Historically, NIL has been preceded by some remarkable events. In the twelfth century, metal type printing techniques were developed in Korea; for example, in 1234 the "Kogumsangjong-yemun" (Prescribed Ritual Text of Past and Present) appeared, while in 1450 Gutenberg introduced his press and printed 300 issues of the two-volume Bible. Somewhat strangely, an extensive time lapse then occurred until the early twentieth century, when the first vinyl records were produced by using hot embossing [11]. The next major step occurred during the 1970s, when compact discs were fabricated by injection molding.

The term "nanoimprint lithography" was most likely used for the first time by Stephen Y. Chou, when referring to patterning of the surface of a polymer film or resist with lateral feature sizes below 10 nm [3]. Previously, within a larger lateral size range, the method was referred to as "hot embossing". At about the same time, Jan Haisma reported the molding of a monomer in a vacuum contact printer and subsequent curing by UV radiation, which was known as "mold-assisted lithography" [12]. The first comprehensive review of these two approaches appeared in 2000 [13], but since then several excellent reviews of NIL have been produced [14, 15]. During recent years these methods have developed further and have become to be known as "thermal nanoimprint lithography" and "UV-nanoimprint lithography" (UV-NIL), respectively.

The question must be asked, however, what is NIL? Nanoimprint lithography is basically a polymer surface-structuring method which functions by making a polymer flow into the recesses of a hard stamp in a cycle involving temperature and pressure. In order to nanoimprint a surface, three basic components are required: (i) a stamp with suitable feature sizes; (ii) a material to be printed; and (iii) the equipment for printing with adequate control of temperature, pressure and control of parallelism of the stamp and substrate. The NIL process is illustrated schematically in Figure 7.1. In essence, the process consists of pressing the solid stamp using a pressure in the range of about 50 to 100 bar, against a thin polymer film. This takes place when the polymer is held some 90–100 °C above its glass transition temperature (T_g), in a time scale of few minutes, during which time the polymer can flow to fill in the volume delimited by the surface topology of the stamp. The stamp is detached from the printed substrate after cooling both it and the substrate. The cycle, which is illustrated graphically in Figure 7.2, involves time, temperature, and pressure. Here, we have the main issues of NIL: polymer flow and rheology. Although these points have been addressed from the materials point of view in Refs. [16, 17], they remain a serious challenge for feature sizes below 20 nm. These aspects will be discussed in the following sections.

In UV-NIL the thermal cycle is replaced by curing the molded polymer by UV light through a transparent stamp. This requires different polymer properties, as will be discussed in the next section.

7.2.2 Polymers for Nanoimprint Lithography

The polymers used in NIL play a critical role. A comparison of the 10 most-often used polymers (resists) used in thermal NIL is provided in Ref. [15]. The resists determine both the quality of printing and the throughput. Quality is achieved via the thickness uniformity of the spin-coated film, the strong adhesion to the substrate, and the weak adhesion to the stamp. Throughput is achieved via the duration of the printing cycle, which in turn is determined by several time scales including:

- the time needed to reach the printing temperature (the higher the T_g, the longer the cycle, unless there is a pre-heating stage)

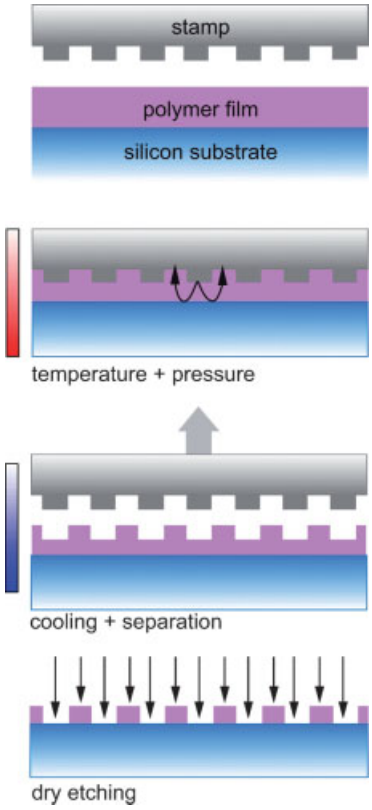


Figure 7.1 Schematics of the thermal nanoimprint concept. Top to bottom: The polymer layer on a solid substrate is heated to a temperature above the glass transition temperature (T_g). The stamp and polymer layer are brought into contact. Pressure is applied to start the polymer flow into the cavities of the stamp. The sample and stamp are cooled down for demolding or separation at a temperature below T_g. The residual polymer layer is removed, typically by dry etching. The end result is a patterned polymer layer on a substrate.

- the hold time for optimum flow at the printing temperature (the more viscous the polymer, the longer the hold time)
- the time needed for cooling and demolding.

Different criteria must be met for thermal NIL and for UV-NIL [18]. For thermal NIL, the polymer is used as a thin film of a few hundred nanometers thickness which is spin-coated onto the support substrate. The key properties are of a thermodynamic nature, and therefore these polymers are of the thermoplastic and thermosetting varieties with varying molecular weights, chemical structures, and rheological and mechanical behaviors. The dependence of the viscosity of a thermoplastic polymer as a function of temperature is shown graphically in Figure 7.3, and illustrates the region where thermal NIL takes place. The mechanical behaviors of polymers in different temperature regimes, in relation to the molecular mobility, are listed in Table 7.1.

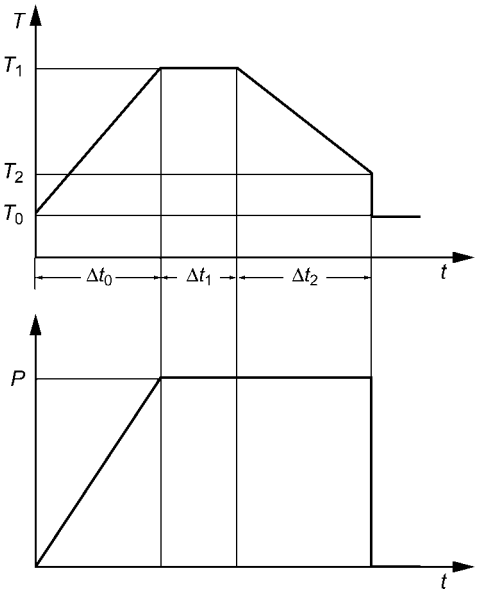


Figure 7.2 Temperature and pressure cycles as a function of time in the thermal NIL process. Typical parameters used for thermoplastic polymers are: Printing temperature T_1 (°C) = 185; demolding temperature T_2 (°C) = 95; printing pressure P (bar) = 30; time to reach printing temperature allowing polymer to go from solid to viscous regime Δt_0(s) = 60; molding time Δt_1 (s) = 60; cooling time Δt_2 (s) = 160.

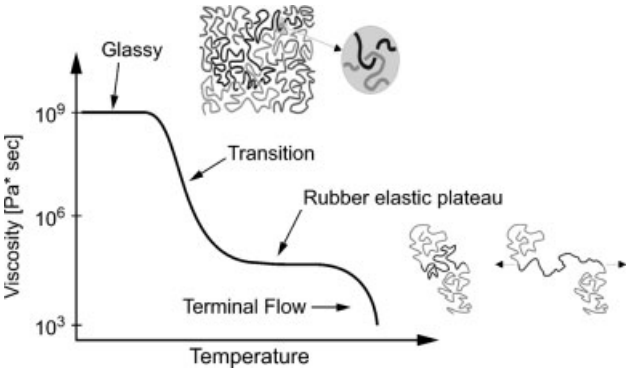


Figure 7.3 Typical dependence of a polymer viscosity on temperature. At room temperature, the polymer is in its solid (glassy, brittle) state. As the temperature increases the short-chain segments become disentangled and the polymer rapidly undergoes a transition from its solid to a rubbery state, changing its viscosity by several orders of magnitude around the glass transition temperature, T_g. Further temperature increases lead to disentanglement of the long polymer chains and resulting in a terminal flow of the viscous melt. Printing takes place in the region where the flow is optimum for filling of the stamp cavities, depending on the molecular weight and stamp design.

Table. 7.1 Relationship between molecular mobility and mechanical behavior of polymers in different temperature regimes.

Temperature regime		$T_{subtransition}$	T_g	T_{flow}
State	Glassy	Rubber elastic	Plastic	
Mechanical appearance	Brittle	Hard elastic, rigid	Rubber elastic	Viscoelastic
Young's modulus ($N\,mm^{-2}$)	about 3000	about 1000	about 1	Too small to measure
Molecular mobility	Molecular conformation completely fixed.	Molecular conformation largely fixed. Occasional change in molecular positions of side groups and chain segments.	Entanglement and physical junction zones prevent movement of entire macromolecules. Entropy-elastic change of molecular position of chain segments. Micro-Brownian motion. Creep, no plastic flow.	No restricted rotation around single bonds. Whole macromolecules change their positions gliding past each other. Plastic flow. Macro-Brownian motion.
Effect of stress	Energy-driven elastic distortions.	Energy-driven elastic distortions.	Entropy-driven elastic distortion. Besides temperature, the deformation rate affects the mechanical behavior.	Pseudoplasticity, shear thinning.
Suitability for printing			Imprinting is possible, but will have memory effects.	Best printing temperature range.

[a] Adapted from Ref. [19]; reproduced with permission.

One of the polymer strategies used to reduce the printing temperature and to improve thermal stability has been to cure prepolymers (special precursors of crosslinked polymers). Here, the term "curing" refers to the photochemical (UV)- or thermal-induced crosslinking of macromolecules to generate a spatial macromolecular network. The prepolymers are low-molecular-weight products, with a low T_g, which are soluble and contain functional groups for further polymerization. Thus, lower printing temperatures of about 100 °C can be used. Curing can take place during the printing time, or thereafter, with the thermal stability enhancement arising from the crosslinking process of the macromolecules.

Polymers for UV-NIL must be suitable for liquid resist processing; that is, they are characterized by a lower viscosity than the polymers for thermal NIL. Naturally, they must also be UV-curable over short time scales [19]. The characteristics of these polymers after printing for their direct use, as in polymer optics or microfluidics, or as a mask for subsequent pattern transfer, by means of dry etching, demand high mechanical, thermal and temporal stability. In photonic applications, stability in

terms of optical properties, such as refractive index, is also essential. Recently, micro resist technology GmbH [20] has developed a whole range of polymers for thermal and UV-NIL (for a discussion, see Ref. [21]). Moreover, tailoring the polymer properties to increase the control of critical dimensions remains an area where, although rapid progress has recently been made [18, 19, 21], further research investigations are still required. The importance of this research may be appreciated especially in a one-to-one filling of the stamp cavities, thereby making the printed polymer features resilient to residual layer removal. Moreover, polymer engineering is also a determinant in larger throughputs, in terms of shorter times for the curing and pressure cycles.

In recent years several reports detailing mechanical studies of thermal NIL have been made, and the interested reader is referred to the data of Hirai [22], the review of Schift and Kristensen [15], and to a recent review of the research on the simple viscous squeeze flow theory [23].

7.2.3 Variations of NIL Methods

To date, four main variations of the NIL process have been developed, and these are briefly described below.

7.2.3.1 Single-Step NIL

This is the most commonly used method to print a polymer in one temperature–pressure cycle, and has been extended to the printing of 150-mm [24] and 200-mm wafers [25]. A scanning electron microscopy (SEM) image of an array of lines of 200 nm width printed over a 200-mm silicon wafer is shown in Figure 7.4. Although one-step thermal NIL can be performed using regular laboratory-scale equipment, commercially available tools include, among others, those of OBDUCAT [26] and EVG [28], which are available in Asia and the Americas. Thermal expansion may cause distortions in the imprinted pattern and to avoid this, strategies for room-temperature NIL have been investigated [28, 29].

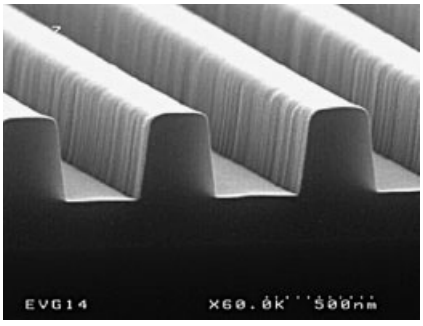


Figure 7.4 A scanning electron microscopy image of 200-nm lines printed in polymer on a 200-mm wafer [25].

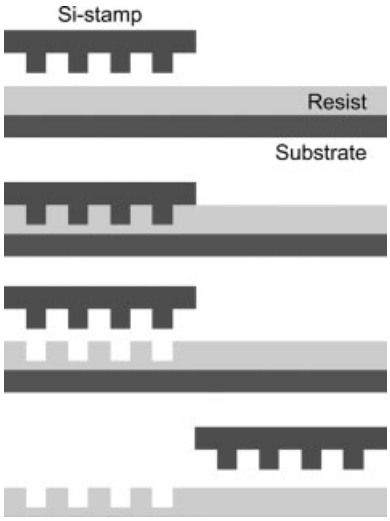


Figure 7.5 The step-and-stamp imprinting lithography process shown schematically. The substrate is patterned in a sequential process by stepping and imprinting across the surface. During the process, the substrate temperature is kept below the glass transition temperature of the resist polymer, while the temperature of the stamp is cycled up and down, above and below the glass transition temperature [30].

7.2.3.2 **Step-and-Stamp Imprint Lithography**

Step-and-stamp imprint lithography (SSIL) is a sequential process, pioneered by Tomi Haatainen and Jouni Ahopelto [30], and is depicted schematically in Figure 7.5. Basically, the system employs thermal NIL and uses a small stamp to print, step and print again, in order to nanostructure the desired area. Initially, SSIL was developed using a commercially available flip-chip bonder, but a dedicated wafer-scale tool from SUSS MicroTec is now available for SSIL [31]. One advantage of SSIL is its capability to achieve a high overlay accuracy, which makes it possible to pattern several consecutive layers or to mix and match with other lithography techniques [32]. An example of a full-patterned wafer is shown in Figure 7.6.

7.2.3.3 **Step-and-Flash Imprint Lithography**

Step-and-flash imprint lithography (SFIL) is also a sequential process, and uses UV radiation instead of temperature to generate relief patterns with line widths below 100 nm. Like NIL, SFIL does not use projection optics but, unlike NIL, it functions at room temperature. SFIL, which is depicted schematically in Figure 7.7, was pioneered by the team of Grant Wilson in the USA [33, 34], with an initial target of meeting the needs of front-end CMOS process fabrication. One of its attractive features is the ability to print over already patterned surfaces. Molecular Imprints Inc. has developed a range of tools for SFIL [35]. As with UV-NIL, SFIL requires transparent stamps

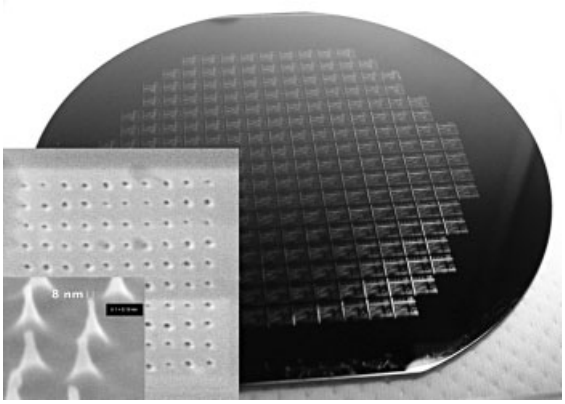


Figure 7.6 A full 100-mm wafer patterned by SSIL, consisting of a matrix of more than 200 imprints into mr-I 7030 resist. The inset shows scanning electron microscopy images of a silicon stamp with sub-10-nm pillars, together with the corresponding imprint.

(usually quartz), the fabrication of which is at present less straightforward than for thermal NIL.

7.2.3.4 **Roll-to-Roll Printing**

This is an advanced sequential method stemming from production method used in, for example, the newspaper industry. Roll-to-roll nanoimprinting is a versatile method that can be combined with other continuous printing techniques, as shown schematically in Figure 7.8. Its extension to 100-nm lateral resolution has been reported [36]. Recent developments suggest that roll-to-roll nanoimprinting is the closest to an industrial technology for organic opto- and nano-electronics, as well as for lab-on-chip device fabrication. The challenge is to fabricate the round stamps; that is, the printing rolls with nanometer-scale features. Moreover, due to the nature of the continuous process, some restrictions may arise in applications requiring multilevel patterning with high alignment accuracy between the layers. Examples of the feature size that can be obtained with a laboratory-scale roll-to-roll printer are shown in Figure 7.9.

7.2.4 **Stamps**

Stamps for NIL have been extensively discussed in Ref. [15]. The main considerations from the materials aspect include:

- Hardness (e.g., typically from 500 to thousands of kg mm^{-2}), which determines the stamp lifetime and the way in which it wears out.

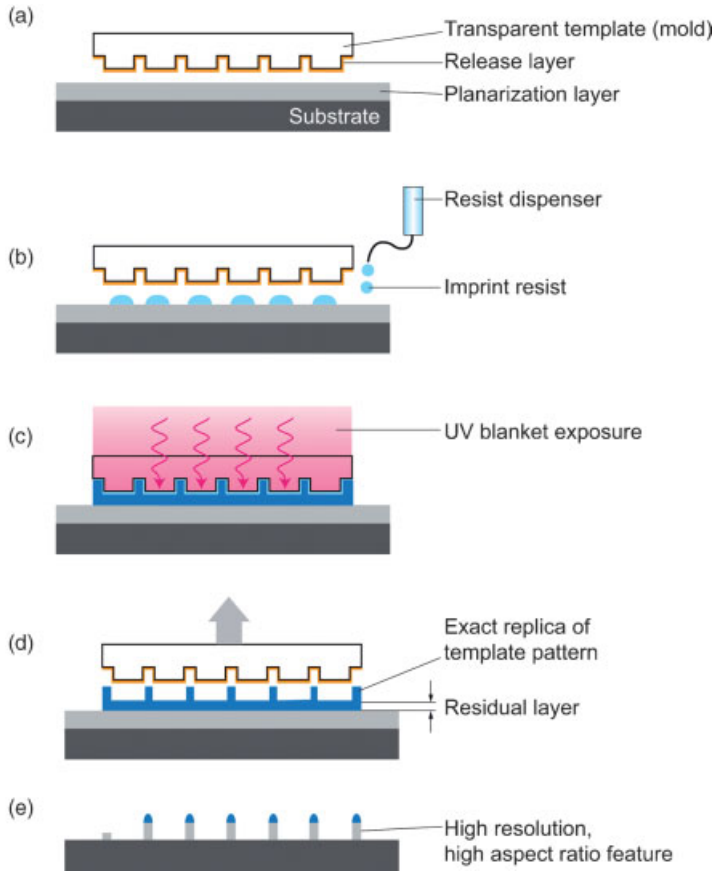


Figure 7.7 Schematics of the step-and-flash imprint lithography (SFIL) process. (a) The pre-planarized substrate and treated stamp are oriented parallel to each other. (b) Drops of UV-curable, low-viscosity imprint resist are dispensed on specified places. (c) The stamp is lowered to fill the patterns and the imprint fluid is polymerized (cured) with UV light at room temperature and low pressure. (d) The stamp is separated from the imprinted substrate. (e) A halogen breakthrough etch to remove the residual layer is performed, followed by an oxygen reactive ion etch [35].

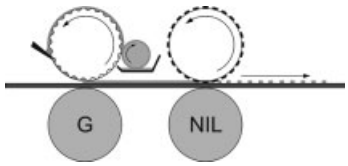


Figure 7.8 The advanced roll-to-roll nanoimprinting process, shown schematically. The gravure unit on the left (G) spreads a film of the conducting polymer on the web. This is followed by patterning of the film, using the nanoimprinting unit (NIL) on the right. The combining of different techniques allows the fabrication of complex layered structures in a single pass [36].

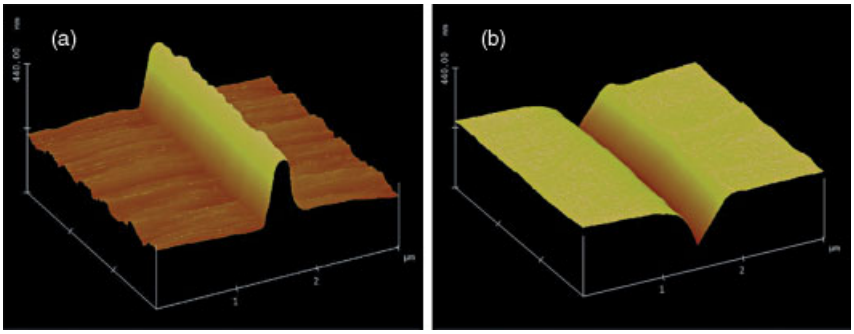


Figure 7.9 (a) Atomic force microscopy (AFM) image of a 100 nm-wide and 170 nm-high ridge on an electroplated roll-to-roll nanoimprinting stamp. (b) AFM image of a trench imprinted into cellulose acetate using the stamp shown in (a). The process temperature is 110 °C and the printing speed 1 m min^{-1}. There was no significant difference between the results obtained at speeds ranging from 0.1 to 5 m min^{-1}.

- Thermal expansion coefficient (e.g., typically from 0.6 to 3×10^{-6} K^{-1}), as well as Poisson's ratio (e.g., typically from 0.1 to 3.0), which will have a strong impact on distortion while demolding.
- Surface smoothness (e.g., better than 0.2 nm), as a rough surface will require large demolding forces and may lead to stronger than needed adhesion.
- Young's modulus (e.g., typically from 70 to hundreds of GPa), which in turn will control possible stamp bending. The latter effect may lead to uneven residual layer thickness, and thus compromise critical dimensions.
- Thermal conductivity (e.g., typically from 6 to hundreds of Wm^{-1} K^{-1}), which determines the duration of the heating and cooling cycles.

With regards to fabrication, the parameters to be considered include the minimum lateral feature size or resolution, the aspect ratio (feature lateral size: feature height), the homogeneity of the feature height across the stamp, as well as depth homogeneity, sidewall roughness, and inclination. For thermal NIL, stamps are usually fabricated in silicon by using EBL and reactive ion etching for the highest resolution and versatility. Unfortunately, these procedures are rather expensive, so that strategies for lower-cost replication have been developed, such as SSIL, the use of a master stamp and a (negative) first-generation replica using NIL or a (positive) second-generation replica, again using NIL. Replication while maintaining a resolution of 100 nm or better requires electroplating to replicate the original in metal. Current developments in the replication of a master stamp in thermosetting polymers show great promise, as they are expected greatly to reduce the cost of stamp replication [37]. As nanoimprint is a 1-to-1 replication technology, it is essential that the stamp has the correct feature sizes required on the wafer, thus emphasizing the need for quality stamps.

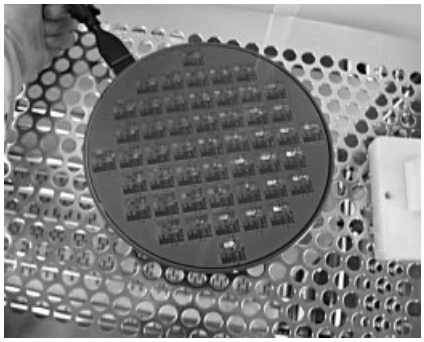


Figure 7.10 Optical image of a 200-mm silicon stamp fabricated by electron beam lithography [25].

In the past, stamps have been realized for wafer-scale thermal NIL (see Figure 7.10). In addition, electron-beam-written silicon stamps for thermal NIL are commercially available from NILTechnology [38], and an example is shown in Figure 7.11.

In recent years, the adhesion between the stamp and the printed polymer film has been the subject of significant research effort in thermal NIL. Here, the main issue is to ensure that the interfacial energy between the stamp and the polymer film to be printed is smaller than the respective interfacial energy between the substrate and the polymer film [39]. However, based on the materials commonly used, this matching is not sufficient for easy detachment, in which the frozen strain also plays a role. The normal practice here in order to facilitate demolding and to prolong the stamp

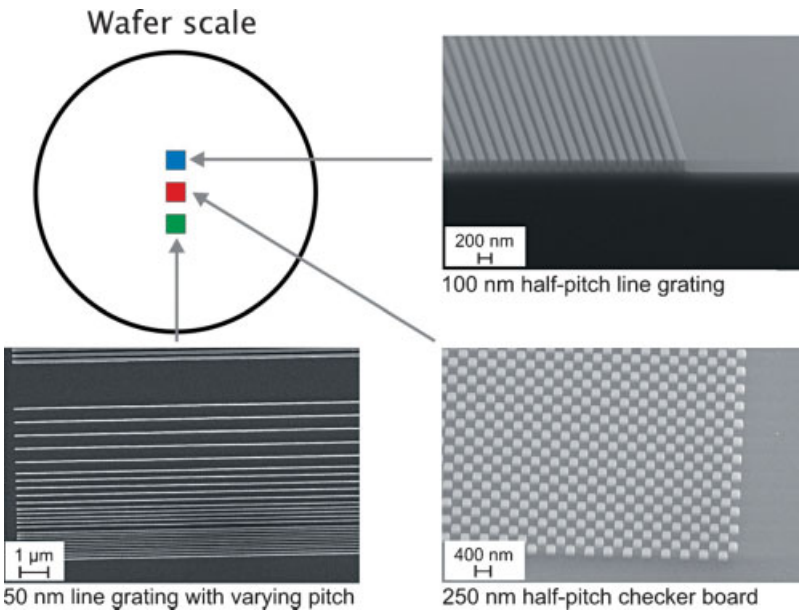


Figure 7.11 Silicon stamp from NILTechnology [38]. (Illustration courtesy of NIL Technology.)

Table. 7.2 Surface energies of common materials used in nanoimprint lithography.

Material	Surface energy (mN m^{-1})
PMMA	41.1
PS	40.7
PTFE	15.6
–CF3 and –CF2	15–17
Silicon surface	20–26

Values are taken from Reference [39].

lifetime, is to coat the stamp with an anti-adhesive layer to minimize the interfacial energy and, therefore, the adhesion. Values of the surface energies of materials commonly used in the NIL process are listed in Table 7.2. These data show that a fluorinated compound can dramatically reduce the surface energy and minimize adhesion while demolding a stamp from the printed polymer.

For both UV-NIL and SFIL, UV-transparent stamps are required, and these are typically constructed from quartz. Although, the fabrication of quartz stamps for high resolution has not yet been standardized, various efforts have been made to use photomask fabrication methods to prepare stamps or templates for UV-NIL [40]. A schematic overview of the stamp fabrication process is shown in Figure 7.12. In a first lithography step, the stamp is structured using EBL (first level writing), while in a second lithography step the pedestal requirement for imprint is made (second level

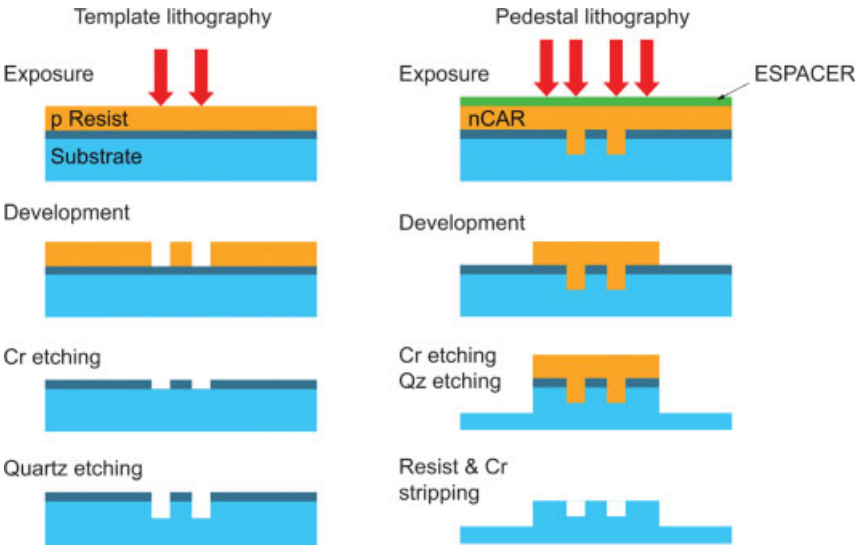


Figure 7.12 Schematics of the fabrication process of two-dimensional stamps for step-and- stamp and or step-and-flash UV-NIL [41, 42].

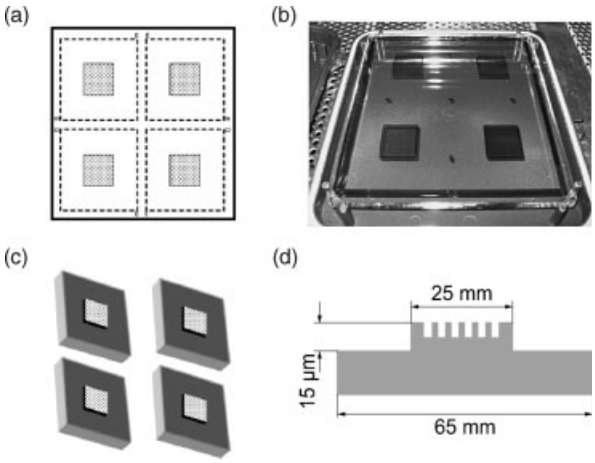


Figure 7.13 Photomask fabrication methods for UV-NIL stamp fabrication [43].

writing). Further details on the process flow for UV-NIL stamps can be found in Refs. [41, 42]. By using photomask fabrication methods, four imprint stamps or templates may be structured on a single photomask blank (see Figure 7.13a and b). The photomask blank is then diced into separate stamps (Figure 7.13c). As dicing introduces some contamination and mechanical strains, a modified fabrication process must be introduced before step-and-repeat- and step-and-flash- UV-NIL can be employed in volume production. The size standard for stamps resembles the exposure field of current optical lithography steppers (see Figure 7.13d).

Recently, stamps for 3-D structuring tests of several layers of functional films using UV-NIL targeting the back-end CMOS processes have been developed, using stamps similar to that shown schematically in Figure 7.13d.

The fabrication of stamps with high-resolution features for roll-to-roll nanoimprinting is more complicated because patterning of the curved surfaces is not straightforward. One possibility way to overcome this is to make a bendable shim that is wrapped around the printing roll. Such bendable large area stamps can be fabricated by electroplating, and exploiting SSIL in large-area pattering has been shown to reduce the fabrication time remarkably [44]. A 100 mm-diameter bendable Ni stamp is shown in Figure 7.14; this figure also shows that sub-100-nm features can be easily reproduced by using an electroplating process.

The details of stamps used for 3-D printing are discussed later in the chapter.

7.2.5
Residual Layer and Critical Dimensions

Most of the processes described above yield a nanostructured polymer layer (as shown schematically in Figure 7.1), with a residual layer under the features of the stamp. If the desired nanostructured surface is the polymer itself, with no material

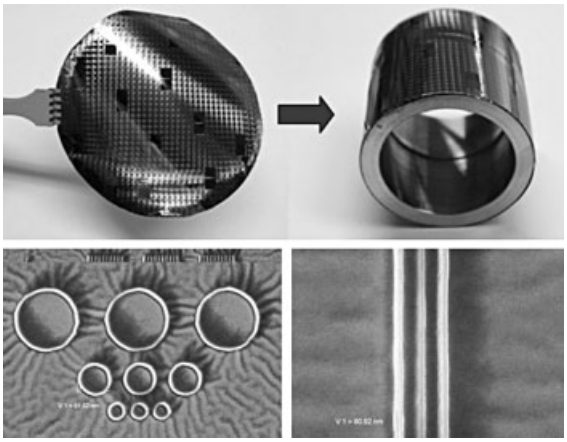


Figure 7.14 (a) An electroplated, bendable 100 mm-diameter Ni shim. The thickness of the shim is 70 μm. The roll-to-roll stamp is made by wrapping the shim around a stainless steel roll. (b) Scanning electron microscopy images of various 80 nm-wide features on an electroplated Ni shim patterned by SSIL. The surface metal layer on the stamp is TiW.

between the features, then the residual layer must be removed. The same applies [44] if the patterned polymer or resist is to be used directly as a mask for pattern transfer into the substrate by, for example, reactive ion etching, or if a metal lift-off step will be needed that will result in a metal mask for further pattern transfer [45]. An etching step of the printed polymer, whether to remove the residual layer or to be used as a mask, necessarily results in the feature sizes experiencing change. This was shown in the variation in the width of printed Aharonov–Bohm ring leads following removal of the residual layer by etching, and after metal lift-off. The leads increased by 15 nm in width from the targeted width of 500 nm, taken over an average of 20 samples [32]. This means that, in order to control the critical dimensions of the printed features, the residual layer thickness uniformity must also be controlled, as its removal leads to a size fluctuation of the resulting nanostructures. Significant efforts have been made to develop non-destructive metrology for nanoimprinted polymers. One of the most salient approaches is to use scatterometry as applied to NIL [46], in order to determine both the feature height and residual layer thickness. Being based on the principles of ellipsometry, a laser spot is used, which is scanned over the region of interest. An example of this is shown in Figure 7.15 (right panel), with a cross-sectional SEM image of the printed ridges and the corresponding scatterometry data and curve fitting. The left panel of Figure 7.15 depicts an optical reflection image of a printed wafer, showing the thickness variation of the residual layer across the wafer [47]. An *in-situ* and non-destructive method was demonstrated by adding chromophores to the printed polymer and using their emission as an indicator of stamp deterioration (such as missing features), mirrored in the printed fields. Although the resolution of this method was poor, it did at least demonstrate the feasibility of the in-line monitoring of printed arrays of nanostructures [48].

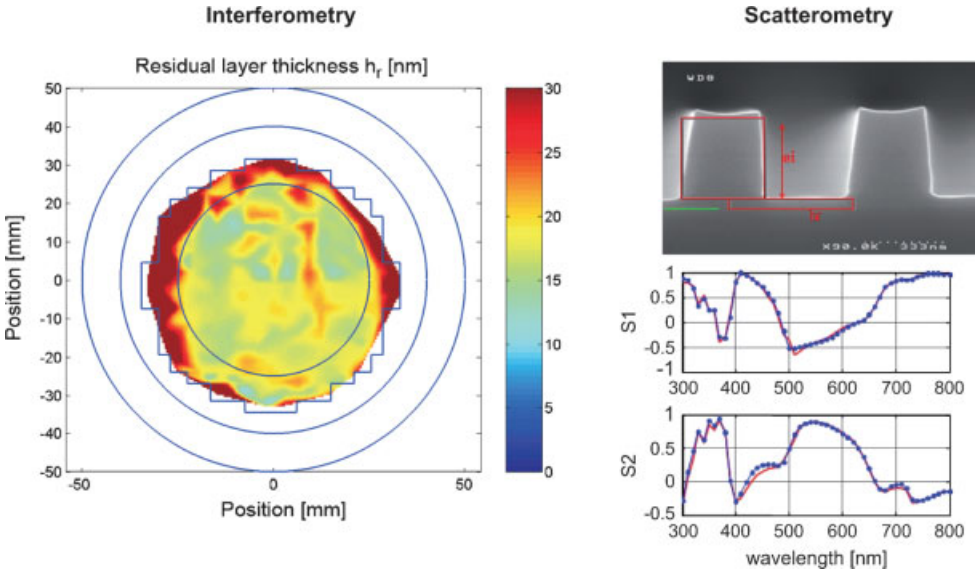


Figure 7.15 Two approaches to metrology. Left: Optical imaging of the contrast resulting from variations of the residual layer thickness over a 100-mm wafer, showing a good uniformity over most of the wafer, except at the edges [47]. Right: The upper image is a cross-section micrograph showing the feature height and residual layer thickness; the lower images show experiments and fit of a scatterometry spectrum recorded on a printed sample, from which the residual layer as well as the feature size can be obtained [46].

Control of the residual layer is necessary due to a need to fill in completely the stamp cavities, whilst achieving as thin and as uniform a residual layer as possible. This is a non-trivial issue which depends not only on nanometer-scale polymer rheology but also on the stamp and substrate deformation.

The polymer challenge in thermal NIL is basically four-fold: (i) to obtain complete filling of the cavities or to ensure a one-to-one transfer; (ii) to obtain as thin a residual layer as possible to control critical dimensions; (iii) to ensure that the printed features do not relax mechanically; and (iv) to achieve a reasonable throughput.

A typical curve of the dependence of viscosity on temperature (e.g., Figure 7.3) shows that T_g occurs in a regime where the viscosity is changing by several orders of magnitude. This poses a non-negligible challenge to understanding polymer flow in the context of NIL. Initially, the polymer flow has been approximated to that of a Newtonian fluid in the gap between two parallel disks of radius R (as discussed in Ref. [49]). The discussion of Ref. [49], which is summarized below, is probably the most complete account to date covering the simple case and providing an insight into the scope of the problem. By using the Stefan equation for the quasi steady-state solution (this is a simplified version of the non-stationary Navier–Stokes equation), the force is found to be proportional to the viscosity, the fourth power of the disk radius, the speed of the disks coming together, and inversely proportional to the cube of the initial layer thickness. In other words, a huge force is needed for a fast fluid

motion in thin films over large distances. There are two basic considerations to this point:

- The force has only a linear dependency on viscosity, which changes by orders of magnitude when the temperature is in the vicinity of T_g.
- Although a Newtonian fluid, or a fluid in the limit of small shear rates, the viscosity does not depend on the shear rate. However, at moderate or high shear rates, a non-linear flow can lead to a decrease in viscosity by several orders of magnitude. The effect of this on the NIL process would be seen as a reduction either in the pressure needed or in the processing time.

The question is, therefore, what are the contributions to the force from pressure and shear stress of the fluid motion? Clearly, pressure is related to the contact area of the stamp and the fluid (polymer above T_g), whereas the shear stress is related to the flow velocity, which in turn depends on the distances over which the fluid must be transported, and therefore on the particular stamp design. At any given time, the condition of continuity and the conservation of momentum of an incompressible liquid requires that the velocity must increase with radial distance, which would result in a parabolic velocity profile in the z-direction. The velocity would be least at the interface with the disk walls, and greatest in the middle of the gap. To this, the inversely proportional cubic dependence on liquid layer thickness must be added. The calculated values of viscosity and transport thickness tend to agree with the observed experimental values for polymethyl methacrylate (PMMA) and, rather simplistically, some basic trends can be obtained:

- Thermal NIL works best for smallest features (sub-100 nm) which are close together and in which a local flow takes place, allowing easy and reliable filling of the stamp cavities.
- Conversely, large features (>10 μm) separated by large distances require a large displacement of material, and larger forces.

Here, the force–displacement curve results reviewed in Ref. [23] are highly illuminating, as a more complete model requires the consideration of several flow fields arising from the different shapes, depths, and separation of cavities in the stamp. Schift and Heyderman carried out a thorough analysis in this respect in the linear micrometer regime [50].

In the linear regime, the temperature dependence of the viscosity is viewed as a thermally activated process [49, 50], following a formalism of amorphous polymers and remaining within the limit of small shear rates. Such a non-linear regime is substantially more complex, and is basically exemplified by shear thinning and extrudate swelling. Hoffmann suggested that shear thinning, with its inherent shear rate-dependent viscosity, may influence the thermal NIL process, especially for small features [49].

A key remaining issue is the understanding of how the stored deformation energy depends on the rate at which the temperature and pressure are applied and released,

and to what degree these influence the mechanical stability of the printed polymer features in time scales of weeks, months, and years.

In practice, on order to gain an understanding of the filling dynamics of a stamp cavity under the combined effects of squeezing flow, polymer rheology, surface tension and contact angle in typical NIL experiments, full fluid–solid interaction models based on the continuum approach have been devised [51, 52]. In these, both the fluid bed and the solid stamp are represented and a continuity of displacement and pressure is applied at the interface. As these are based on finite elements, there are almost no limits to the choice of the materials' constitutive behaviour, and these clearly reflect the effects of stamp anisotropy and the shear thinning behavior of the polymer. In particular, they are especially efficient at predicting the shape of the polymer in partially filled cavities.

Coarse grain methods have proven powerful in computing the residual layer thickness of the embossing process [53]. Based on the Stokes equation, they solve the simple squeeze flow equation for Newtonian fluids and embossed areas of up to

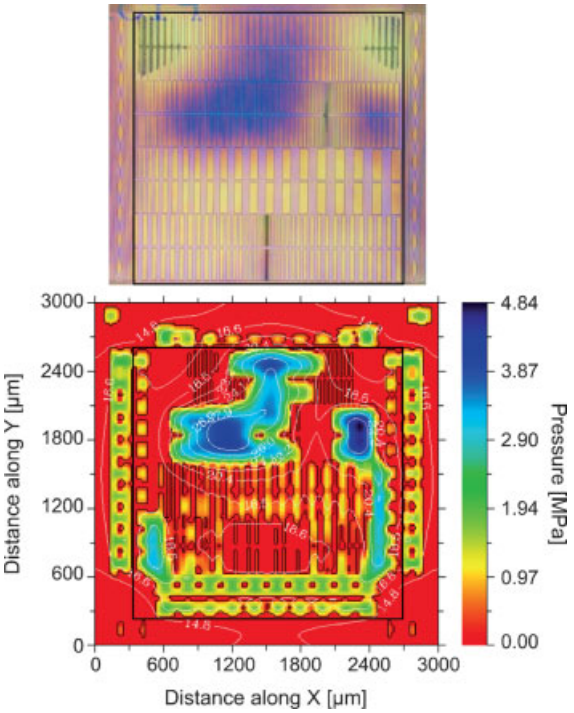


Figure 7.16 Experimental (top) and simulated (bottom) residual layer thickness. The colors correspond to different thicknesses, as observed in an optical microscope. The self-consistent coarse-grain model considers that the stamp is flexible; thus, the resulting contour lines are the variation with respect to the imposed average residual layer thickness. (Illustration courtesy of D.-A. Mendels and S. Zaitsev.).

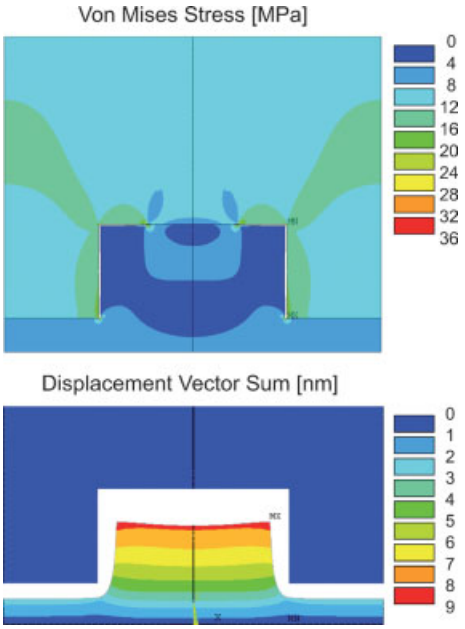


Figure 7.17 Upper: Von Mises stress and stamp/polymer interfacial separation during cool-down of a $200 \times 100\,nm^2$ single polymer cavity obtained by embossing, with a poor interfacial adhesion. Lower: Residual displacement and shape of the same structure after stamp removal in the case of high interfacial adhesion. (Illustration courtesy of D.-A. Mendels.)

several square millimeters within a matter of minutes. Here, the calculation is based on the determination of a homogenized depth which is representative of the average pressure applied to the area. The quantitative agreement has proven excellent, and is generally acceptable when the polymer layer is embossed well above T_g [54]. Freezing of the embossed structures through the T_g has also been modeled [55, 56], and has provided precious insight into the build-up of internal stresses prior to stamp release and of the polymer–stamp interface. It has also been possible to simulate the demolding process, and thus the final shape of the embossed structures, both after stamp release and after relaxation for a given period of time [57]. Two examples of the models described in this section are shown in Figures 7.16 and 7.17.

7.2.6 Towards 3-D Nanoimprinting

One special aspect of NIL and SFIL is their ability to pattern in three dimensions compared to other lithographies. Several applications require this ability, from MEMS to photonic crystals, including a myriad of sensors. One of the first demonstrations of 3-D patterning by NIL was the realization of a T-gate for microwave transistors with a footprint of 40 nm by a single-step NIL and metal lift-off [58]. SFIL

also showed its 3-D pattering ability in the fabrication of multitiered structures, maintaining a high aspect ratio [59].

If metal lift-off is to be avoided, then 3-D NIL requires 3-D stamps. These are produced by gray-scale lithography with sub-100 nm resolution, but are limited in depth and volume production due to the sequential nature of EBL [60]. One recent variation of this approach consisted of using inorganic resists and low-acceleration electron-beam writing, thus allowing the control of the depth to tens of nanometers [61].

A combination method which was based on focused ion beam and isotropic wet etching has been demonstrated by Tormen *et al.* [62], and resulted in tightly controlled 3-D profiles in the range from 10 nm to 100 μm.

Within the microelectronics industry, one of the main expectations from NIL was its application as a lithography method in the dual damascene process for back-end CMOS fabrication [1], and this process is still undergoing testing today.

Bao *et al.* showed that it is possible to print over non-flat surfaces using polymers with different mechanical properties using thermal NIL and polymers with progressively lower T_g-values for each subsequent layer [63]. This meant that a different polymer must be used for each layer. In order to overcome this situation, several other variations and combinations of methods based on NIL have been developed. One such development is that of reversed contact ultraviolet nanoimprint lithography (RUVNIL) [64], which combines the advantages of both reverse nanoimprint lithography (RNIL) and contact ultraviolet (UV) lithography. In this process, a UV crosslinkable polymer and a thermoplastic polymer are spin-coated onto a patterned

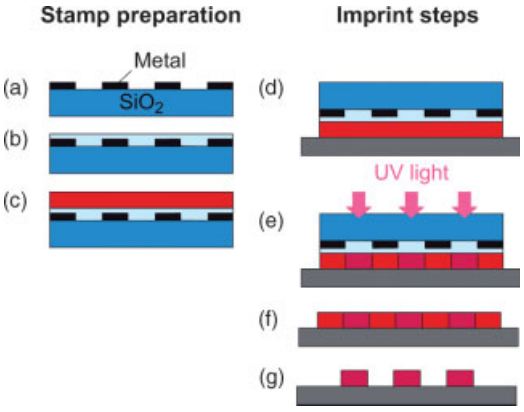


Figure 7.18 Schematics of the reverse contact UV NIL (RUVNIL) process. Left panel: Steps to prepare a stamp. (a) A hybrid mask of SiO_2 with metal feature; (b) a thermoplastic polymer, for example, mr-I 7030, is spin coated; (c) a UV crosslinkable polymer, for example, mr-I 6000 is spin-coated. Right panel: Steps to obtain nanostructures by this method. (d) Reverse imprinting on a Si substrate is carried out; (e) the silicon substrate is heated to heat the polymer above T_g, and pressure is applied; (f) the polymer is cooled down and exposed to UV light; (g) the stamp is separated from the substrate; (h) the exposed polymer layer is developed in acetone, resulting in a polymer pattern with no residual layer. The fabrication time per printed layer is just under 2 min [64].

hybrid metal–quartz stamp. The thin polymer films are then transferred from the stamp to the substrate by contact at a suitable temperature and pressure, after which the whole assembly is exposed to UV light. Following separation of the stamp and substrate, the unexposed polymer areas are rinsed away with a suitable developer, leaving behind the negative features of the original stamp. The process is shown schematically in Figure 7.18.

By using the same UV-curable polymer for each layer, 3-D nanostructures have been obtained (Figure 7.19). This technique offers a unique advantage over reverse-contact NIL and thermal NIL, as no residual layer is obtained by controlling the UV light exposure. This avoids the normal post-imprinting etching step, and therefore results in a much better control of the critical dimensions. Another interesting feature here is that it is not necessary to treat the stamp with an anti-adhesive coating.

Three-dimensional UV-NIL can be potentially used in the fabrication of modern integrated circuits, which employ several layers of copper interconnects, separated by an interlayer dielectric (ILD) and connected by copper vias (Figure 7.20). An imprintable and curable UV-ILD material is deposited on an existing interconnect layer (Figure 7.20a), this material having been structured by a 3-dimensional

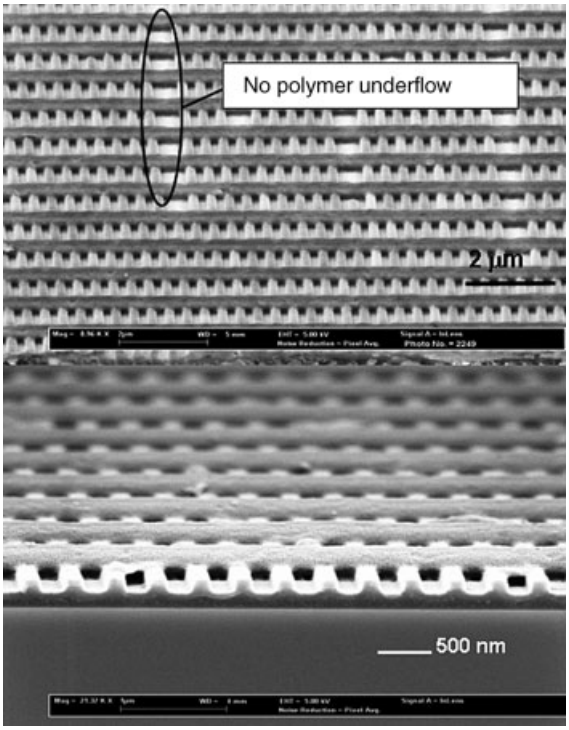


Figure 7.19 RUVNIL prints showing two layers printed without leaving a residual layer, and avoiding polymer overflow of the second layer [64].

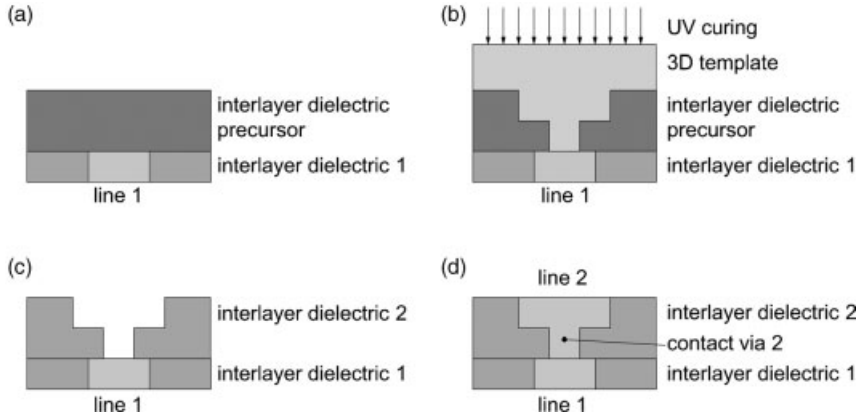


Figure 7.20 Schematics of direct-printing of interconnect layers using a three-dimensional stamp. (Illustration courtesy of L. Berger.)

stamp (Figure 7.20b). After UV-curing, the materials resembles a structured ILD (Figure 7.20c), which is then filled with copper to form two layers of vias and interconnects (Figure 7.20d).

7.2.7 The State of the Art

A comparison of the methods discussed to date, in addition to some relevant data, are displayed in Table 7.3. This information forms part of the studies of the European integrated project "Emerging Nanopatterning Methods (NaPa)" [65], which is exploring several non-optical lithographic methods with the purpose of gathering a library of processes that employ some of these newly emerging patterning technologies.

7.3 Discussion

Nanoimprint lithography, as an example of non-optical lithographies, has proven to be a versatile patterning method in several fields of application where a rather rapid development has been demonstrated, in addition to sole pattern transfer, notably in the areas of optics [66] (some of them at 200 mm wafer scale [67]) and microfluidics [68]. The versatility of NIL opens new possibilities for the nanostructure of various types of functional material, such as conducting polymers [69], light-emitting polymers [70], polymers loaded with nanocrystals [71], and biocompatible polymers [72]. An example of photonic applications is shown in Figure 7.21, which depicts a printed two-dimensional photonic crystal in polymer. The patterning of functionalized materials may be difficult when using traditional methods such as optical or electron beam

Table. 7.3 Comparison of the different printing techniques.

Technique	Smallest/largest features in same print	Min pitch (nm)	Largest wafer printed (mm)	Overlay Accuracy (nm)*	T align, T print, T release, T cycle	No. of times stamp used	Materials
NIL	5 nm[a]/N/A	14	200[b]	500	Minutes, 10 s, Min, 10–15 min	>50[c]	Various
SSIL[d]	8 nm min features. 50 nm/5 μm on same stamp	50	200	<250	Full cycle 2.5 min with, 20 s without full auto-collimation.	1000	mr-I 8000, mr-I 7000
SFIL[e]	25 nm[f]/μm	50[f]	300[g,h,m] stamp size: ~26×26 mm^2	50[i] (about 20[i])	20 wafers/h[j]	800[g]	Various NILTM105, AMONIL, PAK 01
UV-NIL[k]	9 nm/100 μm[l]	12[m]	200[n]	about 20[o]	20 s/step[p] three wafers/h[q]	>1000[r]	MRT07xp, PAK01, AMONIL1, AMONIL2, NXR, Laromer
Soft UV-NIL	25 nm/20 μm[s]	150[t]	200	1–50 μm[u]	4–5 min; about 12 wafers/hr[v]	>50[w]	AMONIL1, NXR-Mod, Laromer

Data from several sources.

[a]Depends more on the equipment than on the imprinting method.

[b]M.D. Austin, *et al.*, *Appl. Phys. Lett.* **2004**, *84*, 5299.

[c]From Ref. [25].

[d]This value is from manual tests. A cassette-loading tool will have better values.

[e]Step-and-stamp imprint lithography is based on thermal NIL using the step-and-stamp imprinting tool, NPS300 by SUSS MicroTec.

[f]Step-and-flash imprint lithography.

[g]D.J. Resnick, G. Schmid, E. Thompson, N. Stacey, D.L. Olynick and E. Anderson, *Step and Flash Imprint Lithography Templates for the 32 nm Node and Beyond*, NNT 06, San Francisco, US, November 15–17, **2006**.

[h]M. Miller, G. Schmid, G. Doyle, E. Thompson and D. J. Resnick, *S-FIL Template Fabrication for Full Wafer Imprint Lithography*, NNT 06, San Francisco, US, November 15–17, **2006**.

(Continued)

[i]T.-Wei Wu, M. Best, D. Kercher, E. Dobisz, Z. Bandic, H. Yang and T. R. Albrecht, *Nanoimprint Applications on Patterned Media*, NNT 06, San Francisco, US, November 15–17, **2006**.
[j]S.V. Sreenivasan, P. Schumaker, I. McMackin and J. Choi, *Nano-Scale Mechanics of Drop-On-Demand UV Imprinting*, NNT 06, San Francisco, US, November 15–17, **2006**.
[k]R. Hershey, M. Miller, C. Jones, M. G. Subramanian, X. Lu, G. Doyle, D. Lentz and D. LaBrake, *SPIE***2006**, *6337*, 20.
[l]UV-NIL includes Single-Step & Step&Repeat on a EVG770 tool.
[m]B. Vratzov, *et al.*, *J. Vac. Sci. Technol. B***2003**, *21*, 2760; and http://www.amo/de.
[n]S. Y. Chou, *et al.*, *Nanotechnology***2005**, *16*, 10051.
[o]4-inch (10-cm) Single-Step, 8-inch (20-cm) Step&Repeat on EVG770.
[p]A. Fuchs, *et al.*, *J. Vac. Sci. Technol.* **2004**, *22*, 3242–3245, and to be published.
[q]Without fine alignment nor automation yet.
[r]http://www.molecularimprints.com.
[s]M. Otto, *et al.*, *Microelectronic Eng.* **2004**, *73–74*, 152.
[t]U. Plachetka, *et al.*, *Microelectronic Eng.* **2006**, *83*, 944.
[u]Pitch for Soft UV-NIL tested so far and to be published.
[v]Only coarse alignment available; depending on stamp material used.
[w]70–80% of the given time is to cure the resist.
[x]Based on laboratory tests to date.

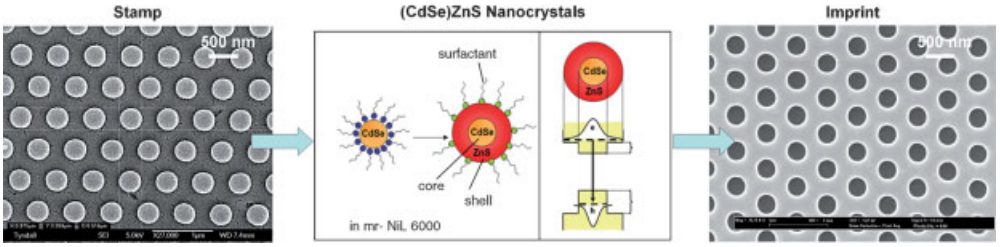


Figure 7.21 Two-dimensional photonic crystals printed in a polymer containing semiconductor quantum dots to control the spontaneous emission. This single-step imprint resulted in a 200% increase of the emission efficiency [71].

lithography, because it may not be possible to incorporate photosensitive components without degrading the functionality; alternatively, the materials may not tolerate the chemical processes associated with these pattern-transfer technologies. As described above, NIL requires only a fairly moderate temperature cycle in order to mediate the patterning process.

One exciting extension of NIL is the potential for patterning curved, 3-D surfaces, and this is yet to be exploited both in research and commercial applications. Somewhat surprisingly, it has been the lack of straightforward ways to provide curved surfaces that has hindered progress in this area. Nonetheless, NIL provides a simple means of realizing various types of curved 3-D surface which, of course, require the fabrication of a master stamp. This ability can be used, for example, in optics [72], cell cultivation [73], and plasmonics [74, 75]. The possibility of aligning to already existing patterns in SSIL and SFIL allows the use of mix-and-match approaches and the combination of more than one technique to build up multifunctional structures. The promise of low-cost and high-throughput uses of NIL in nanofabrication may be fulfilled by the roll-to-roll type of continuous approaches. During the late 1990s, the tools used for nanoimprinting were mainly commercial presses with heating units, but some time later tools based on modified optical mask aligners emerged, both for thermal and UV NIL. Today, several commercially available machines are dedicated to the nanoimprinting processes. The development of materials intended for NIL has also witnessed similar progress since the late 1990s, with not only methods but also instruments and software for non-destructive characterization and metrology having been introduced [46]. Clearly, whilst NIL is becoming a mature and capable technology for nanofabrication, its prospective roles are reaching even further, and have been included among the top ten technologies considered capable of "changing the world" [76]. This situation is reflected in the dissemination of information pertaining to NIL, with the numbers of published reports increasing at breath-taking pace, along with numerous conferences and discussion sessions of microfabrication and nanofabrication systems dedicated to nanoimprinting.

Despite these many advances, much remains to be understood and achieved. One such example is the need to pin down design rules based on an understanding of non-linear processes in viscous flow. If NIL is to improve its throughput, then by necessity faster processes will have to be partly non-linear and undergo concomitant modeling

challenges. Here, the impact will be upon stamp layout, and on the design rules. With regards to critical dimensions, those applications with a strict control of periodicity and smoothness of features will serve as the "acid test" for NIL. For example, if NIL is to be used as a mask-making method for the transfer of 2-D photonic crystal patterns into a high-refractive index material then, in addition to alignment, the "disorder" must be controlled to better than a few nanometers after pattern transfer – that is, after reactive ion etching. Although these critical dimensions are more relaxed in lower-refractive index photonic crystals, such as those printed directly in polymers [71], the verticality of the side walls is still of paramount importance. While the current resolution of NIL is already of a few nanometers, such demands will need to be even more stringent in order to fabricate hypersonic phononic crystals and nanoplasmonic structures, where the relevant wavelengths are of the order of only a few nanometers.

The versatility of the printable polymers, the resolution of NIL and the ability to realize 3-D structures open further possibilities. One such advance is the use of 3-D templates or scaffolds to provide not only the support but also input and output contacts for supramolecular structures. To be successful, this has two requirements: first, the feature sizes must be commensurate, and the structured polymer surface site-selectively functionalized. Whilst the complete proof of concept is still missing, some degree of progress has been made towards spatially selective functionalization by means of NIL, chemical functionalization and lift-off, and this has resulted in electrical contacts to 150 nm-wide arrays of polypyrrole nanowires [76]. The second requirement is in the use of 3-D nanostructures, beyond the face-centered cubic (fcc) and cubic symmetries, the properties of which can be modified by subsequent surface treatment, while preserving the symmetry. Modifications may include coating with oxides, and also removal of the polymer template, followed by subsequent infilling with another material. In this way, an artificial 3-D superlattice may be realized, thereby providing a periodic or quasi-periodic arrangement for electronic and or optical excitations.

7.4 Conclusions

In this chapter we have reviewed some of the key developments of NIL, as a non-optical lithography method, paying particular attention to the schematics of the process, and discussing: (i) materials issues and their impact on the process; (ii) the variations of NIL (among which roll-to-roll appears particularly promising for volume production); (iii) stamp design, both in terms of robustness and adhesion; (iv) the issue of residual layer thickness and its impact on critical dimensions; and (v) finally briefly reviewing the latest progress in 3-D NIL. In addition, we have discussed the importance of understanding polymer flow as an enabling knowledge to optimize stamp design, and thereby throughput. When discussing 3-D NIL, the scope for further progress was outlined as to date this is largely unexplored.

NIL has been said to have short-term prospects in back-end CMOS fabrication processes, but more so in the fabrication of photonic structures and circuits, with a

proviso in case of photonic crystals where stringent tolerances are still to be met. On the other hand, applications in less-demanding areas, such as gene chips for diagnostic screening, appear to have a very bright future.

Acknowledgments

he authors gratefully acknowledge the support of the EC-funded project NaPa (Contract no. NMP4-CT-2003-500120) and of Science Foundation Ireland (Grant No. 02/IN.1/172). They are also grateful to A. Kristensen, H. Schift, M. Tormen, T. Haatainen, P. Majander, T. Mäkelä, N. Kehagias, V. Reboud, M. Zelsmann, F. Reuther, D.-A. Mendels and many other colleagues of the NaPa project for fruitful discussions and joint investigations over the years. Note: The content of this work is the sole responsibility of the authors. Part of the sections on stamps (Photomasks) and on 3-D stamps for integrated electronic circuits were kindly provided by L. Berger.

References

1 http://www.itrs.net/Links/2006/Update/Final/ToPost/08_Lithography2006.Update.pdf.

2 Sotomayor Torres, C.M. (Ed.) (2003) *Alternative Lithography: Unleashing the Potentials of Nanotechnology*, Kluwer Academic Plenum Publishers, New York.

3 (a) Chou, S.Y., Krauss, P.R. and Renstrom, P.J. (1995) *Applied Physics Letters*, **67**, 3114. (b) Chou, S.Y. (1998) United States Patent, No. 5,772,905, 57.

4 Garcia, R. (2003) *Alternative Lithography: Unleashing the Potentials of Nanotechnology* (ed. C.M. Sotomayor Torres), Kluwer Academic Plenum Publishers, New York, p. 213.

5 Piner, D., Zhu, J., Xu, F., Hong, S. and Mirkin, C.A. (1999) *Science*, **283**, 661.

6 Xia, Y., Zhao, X.-M. and Whitesides, G.M. (1996) *Microelectronic Engineering*, **32**, 255.

7 Michel, B., Bernard, A., Bietsch, A., Delamarche, E., Geissler, M., Juncker, D., Kind, H., Renault, J.-P., Rothuizen, H., Schmid, H., Schmidt-Winkel, P., Stutz, R. and Wolf, H. (2001) *IBM Journal of Research and Development*, **45**, 697.

8 Brugger, J., Berenschot, J.W., Kuiper, S., Nijdam, W., Otter, B. and Elwenspoek, M. (2000) *Microelectronic Engineering*, **53**, 403.

9 Muetzel, M., Tandelr, S., Haubrich, D., Meschede, D., Peithmann, K., Flaspoehler, M. and Buse, K. (2002) *Physical Review Letters*, **88**, 083601.

10 Zubarev, E.R., Xu, J., Sayyad, A. and Gibson, J.D. (2006) *Journal of the American Chemical Society*, **128**, 15098.

11 Ruda, J.C. (1977) *Journal of the Audio Engineering Society*, **11/12**, 702.

12 Haisma, J., Verheijen, M., van den Heuvel, K. and van den Berg, J. (1996) *Journal of Vacuum Science & Technology B*, **14**, 4124.

13 Scheer, H.-C., Schulz, H., Hoffmann, T. and Sotomayor Torres, C.M. (2002) *Handbook of Thin Film Materials* (ed. H.S. Nalwa,), Vol. **5** Academic Press, New York, p. 1.

14 Guo, L.J. (2004) *Journal of Physics D-Applied Physics*, **37**, R123.

15 Schift, H. and Kristensen, A. (2007) *Handbook of Nanotechnology*, (ed. B. Bhushan), 2nd edn., Springer, Berlin, Heidelberg, p. 239.

16 Hirai, Y., Fujiyama, M., Okuno, T., Tanaka, Y., Endo, M., Irie, S., Nakagawa, K. and

Sasago, M. (2001) *Journal of Vacuum Science & Technology B*, **19**, 2811.

17 (a) Heyderman, L.J., Schift, H., David, C., Gobrecht, J. and Schweizer, T. (2000) *Microelectronic Engineering*, **54**, 229. (b) Schift, H., Heyderman, L.J., Auf der Maur, M. and Gobrecht, J. (2001) *Nanotechnology*, **12**, 173.

18 (a) Reuther, F., Fink, M., Kubenz, M., Schuster, C., Vogler, M. and Gruetzner, G. (2005) *Microelectronic Engineering*, **78–79**, 496. (b) Pfeiffer, K., Reuther, F., Carlsberg, P., Fink, M., Gruetzner, G. and Montelius, L. (2003) *Proceedings of SPIE*, **5037**, 203.

19 Reuther, F. (2005) *Journal of Photopolymer Science and Technology*, **18**, 523.

20 http://microresist.de/, micro resist technology GmbH, Koepenicker Str. 325, 12555 Berlin, Germany.

21 Vogler, M., Wiedenberg, S., Mühlberger, M., Glinsner, T. and Grützner, G. (2006) poster P_NIL_31 3C-6, presented at the Micro- and Nano-Engineering International Conference MNE 2006, 17–20 September, Barcelona, Spain.

22 Hirai, Y., Konishi, T., Yoshikawa, T. and Yoshida, S. (2002) *Journal of Vacuum Science & Technology B*, **22**, 3288.

23 Cross, G.L.W. (2006) *Journal of Physics D-Applied Physics*, **39**, R363.

24 Heidari, B., Maximov, I., Sarwe, E.-L. and Montelius, L. (2000) *Journal of Vacuum Science & Technology B*, **18**, 3552.

25 Gourgon, C., Perret, C., Tallal, J., Lazzarino, F., Landis, S., Joubert, O. and Pelzer, R. (2005) *Journal of Physics D-Applied Physics*, **38**, 70.

26 http://www.obducat.com, Obducat AB, Box 580, 20125 Malmö, Sweden.

27 http://evgroup.com, EV Group, E. Thallner GmbH, DI Erich Thallner Strasse 1, A-4782 St. Florian/Inn, Austria.

28 Matsui, S., Igaku, Y., Ishigaki, H., Fujita, J., Ishida, M., Ochiai, Y., Komuro, M. and Hiroshima, H. (2001) *Journal of Vacuum Science & Technology B*, **19**, 2801.

29 Matsui, S., Igaku, Y., Ishigaki, H., Fujita, J., Ishida, M., Ochiai, Y., Namatsu, H. and Komuro, M. (2003) *Journal of Vacuum Science & Technology B*, **21**, 688.

30 Haatainen, T., Ahopelto, J., Gruetzner, G., Fink, M. and Pfeiffer, K. (2001) *Proceedings of SPIE*, **3997**, 874.

31 http://www.S.E.T.-SAS.fr, S.E.T.S.A.S., 131, impasse Bartheudet, BP24, F-74490, Saint-Jeoire, France.

32 Ahopelto, J. and Haatainen, T. (2003) *Alternative Lithography: Unleashing the Potentials of Nanotechnology* (ed. C.M. Sotomayor Torres), Kluwer Academic Plenum Publishers, New York, p. 103.

33 Colburn, M., Johnson, S.C., Stewart, M.D., Damle, S., Bailey, T.C., Choi, B., Wedlake, M., Michaelson, T.B., Sreenivasan, S.V., Ekerdt, J.G. and Grant Willson, C. (1999) *Proceedings of SPIE*, **3676** (I), 379.

34 Bailey, T.C., Colburn, M., Choi, B.J., Grot, A., Ekerdt, J.K., Sreenivasan, S.V. and Wilson, C.G. (2003) *Alternative Lithography: Unleashing the Potentials of Nanotechnology* (ed. C.M. Sotomayor Torres), Kluwer Academic Plenum Publishers, New York, p. 117.

35 http://www.sfil.org and http://www.molecularimprints.com.

36 Mäkelä, T., Haatainen, T., Majander, P. and Ahopelto, J. (2007) *Microelectronic Engineering*, **84**, 877–879.

37 Schultz, H., Lyebyedyev, D., Scheer, H.-C., Pfeiffer, K., Bleidiessel, G. and Gruetzner, G. (2000) *Journal of Vacuum Science & Technology B*, **18**, 3582.

38 http://www.nilt.com/NILTechnology, Oersteds Plads, DTU-Building 347, DK-2800 Kongens Lyngby, Denmark.

39 Brandrup, J. and Immergut, E.H. (1975) *Polymer Handbook*, 2nd edn., John Wiley & Sons, New York.

40 Sasaki, S., Itoh, K., Fujii, A., Toyama, N., Mohri, H. and Hayashi, N. (2005) *Proceedings of SPIE*, **5853**, 277.

41 Hudek, P., Beyers, D., Groves, T., Fortagne, O., Dauksher, W.J., Mancini, D., Nordquist, K. and Resnick, D.J. (2004) *Proceedings of SPIE*, **5504**, 204.

42 Dauksher, J., Mancini, D., Nordquist, K., Resnick, D.J., Hudek, P., Beyer, D. and

Fortagne, O. (2004) *Microelectronic Engineering*, **75**, 345.
43 Institut fuer Mikroelektronik Stuttgart, www.ims-chips.de.
44 Haatainen, T., Majander, P., Riekkinen, T. and Ahopelto, J. (2006) *Microelectronic Engineering*, **83**, 948.
45 Arakcheeva, E.M., Tanklevskaya, E.M., Nesterov, S.I., Maksimov, M.V., Gurevich, S.A., Seekamp, J. and Sotomayor Torres, C.M. (2005) *Technical Physics*, **50**, 1043. (Translated from Zhurnal Tekhnichesko Fiziki **2005**, 75, 80–84).
46 Fuard, D., Perret, C., Farys, V., Gourgon, C. and Schiavone, P. (2005) *Journal of Vacuum Science & Technology B*, **23**, 3069.
47 Nielsen, T., Pedersen, R.H., Hansen, O., Haatainen, T., Tollki, A., Ahopelto, J. and Kristensen, A. (2005) Technical Digest 18th IEEE Conference Micro Electro Mechanical Systems, MEMS 2005, Miami, FL, USA, January 30–February 3, 2005, pp. 508.
48 Finder, Ch., Beck, M., Seekamp, J., Pfeiffer, K., Carlberg, P., Maximov, I., Reuther, F., Sarwe, E.-L., Zankovych, S., Ahopelto, J., Montelius, L., Mayer, C. and Sotomayor Torres, C.M. (2003) *Microelectronic Engineering*, **67–68**, 623.
49 Hoffmann, T. (2003) *Alternative Lithography: Unleashing the Potentials of Nanotechnology* (ed. C.M. Sotomayor Torres), Kluwer Academic Plenum Publishers, New York, p. 103.
50 Schift, H. and Heyderman, L. (2003) *Alternative Lithography: Unleashing the Potentials of Nanotechnology* (ed. C.M. Sotomayor Torres), Kluwer Academic Plenum Publishers, New York, p. 47.
51 Rowland, H.D., Sun, A.C., Schunk, P.R. and King, W.P. (2005) *Journal of Micromechanics and Microengineering*, **15**, 2414.
52 Mendels, D.-A. (2006) *Proceedings of SPIE*, **6151**, 615113.
53 Sirotkin, V., Svintsov, A., Zaitsev, S. and Schift, H. (2006) *Microelectronic Engineering*, **83**, 880.
54 Sirotkin, V., Svintsov, A. and Zaitsev, S. Paper presented at the Micro and NanoEngineering 2006, MNE'06, 17–20 September 2006, Barcelona Spain.
55 Mendels, D.-A. (2006) in: Proceedings Nanoimprint and Nanoprint Technology (NNT'06), San Francisco, USA, November 17–20.
56 Worgull, M., Heckele, M., Hétu, J.F. and Kabanemi, K.K. (2006) *Journal of Microlithography, Microfabrication, and Microsystems*, **5**, 011005.
57 Ishii, Y. and Taniguchi, J. (2007) *Microelectronic Engineering*, **84**, 912–915.
58 Li, M., Chen, L. and Chou, S.Y. (2001) *Applied Physics Letters*, **78**, 3322.
59 (a) Colburn, M., Grot, A., Amistoso, M.N., Choi, B.J., Bailey, T.C., Ekerdt, J.G., Sreenivasan, S.V., Hollenhorst, J. and Willson, C.G. (2000) *Proceedings of SPIE*, **3997**, 453. (b) Johnson, S., Resnick, D.J., Mancini, D., Nordquist, K., Dauksher, W.J., Gehoski, K., Baker, J.H., Dues, L., Hooper, A., Bailey, T.C., Sreenivasan, S.V., Ekerdt, J.G. and Willson, C.G. (2003) *Microelectronic Engineering*, **67–68**, 221.
60 Yamazaki, K. *et al.* (2004) *Microelectronic Engineering*, **73–74**, 85.
61 (a) Jun Taniguchi, *et al.* (2004) *Applied Surface Science*, **238**, 324. (b) Ishii, Y. and Taniguchi, J. (2006) Paper P-NIL09, presented at the Micro and NanoEngineering 2006, MNE,(06, 17–20 September, Barcelona Spain.
62 Tormen, M. *et al.* (2005) *Journal of Vacuum Science & Technology B*, **23**, 2920.
63 Bao, L.-R., Cheng, X., Huang, X.D., Guo, L.J., Pang, S.W. and Lee, A.F. (2002) *Journal of Vacuum Science & Technology B*, **20**, 2881.
64 (a) Kehagias, N., Zelsmann, M., Pfeiffer, K., Ahrens, G., Gruetzner, G. and Sotomayor Torres, C.M. (2005) *Journal of Vacuum Science & Technology B*, **23**, 2954. (b) Kehagias, N., Reboud, V., Chansin, G., Zelsmann, M., Jeppesen, C., Schuster, C., Kubenz, M., Reuther, F., Gruetzner, G. and Sotomayor Torres, C.M. (2007) *Nanotechnology*, **18**, 175303.
65 http://www.phantomsnet.net/NAPA/index.php.

66 Merino, S., Retolaza, A. and Lizuain, I. (2006) *Microelectronic Engineering*, **83/4-9**, 897.

67 Chaix, N., Landis, S., Gourgon, C., Merino, S., Lambertini, V.G., Repetto, P.M., Durand, G. and Perret, C. (2006) Paper P_NIL01, presented at the Micro- and Nano-Engineering International Conference MNE 2006, 17–20 September, Barcelona, Spain.

68 Kristensen, A., Balsev, S., Gersbrog-Hansen, M., Bilenberg, B., Rasmussen, T. and Nilsson, D. (2006) *Proceedings of SPIE*, **6329**, 632901.

69 Mäkelä, T., Haatainen, T., Ahopelto, J. and Isotalo, H. (2001) *Journal of Vacuum Science & Technology B*, **19**, 487.

70 Kim, C., Stein, M. and Forrest, S.R. (2002) *Applied Physics Letters*, **80**, 4051.

71 Reboud, V., Kehagias, N., Zelsmann, M., Striccoli, M., Tamborra, M., Curri, M.L., Agostiano, A., Fink, M., Reuther, F., Gruetzner, G. and Sotomayor Torres, C.M. (2007) *Applied Physics Letters*, **90**, 011115.

72 Bilenberg, B., Hansen, M., Johansen, D., Ozkapici, V., Jeppesen, C., Szabo, P., Obieta, I.M., Arroyo, O., Tegenfeldt, J.O. and Kristensen, A. (2005) *Journal of Vacuum Science & Technology B*, **23**, 2944.

73 Martines, E., Seunarine, K., Morgan, H., Gadegaard, N., Wilkinson, C.D.W. and Riehle, M.O. (2005) *Nano Letters*, **5**, 2097.

74 Tormen, M., Carpentiero, A., Ferrari, E., Cabrini, S., Cojoc, D. and Di Fabrizio, E. (2006) *Proceedings of SPIE*, **6110**, 611055.

75 Pedersen, R.H., Boltasseva, A., Johansen, D.M., Nielsen, T., Jørgensen, K.B., Leosson, K., Østergaard, J.E. and Kristensen, A. Paper P_NIL04, presented at the Micro- and Nano-Engineering International Conference MNE 2006, 17–20 September, 2006, Barcelona Spain.

76 MIT Technology Review (2003) February, 33.

77 Dong, B., Lu, N., Zelsmann, M., Kehagias, N., Fuchs, H., Sotomayor Torres, C.M. and Chi, L. (2006) *Advanced Functional Materials*, **16**, 1937.

8
Nanomanipulation with the Atomic Force Microscope

Ari Requicha

8.1
Introduction

The Scanning Probe Microscope (SPM) provides a direct window into the nanoscale world, and is one of the primary tools that are making possible the current development of nanoscience and nanoengineering. The first type of SPM was the Scanning Tunneling Microscope (STM), invented at the IBM Zürich laboratory by Binnig and Rohrer [1], who received the Nobel Prize for it only a few years later (1986). The STM provided for the first time the ability to image individual atoms and small molecules, and it is still widely used, especially to study the physics of metals and semiconductors. Much of the STM work is conducted in ultra high vacuum (UHV) and often at low temperatures. The STM main drawback is the need for conductive samples, which rules out many of its potential applications in biology and other important areas.

The next instrument to be developed in the SPM family was the Atomic Force Microscope (AFM), sometimes also called Scanning Force Microscope [2]. The AFM has become the most popular type of SPM because, unlike the STM, it can be used with non-conductive samples, and therefore has broad applicability. Today, there are many other types of SPMs. All of these instruments scan a surface with a sharp tip (with apex radius on the order of a few nm), placed very close to the surface (sometimes at distances $\sim$1 nm), and measure the interaction between tip and surface. For example, STMs measure the tunneling current between tip and sample, and AFMs measure interatomic forces – see Section 8.2 below for a lengthier discussion of SPM principles.

It was noticed from the beginnings of SPM work that scanning a sample with the tip often modified the sample. This was initially considered undesirable, but researchers soon recognized that the ability to modify a surface could be exploited for nanolithography and nanomanipulation. In SPM nanolithography one writes

Nanotechnology. Volume 3: Information Technology I. Edited by Rainer Waser

ISBN: 978-3-527-31738-7

lines and other structures directly on a surface by using the SPM tip. A well-known technique is called local oxidation, first demonstrated by passing an STM tip in air over a surface of hydrogen-passivated silicon [3]. Other materials can be used, as well as a conductive AFM tip in lieu of an STM [4]. Another STM method involves removing atoms from a silicon surface by applying voltage pulses to the tip [5]. Lines as narrow as a silicon dimer have been produced by this method. Lithography by material deposition, as opposed to material removal, has also been demonstrated in early work. For example, in [6] atomic-level structures of germanium were deposited on a Ge surface in UHV by pulsing the voltage on an STM tip, whereas in [7] gold clusters were deposited on a gold surface also by applying voltage pulses to the STM tip, but in air and at room temperature. Many other SPM nanolithography approaches have been demonstrated – see [8] for a survey of early work.

More recently, several other SPM nanolithographic techniques have been developed. Some examples follow. Dip Pen Nanolithography [9] involves depositing material on a surface much like one writes with a pen on paper. A pen (the AFM tip) is inked by dipping it into a reservoir containing the material to be deposited, and then it is moved to the desired locations on the sample. As the tip approaches the sample, a capillary meniscus is formed, which drives the material onto the sample. Other approaches are discussed for example in [10–15]. For a recent review see [16].

Nanomanipulation is defined in this chapter as the motion of nanoscale objects from one position to another on a sample under external control. Precise, high-resolution nanolithography shares with nanomanipulation the need for accurately positioning the tip on the sample. This is a challenging issue, which we will discuss later in this article.

Given the atomic resolution achieved by SPM imaging, one would expect also that atoms might be moved individually. This is indeed the case, and it was demonstrated in the early 1990s [17]. At the IBM Almadén laboratory, Eigler's group has been able to precisely position xenon atoms on a nickel surface, platinum atoms on platinum, carbon monoxide molecules on platinum [18], iron on copper [19], and so on, by using a sliding, or dragging process. The tip is brought sufficiently close to an adsorbed atom for the attractive forces to prevail over the resistance to lateral motion. The tip then moves over the surface, and the atom moves along with it. Tip withdrawal leaves the atom in its new position.

Eigler also has succeded in transferring to and from an STM tip xenon atoms on platinum and nickel, platinum on platinum, and benzene molecules. This was done by approaching the atoms or molecules with the tip until contact (or near contact) was established. In addition, xenon atoms on nickel were transferred to the tip by applying a voltage pulse to the tip. All of Eigler's work cited above has been done in ultra high vacuum (UHV) at 4 K.

Avouris group, at the IBM Yorktown laboratory, and Aono's group in Japan have transferred silicon atoms between an STM tip and a surface in UHV at room temperature, by aplying voltage bias pulses to the tip [20, 21].

Atomic manipulation with SPMs continues to be studied today and is providing new insights into nanoscience. For example, in the late 1990s, Rieder's group in Berlin conducted a series of experiments in which they showed that xenon atoms can be pushed and pulled across a copper surface, in UHV and at low temperature, and that the tunnel current during the motion has distinct signatures that correspond to the pushing and pulling modes [22]. As an example of very recent work, a NIST group has shown that a cobalt atom can be moved on a copper surface by exciting it electronically with and STM tip in UHV and at low temperature [23].

Molecules have been arranged into prescribed patterns at room temperature by Gimzewski's group at IBMs Zürich laboratory (now at UCLA). They push molecules at room temperature in UHV by using an STM. They have succeeded in pushing porphyrin molecules on copper [24], and they have arranged bucky balls (i.e., C_{60}) in a linear pattern, using an atomic step in the copper substrate as a guide [25]. They approach the molecules, change the voltage and current values of the STM so as to bring the tip closer to the sample than in imaging mode, and push with the feedback on. C_{60} molecules on silicon also have been pushed with an STM in UHV at room temperature by Maruno and co-workers in Japan [26], and Beton and co-workers in the UK [27]. In Maruno's approach the STM tip is brought closer to the surface than in normal imaging mode, and then scan across a rectangular region with the feedback (essentially) turned off. This may cause probe crashes. In Beton's approach the tip also is brought close to the surface, but the scan is done with the feedback on and a high value for the tunneling current; the success rate is low.

It should be clear from this brief review, which is not meant to be exhaustive, that much work has been done in nanomanipulation and related topics. In this article our focus is on manipulation by using AFMs, in air or a liquid, of objects such as nanoparticles or nanowires, which are larger than atoms or small molecules. The remainder of the chapter is organized as follows. First we address in some detail the principles of operation of the AFM, and the spatial uncertainties associated with the instrument. Next we present various protocols for moving nanoobjects with the AFM tip and discuss research aimed at building nanoassemblies. Section 8.4 addresses systems for nanomanipulation, both interactive and automated. We draw conclusions in a final section.

Before we embark on the main discussion of this chapter, we point out that SPM manipulation is not the only way of positioning nanoobjects on a surface. A variety of other approaches has been reported in the literature, using principles from optics, magnetics, electrophoresis and dielectrophoresis, which are beyond the scope of this chapter. The pros and cons of these other approaches are not fully understood. For example, optical, laser traps are normally used to move micrometer-sized particles, but they can also manipulate smaller objects. They can achieve 3-D (three-dimensional) positioning, unlike AFMs, but cannot resolve objects which are at a distance significantly below the wavelength radiation used, which is typically in the hundreds of nanometers.

8.2
Principles of Operation of the AFM

8.2.1
The Instrument and its Modes of Operation

The AFM links the macroworld inhabited by users, computers and displays to the nanoworld of the sample by using a microscopic *cantilever* that reacts to the interatomic forces between its sharp *tip* and the sample. Cantilevers are typically built from silicon or silicon nitride by using MEMS (microelectromechanical systems) mass fabrication techniques. They have typical dimensions on the order of $100 \times 20 \times 5$ μm. Tips are built at one of the ends of the cantilevers and are usually pyramidal or conical with apex diameters on the order of 10–50 nm. The tip apex has dimensions comparable to those of the sample's features and can interact with them effectively.

Figure 8.1 shows diagramatically the interaction between atoms in the tip and in the sample. The main forces involved are the long-range electrostatic force (if the tip or sample are charged), the relatively short-range van der Waals force, the capillary force (when working in ambient air, which always contains some humidity), and the repulsive force that arises when contact is established [28]. The cantilever bends under the action of these forces, and its deflection is usually measured by an optical system, as shown in Figure 8.2. Laser light bounces off the back of the cantilever, opposite to the tip, and is collected in the two halves of a photodetector. In the AFM jargon, the electrical signals output from the two photodetector halves are called A and B, and the diferential signal is called A–B. For zero deflection and a calibrated instrument, the differential signal from the detector is also zero, and in general A–B is approximately proportional to the cantilever deflection. The force between tip and sample is simply the product of the measured deflection and the cantilever's spring constant k, which can be determined in several ways [29].

Relative motion between the tip and the sample is accomplished by means of a *scanner,* which consists of piezoelectric actuators capable of imparting x, y, z displacements to the tip or, more commonly, to the sample. (We assume that the sample is on the x, y horizontal plane, and the tip is approximately aligned with the z axis.) Piezo drives react quickly and are very precise, but require high voltages on the order of 100–200 V, and have a small range of motion. They also are highly nonlinear, as we will discuss later.

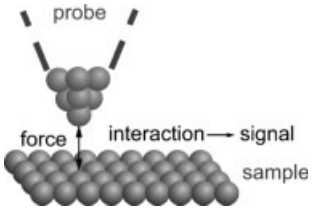


Figure 8.1 Tip sample atomic interactions.

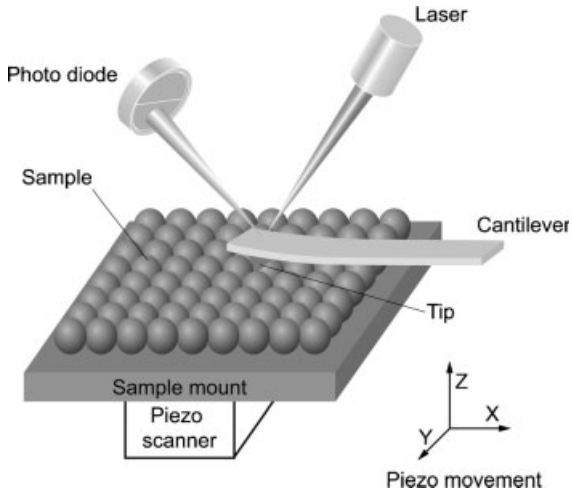


Figure 8.2 Optical detection of cantilever deflection.

In *contact mode* operation, the tip first moves in z until it contacts the sample and a desired value of the tip-sample force, called the *force setpoint*, is achieved. Then it scans the sample by moving in x, y in a raster fashion, while moving also in z under feedback control to maintain a constant force. The feedback circuitry is driven by the deviation (or error) between the (scaled) photodetector signal A–B and the force setpoint. Suppose that the tip moves in straight line along the x direction, at a constant y. When it encounters a change of height Δz of the sample, the scanner must move the sample by the same Δz in the z direction to maintain the contact force (i.e., cantilever deflection). Therefore, the amount of motion of the scanner in z at point x gives us exactly the height of the surface $z(x)$, usually called the *topography* of the surface. Now we move the tip back to the beginning of the line, increment y by Δy and scan again in the x direction. If we do this for a large number of y values we obtain a series of *line scans* that closely approximate the surface of the sample $z(x, y)$. In this example, x is called the *fast scan* direction, and y the *slow scan* direction. In practice, line scan signals are sampled (perhaps after some time-averaging) and discretized, and the output signal becomes a series of values $z(x_i, y_j)$, often called *pixel* values, since the output is a digital image. These images are normally displayed by encoding the height values z as intensities. Of course, other display options are also available, such as perspective images that provide a better feel for the three-dimensional structure of the sample, and so on. A typical AFM image contains 256×256 pixels. For this resolution and a square scan of size 1×1 μm, pixels are $\sim 4 \times 4$ nm, which is a rather large size. Therefore, very small scan sizes are necessary for precise operations.

In contact mode, essentially, the tip is pushed against the sample and dragged across it. This may damage the sample and the tip, and may dislodge nanoobjects from the surface on which they are deposited, thereby making it impossible to image them. An alternative mode of AFM operation that avoids the drawbacks of contact

mode and is kinder to the tip and sample is the *dynamic* mode, also known by other designations such as non-contact, tapping, intermitent-contact, AC, and oscillatory. Here the cantilever is vibrated at a frequency near its resonant frequency, which is typically on the order of 100–300KHz in air and 1–30KHz in water. The vibration is usually generated by a dedicated piezo drive installed at the base of the cantilever. The piezo moves the cantilever endpoint opposite to the tip up and down at a frequency near resonance, and the vibration is mechanically amplified by the cantilever, resulting in a much larger amplitude of oscillation at the tip. Alternatively, the cantilever can be coated with a magnetic material and oscillated by means of an external electromagnet. This is usually called *MAC* mode (for magnetic AC), and does not require operation near the resonance frequency since it does not rely on mechanical amplification.

The amplitude of the vibration at the output of the photodetector is computed, typically by a lock-in amplifier or an analog or digital demodulation technique, and compared with an amplitude setpoint A_{set}. Feedback circuitry drive the z piezo and adjust the vertical displacement to keep the amplitude constant – see Figure 8.3.

The principles of dynamic mode operation may be explained by approximating the vibrating tip with a harmonic oscillator in a nonlinear force field, as follows. Suppose initially that the tip is at some distance z_0 to the sample, with a force $F(z_0)$ between tip and sample. Then, the following equation of motion must be satisfied:

$$m\ddot{z}_0 + c\dot{z}_0 + kz_0 = F(z_0),$$

where m is the effective mass, c is the damping coefficient, k is the spring constant, and the dots denote derivatives with respect to time. Now, consider deviations from this point that are small compared to the tip-sample distance. Denote by z the

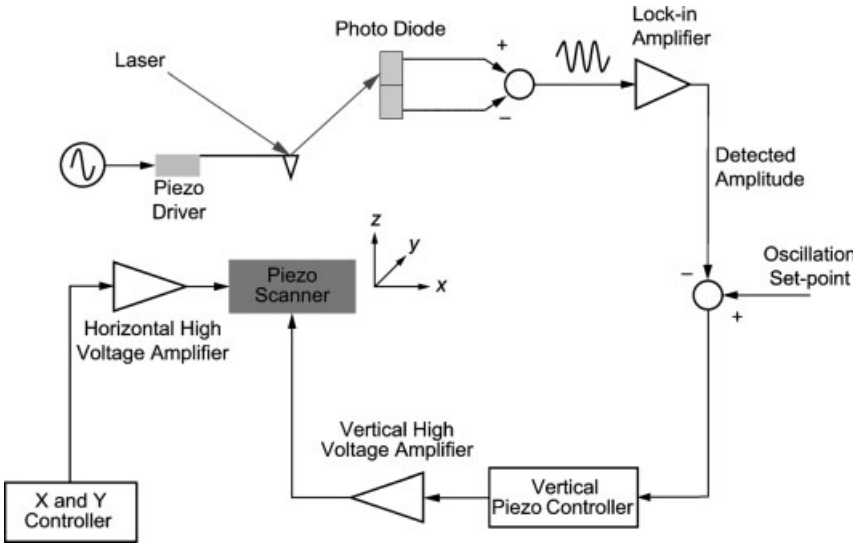


Figure 8.3 Schematic of AFM dynamic mode system.

deviation from the value z_0. The equation of motion may now be written as

$$m(\ddot{z}_0 + \ddot{z}) + c(\dot{z}_0 + \dot{z}) + k(z_0 + z) = F(z_0 + z).$$

Subtracting the two previous equations, expanding F in Taylor series and keeping only the first term yields

$$m\ddot{z} + c\dot{z} + kz = zF'(z_0),$$

where F' denotes the derivative of F with respect to z. This equation may be written as

$$m\ddot{z} + c\dot{z} + k'z = 0,$$

where

$$k' = k - F'(z_0).$$

This means that small deviations of the cantilever satisfy the equations of motion of a simple harmonic oscillator with a spring constant k', which depends on the actual spring constant of the cantilever and the gradient of the tip-sample force at the equilibrium point. Therefore, the resonance frequency changes from its initial value to

$$\varpi'_0 = \sqrt{k'/m}.$$

When the cantilever is at a large distance from the sample, the interaction forces between the two are negligible, the cantilever has a resonance frequency f_{res} corresponding to its spring constant k, and has a resonance curve (amplitude *vs.* frequency) as shown by the red curve in Figure 8.4.

Suppose now that we drive the cantilever at a frequency f_{drive}, which generally is near f_{res}. If the tip is sufficiently far from the sample for the interation force to be negligible, $F = 0$, the cantilever oscillates with the frequency f_{drive} and an amplitude A_{free} that we can read directly from the resonance curve. This is called the *free amplitude*. Now we specify an amplitude setpoint, smaller than A_{free}. For the cantilever to oscillate at this amplitude, the resonance curve must shift as shown by the blue curve in the figure. Therefore, the feedback system must move the

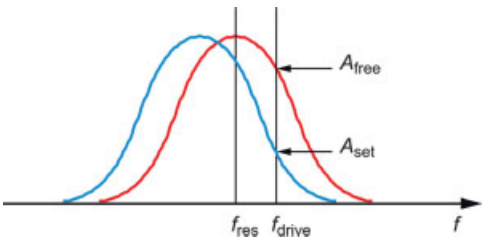


Figure 8.4 Amplitude *vs.* frequency curves when the cantilever is at a large distance from the sample (thick) and when the distance is smaller so that there is significant interaction and a concomitant shift of resonance frequency (thin).

cantilever closer to the sample until the force gradient is such that the spring constant and corresponding resonant frequency shift appropriately. We see that the DC, or average, position of the cantilever is controlled in a rather indirect manner, via the f_{drive}, A_{free}, and A_{set} parameters. (The free amplitude essentially scales the resonance curve.) A_{set} is usually specified as a percentage of A_{free}. Typical values of A_{set} are on the order of 80%. Lower setpoints imply large damping, which means that a considerable amount of the cantilever's oscillation energy is being transferred to the sample. In essence, we are tapping hard on the sample, and this is usually undesirable.

The theory just presented is simple and provides an intuitive understanding of the dynamic mode operation. Unfortunately, it is predicated on a linearization about an operating point, and is valid only for small oscillation amplitudes. Usually, however, the AFM is operated with a setpoint that implies a relatively large amplitude and causes the tip to hit ("tap" on) the surface of the sample at the lower part of each oscillation cycle. The actual behavior of the cantilever when tapping is involved is rather complicated – see for example [30, 31]. But in normal imaging conditions the oscillation amplitude varies approximately linearly with the DC tip position, as shown in Figure 8.5. Note that such A-d (amplitude-distance) curves vary from cantilever to cantilever and depend on several parameters.

The feedback circuitry in dynamic mode maintain a constant amplitude, and therefore a constant distance to the sample. (Here we are assuming that the sample is of a homogeneous material, or at least that the tip-sample forces do not vary over the sample's extent.) Scanning the tip over the sample in dynamic mode produces a topographic image of the sample. There are other modes of AFM operation, but the two discussed above are the most common and important.

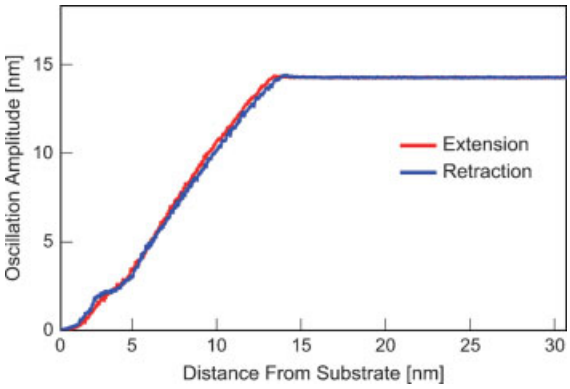


Figure 8.5 Experimental amplitude-distance curve obtained with the MFP-3D AFM (Asylum Research). Cantilever resonance frequency 240.654KHz, drive frequency 240.556KHz, spring constant ~25 N/m, free amplitude 14.2 nm. The zero point for the distance from the surface was inferred by extrapolating the A-d curve to the zero amplitude point.

More information on AFM theory and practice are available for example in [32–34].

8.2.2 Spatial Uncertainties

Let us now turn our attention to the sources of positional errors in AFM operation. These give rise to spatial uncertainty and are important for accurate imaging and especially for nanomanipulation. User intervention is normally used to compensate for spatial uncertainties in nanomanipulation, but extensive user interaction is slow and labor intensive, and therefore is severely limited in the complexity of structures it can construct. Automatic operation is highly desirable but cannot be accomplished without compensating for spatial uncertainties, as we will show below. Compensation techniques are described later in this chapter, in Section 8.4.2.

We noted earlier that the output of a line scan along the x direction is the topography $z(x)$. In the actual implementation this is not quite true. If no error compensation is used, the output is $V_z(V_x)$, where V_z and V_x are the voltages applied to the z and x piezos. In an ideal situation these voltages would be linearly related to the piezo extensions, and the signals $V_z(V_x)$ and $z(x)$ would coincide modulo scale factors. But in practice they don't.

There are many nonlinearities involved. Some of these are normally taken into account by AFM vendors' hardware and software, for example, non-linearities in the voltage-extension relationship for the piezos, coupling between the different axes of motion, and so on. The most pernicious are drift, creep and hysteresis. As far as we know, at the time of this writing (2006), drift is not adequately compensated for in any commercial instrument, and creep and hysteresis are negligible only in top of the line AFMs that have feedback control for the x, y directions, and whose controller noise r.m.s. is under 1 nm. The vast majority of AFMs in use today either have no x, y feedback or have noise levels on their feedback circuitry that are too large for precise lithography and manipulation. Open-loop operation with a small scan size (e.g., 1×1 μm) is preferable for nanomanipulation operations with such instruments.

Drift is caused by changes of temperature in an instrument made from several materials with different coefficients of expansion. At the very low temperatures often used for atomic manipulation the effects are negligible, but at room temperature they can be quite large. Figure 8.6 shows four AFM images of gold nanoparticles with 15 nm diameters, taken at successive times, 8 min apart. The particles appear to be moving, but in reality they are fixed on the substrate surface. The piezos are driven by the same voltage signals in all the panels of the figure, and have the same extensions, but the position of the tip relative to the sample is the sum of the piezo extension and the drift, and therefore changes with time. Experimental observations in our lab indicate that drift is a translation with speeds on the order of 0.01–0.1 nm/s. The drift velocity remains approximately constant for several minutes, and then appears to change randomly to another value.

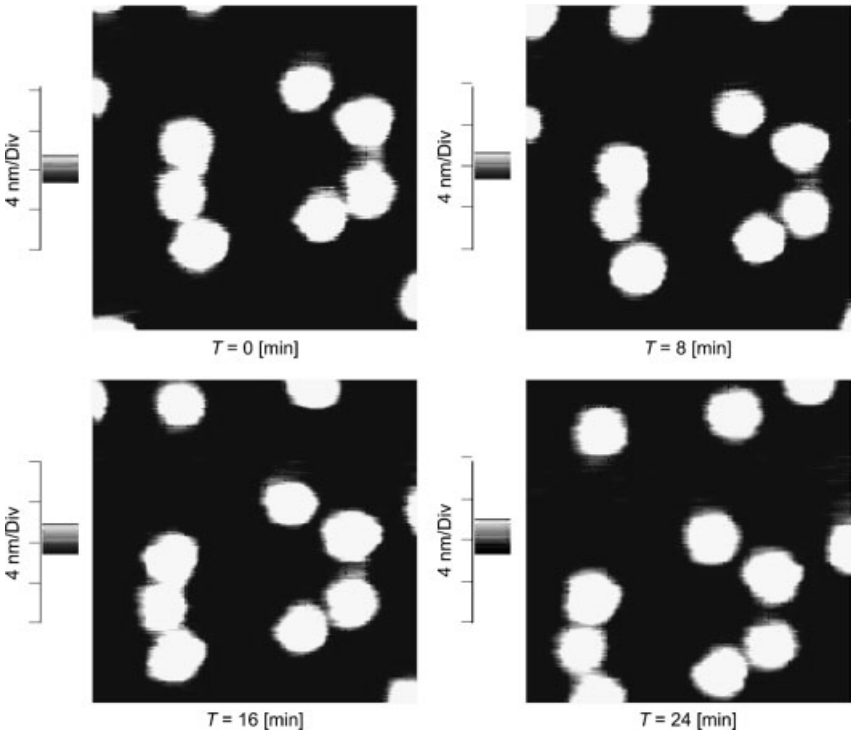


Figure 8.6 Successive images taken at 8 min intervals showing the effects of thermal drift. Gold nanoparticles with nominal diameter 15 nm on mica coated with poly-L-lysine, 1Hz scan rate, Autoprobe CP-R AFM (Veeco). Reproduced with kind permission from [71].

Suppose now that we were trying to manipulate a large set of nanoparticles similar to those in Figure 8.6, and that the manipulation operations would take a total time of 1 hour. Assuming an average drift velocity of 0.05 nm/s, the total drift after 1 hour would be 180 nm. If we relied on the original images and did not compensate for drift, it is clear that as time went by we would completely miss most of the 15 nm particles.

After drift, creep is another major source of spatial uncertainty in AFMs. A piezo actuator commanded to move by a certain distance first responds very quickly and moves by ~70–90% of the commanded distance, and then slowly "creeps" to the final position – see Figure 8.7. Creep is especially noticeable for large motions. Successive images of the (nominally) same area taken after a large tip displacement show an apparent motion of the features in the area. The effects of creep can last several minutes and are sufficiently large to foil manipulation attempts, especially for small particles with dimensions on the order of 10 nm. Creep can be avoided by waiting for several minutes after any large tip motion, but this is obviously a very inefficient approach.

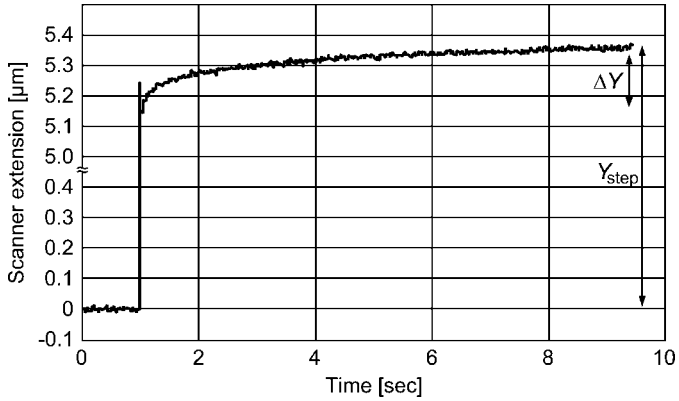


Figure 8.7 Experimental curve showing a scanner response to a 5.4 µm step function. Reproduced with kind permission from [95].

Hysteresis is also present in piezo actuators and has non-negligible effects – see Figure 8.8. Hysteresis is a nonlinear process with memory. The extension of a piezo depends not only on the currently applied voltage, but also on past extremal values.

Finally, the images of the features that appear in a topography scan differ from the actual physical features in the sample because of tip effects. The tip functions as a low-pass filter, and broadens the images. To a first approximation, the image's lateral dimensions of a feature equal the true dimensions plus the tip diameter. Algorithms are known for combating this effect [35]. Note that the vertical dimensions of a feature's image are not affected by the tip's dimensions.

Compensation of spatial uncertainties due to drift, creep and hysteresis in AFMs will be discussed later, in Section 8.4.2, in the context of automated nano-manipulation.

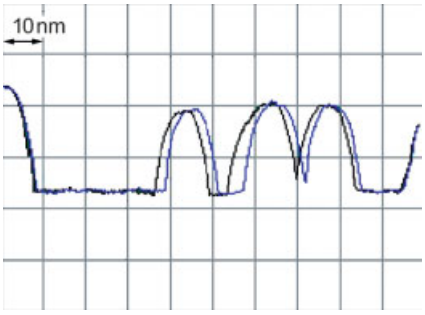


Figure 8.8 Hysteresis effects. Left-to-right vs. right-to-left single-line scans of 15 nm Au particles on mica. Scan size 100 nm.

8.3
Nanomanipulation: Principles and Approaches

8.3.1
LMR Nanomanipulation by Pushing

Here we discuss the approach to nanomanipulation that has been under development at USCs Laboratory for Molecular Robotics (LMR) over the last decade. It was first presented at the fourth International Conference on Nanometer-Scale Science and Technology, Beijing, P.R. China, September 8–12, 1996, and later reported in a string of papers [36–41]. Other approaches are considered in the next subsection.

We begin by preparing a sample with nanoparticles or other structures to be manipulated. A typical sample consists of a mica surface coated with poly-L-lysine, on which we deposit Au nanoparticles. The coating is needed because freshly cleaved mica is negatively charged, and so are the nanoparticles; the poly-L-lysine attaches to the mica and offers a positively-charged surface to the nanoparticles. We have also used other surfaces such as (oxidized) silicon, glass and ITO (indium tin oxide), other coatings such as silane layers [42], other particles such as latex, silver or CdSe, and rods or wires of various kinds. We typically manipulate particles with diameters between 5 and 30 nm, but have occasionally moved particles as small as 2 nm and as large as 100 nm. In all cases the structures to be moved are weakly attached to the underlying surfaces and cannot be imaged by contact mode AFM. We image them in dynamic mode, apply a flattening procedure to remove any potential surface tilt, and then proceed with the manipulation. The bulk of our experiments have been conducted in ambient air at room temperature and without humidity control, but we also have demonstrated manipulation in a liquid environment [43]. We use stiff cantilevers, with spring constants >10 N/m.

The nanomanipulation process is very simple. We move in a straight line with an oscillating tip towards the center of a particle and, before reaching the particle, turn off the z feedback. We turn the feedback on when we reach the desired end of the particle trajectory. With the feedback off, the tip does not move up to keep constant distance to the sample when it encounters a nanoparticle. Rather, it hits the particle and pushes it. We use the same dynamic AFM parameters for pushing as for imaging, but sometimes we force the tip to approach the surface by applying directly a command to move by Δz immediately after turning off the feedback.

Figure 8.9 shows experimental data acquired during a pushing operation for a 15 nm Au particle on mica. The two vertical dashed lines indicate the points where the feedback is turned off and on. The top trace (A) is simply the topography signal acquired by a single line scan in dynamic mode. The next trace (B) is the topography signal during the push. The topography signal is flat while the feedback is off because the tip does not move up and down to follow the sample topography. Observe that as soon as the feedback is turned back on we immediately get a non-null topography signal that indicates that the tip was somewhat below the top of the particle at the end

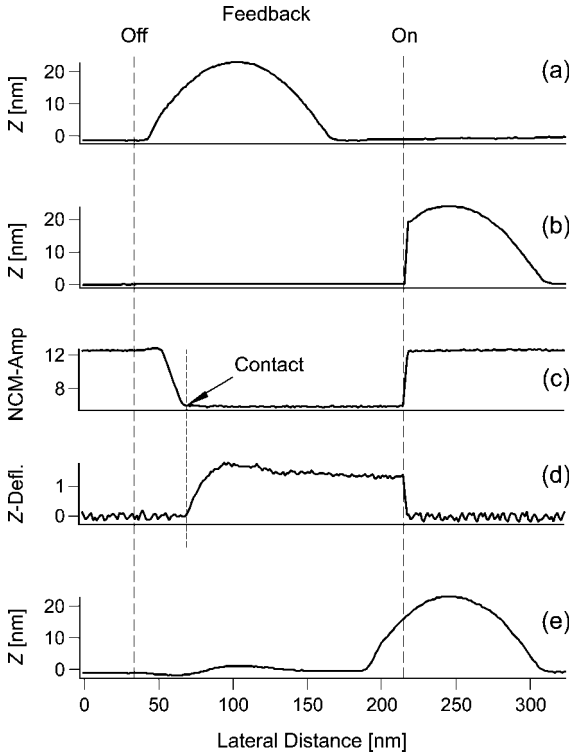


Figure 8.9 Data acquired during a manipulation operation. (a) Dynamic-mode single line scan image of the particle before manipulation. (b) Topography signal during the push. (c) Vibration amplitude during the push. (d) Average position of the antilever during the push. (e) Image of the particle after the manipulation. AutoProbe AFM, 15 nm Au particles on mica with poly-L-lysine, cantilevers with (nominal) spring constant 13 N/m. Reproduced with kind permission from [37].

of the manipulation. We conclude from these data that the tip is pushing the particle forward rather than dragging it behind itself.

Trace C shows the amplitude of the vibration during the manipulation. The amplitude is constant at the setpoint value when the feedback is on, but it decreases as the tip approaches the particle with the feedback off, and eventually reaches zero and stays at zero for the reminder of the push. At the same time that the amplitude decreases, the average (DC) value of the cantilever deflection increases and then reaches an approximately constant level – trace D in the figure. Finally, trace E is a single line scan after the push.

We interpret the data in Figure 8.9 as follows. When the tip approaches the particle with the feedback off, it starts to exchange energy with the particle and the vibration amplitude decreases, much like in standard A-d curves (see Section 8.2.1). When the vibration goes to zero, the tip touches the particle, and remains in contact with it until the feedback is turned on. While the tip contacts the particle the cantilever starts to

"climb" the particle, and the DC deflection increases. When enough force is exerted on the particle for it to overcome the surface adhesion forces, the particle moves, and the deflection (and hence the force) remains approximately constant. Our experiments reveal that there is a deflection (or force) threshold above which the particle moves.

For successful pushing, the trajectory has to pass close to the center of the particle. Here the spatial uncertainties discussed in Section 8.2.2 are a major source of problems. We address these problems in the interactive version of our Probe Control Software (PCS) as follows. The user draws a line over the image and instructs the AFM to scan along that line and output the corresponding topography signal. The user moves the line over (or perhaps near) the particle to be pushed until he or she detects the maximum height of the particle – see Figure 8.10. This indicates that the line is going through the center of the particle. Usually this is several nm away from the apparent center because of drift and other spatial uncertainties. After the center is found, the user sets two points along the trajectory for turning the feedback off and on, and instructs the AFM to proceed with the push. The result can be assessed immediately by looking at the single line scan after the push – see trace E in Figure 8.9.

The amplitude and deflection signals – see traces C and D in Figure 8.9 – are useful to assess whether a pushing operation is proceeding normally. For example, we have observed experimentally that when we "loose" a particle (i.e., when it does not move as far as specified) the deflection drops to zero prematurely. In principle, one could monitor the amplitude and deflection signals while pushing and, for example, stop and locate the particle when the signals are not as expected. In practice, however, this may require substantial modifications to the controller, if decisions are to be made automatically based on this information while the manipulation operation is taking place.

How reliably can particles be moved by the LMR aproach? Figure 8.11 attempts to answer this question. Observe that operations in which the commanded motion is below 50 nm are very successful, whereas for distances ~80 nm the actual and

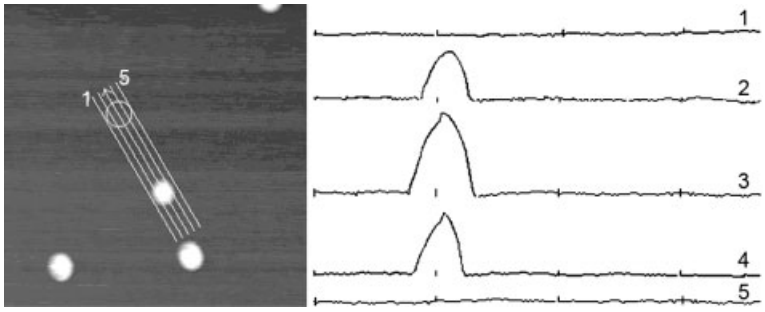


Figure 8.10 Interactive search for the center of a particle with single line scans. Line 3 has the largest peak and is chosen as trajectory for the pushing operation. Reproduced with kind permission from [36].

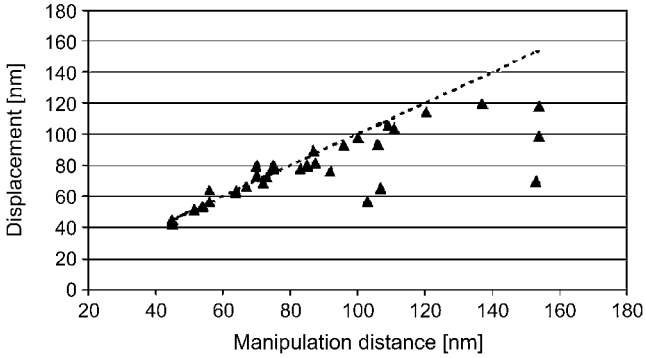


Figure 8.11 Reliability plot, showing the actual distance a particle moved as a function of the commanded manipulation distance. AutoProbe CP-R, 15 nm Au particles on mica, interactive pushing. The dashed line corresponds to perfect pushing, with equal actual and commanded displacements. Reproduced with kind permission from [90].

desired displacements of the particles begin to differ considerably. The reasons why pushing over large distances is unsuccessful are not fully understood.

8.3.2 Other Approaches

At the LMR we have experimented with several other approaches to pushing nanoparticles. We had occasional success with all of these approaches, but did not achieve reproducible, controlled manipulation. The reliability of these protocols has been until now much lower than that of the standard method discussed in the previous subsection. Here is a short description of these various protocols. They all begin with imaging in dynamic mode and finding a particle's center interactively, as discussed earlier.

- As the tip approaches the particle, instead of turning the feedback off and on, change the amplitude setpoint so that the tip gets closer to the surface.
- Move towards the particle while tapping hard on the substrate and then turn the feedback off an on. This appears to induce a "lateral push", in which the cantilever deflection does not increase as in the standard pushing protocol of the previous subsection.
- Approach the particle while moving in a zig–zag pattern, in a direction normal to the desired trajectory and with the feedback off. This appears to simulate pushing with a linear edge rather than a round tip, and has been reported in [44, 45].

The first published reports that demonstrated manipulation of nanoparticles with the AFM came from Purdue University in the US [46] and the University of Lund in

Sweden [47]. The Purdue group pushed 10–20 nm gold clusters on graphite or WSe_2 substrates with an AFM, in a nitrogen environment at room temperature [46]. They image with non-contact AFM, but then stop the cantilever oscillation, approach the substrate until contact, disable the feedback, and push. Samuelson's group at the University of Lund succeeded in pushing galium arsenide (GaAs) nanoparticles of sizes in the order of 30 nm on a GaAs substrate at room temperature in air [47]. They use an AFM in non-contact mode, approach the particles, disable the z feedback and push. This is the protocol investigated in detail later on by the LMR, and discussed in the previous subsection. Essentially the same protocol is used in [48] to push Ag nanoparticles. They observe that the vibration amplitude decreases as the particle is approached, and then essentially vanishes during pushing, which agrees with the findings in our own laboratory.

Lieber's group at Harvard has moved nanocrystals of molybdenium oxide (MoO_3) on a molybdenium disulfite (MoS_2) surface in a nitrogen environment by using a series of contact AFM scans with large force setpoints [49]. The nanoManipulator group at the University of North Carolina at Chapel Hill moves particles by increasing the contact force, under user control through a haptic device [50]. Sitti's group reports manipulation of Au-coated latex particles with nominal diameters 242 and 484 nm on a Si substrate with accuracies on the order of 20–30 nm [51]. First, they move the tip until it contacts the surface, and then move it horizontally to a point near the particle, and up by a predetermined amount. Next, they move against the particle using feedback to maintain either constant height or constant force on the particle. In constant-height pushing, the force signal exhibits several characteristic signatures that may be interpreted as signifying that the particle is sliding, rolling or rotating. Constant-force pushing is equivalent to contact-mode manipulation. Xi's group at Michigan State University has demonstrated pushing of latex nanoparticles with 110 nm diameters on a polycarbonate surface by two methods [52]. The first consists of scanning in contact mode with a high force. The second disables the feedback and moves the tip open loop along a computed trajectory based on a model of the surface acquired by a previous scan. This requires an accurate model of the surface.

Theil Hansen and coworkers moved Fe particles on a GaAs substrate by approaching a particle in dynamic mode, switching to contact mode and pushing with the feedback on [53]. This has the advantage that the pushing force can be controlled by the AFM. However, switching modes is a non-trivial operation that can cause damaging transients. (In the AutoProbe CP-R AFM that we use routinely for nanomanipulation at LMR, such a switch is not allowed by the vendor's software.) Furthermore, switching from tapping to contact mode implies that the tip in contact mode does not touch the same point of the surface it was tapping on, because the cantilever normally is not horizontal – see Figure 8.12.

The LMR and related protocols are essentially "open-loop", because it is virtually impossible to incorporate the force sensed by the cantilever into a feedback loop during actuation for commercial AFMs. For example, in the AutoProbe CP-R we are completely "blind" during a pushing operation. We can

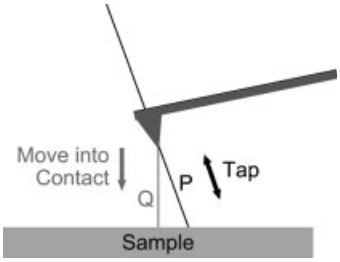


Figure 8.12 A cantilever is initially tapping on a sample at point P. When the AFM is switched into contact mode and the vibration stops, the tip must approach the sample to maintain contact. However, the contact will be at point Q, not P.

record the force (deflection) signal while pushing, but cannot do anything with it until the motion stops. To do otherwise would require a major change to the controller. On the other hand, it is not difficult to make the force signal available for visualization in interactive pushing, and several research groups have reported such capabilities. By developing their own controllers, Sitti and co-workers have been able to use the force signal during pushing, primarily to determine when an operation should be stopped because it is not going to succeed. When the tip-particle force drops to zero the particle is no longer being pushed. One should stop the motion, locate the particle and schedule a new operation to deliver it to its target position.

The AFM is both an imaging device and a manipulator but not both simultaneously. For example, it would be useful to see a particle while it is being pushed, but this cannot be done solely with an AFM. An interesting approach that provides real-time visualization consists of operating the AFM within the chamber of a Scanning Electron Microscope (SEM), or sometimes a Transmission Electron Microscope (TEM). The motion of the tip can then be monitored by the electron microscope and known techniques for visual feedback developed at larger spatial scales can be deployed [54].

The AFM-SEM approach was pioneered by Sato's group for microscopic objects [55, 56], and has been used successfully by several groups [50, 57–59]. In some of this work an AFM cantilever is used as an end effector for a specially-built micromanipulator. Working inside an SEM has its drawbacks: electron microscopes are expensive instruments, they are less precise than AFMs, require more elaborate sample preparation, and normally operate in a vacuum environment, which precludes their use for certain applications, for example, in biology.

All of the work on AFM nanomanipulation discussed above involves essentially pushing objects on a flat surface. Pushing nanoparticles over steps [37] and onto other particles [60] has been demonstrated by our group, but this is a very rudimentary 3-D capability. More sophisticated 3-D tasks would be feasible if there was the equivalent of a macroscopic "pick-and-place" operation for nanoparticles. (Pick-and-place is possible with atoms and small molecules, as noted in the

Introduction.) We know of only one report in which nanoparticles are controllably picked up by the AFM tip and then deposited elsewhere [61]. They succeed in picking up Si nanocrystals deposited by silane CVD (chemical vapor deposition) on a Si surface. They place the tip in contact with the particle and apply successive voltage pulses of opposite polarity to the AFM tip. The tip is then moved to a target location, lowered until there is contact with the surface, and again a series of pulses is applied. This work requires a dry atmosphere and the experiments were performed in a nitrogen environment. The process appears to have limited reproducibility. Diaz and co-workers report a pick-and-place process that uses redox reactions on the tip to pick up and deposit particles, but they only demonstrate it for large, micrometer scale particles [10]. Note that it is not difficult to pick up a nanoobject with a tip, even by just using van der Waals forces. What is very hard to do is to controllably release the object at a target location.

8.3.3
Manipulation and Assembly of Nanostructures

The majority of the experimental work on nanomanipulation has been conducted with nanoparticles. These are simple but interesting nanobjects because many types of them are available (metalic, semiconducting, magnetic, etc.), they often are monodisperse (i.e., have the same size), they can be smaller and more uniform than structures made by other means such as electron-beam lithography, and can be arranged into arbitrary patterns by nanomanipulation. Patterns of nanoparticles may be useful in themselves, or they may serve as templates for constructing other structures [62]. We present examples of both below.

Figure 8.13 illustrates the use of particle patterns to store digital information. The particles are located at the nodes of a uniform grid. We interpret the presence of a particle at a node as a digital "1" and its absence as a "0". Each row of particles represents an ASCII character. From top to bottom, this structure encodes the characters "LMR". These are 15 nm Au nanoparticles, and the internode distance along a row or column is $\sim$100 nm, which corresponds to a density of 10 Gbit/cm^2. This is considerably higher than the current compact disk (CD) density. By using smaller particles and closer spacing, higher densities are achievable. In addition, this structure is editable, simply by moving the particles.

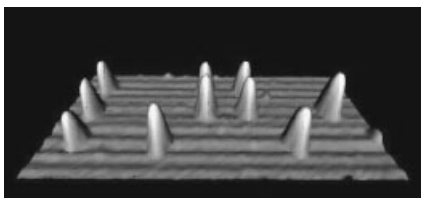


Figure 8.13 The ASCII characters "LMR" encoded in the positions of 15 nm Au nanoparticles on mica.

Nanoparticle patterns can also be used as a resist, as shown by research at the University of Konstanz, Germany [63]. They started with a self-assembled regular pattern and deposited material so as to fill the space between the particles. By etching the particles away, they obtained the complement of the original pattern. In a similar vein, a Japanese/English group used regular nanoparticle patterns as templates in a process that involves both etching and growth [64]. They deposited nanoparticles on a Si substrate and then etched the Si. As the substrate was etched away, reaction products condensated around the nanoparticles and formed regular patterns of pillars with diameters on the order of a few nm. Similar results were obtained in [65]. Another use of Au nanoparticles as a mask is reported in [66], where they demonstrate that the particles can serve as an anti-oxidation mask to prevent the AFM-tip induced oxidation of a modified Si surface. Subsequent etching produces Si nanopillars. In all of these cases, it should be possible to perform similar operations for arbitrary patterns constructed by nanomanipulation. Nanoparticle manipulation may also be an effective way to build templates or molds, for example for imprinting techniques that construct a large number of structures in a parallel fashion by pressing a template against a substrate [67]. Nanoparticle-based templates can be more uniform and have smaller features than those built by other means such as electron-beam lithography. Applications to template or mold making, however, have not yet been demonstrated.

Nanoparticle manipulation with the AFM has been used to build prototype structures for several nanodevices such as single-electron transistors [44, 68], plasmonic waveguides [69, 70], and quantum-dot cellular automata gates [71].

Most nanoparticles are approximately spherical, although some of the nanoparticles used in the research cited above (e.g., [47]) are more like "islands", and have vertical dimensions smaller than their horizontal counterparts. Structures with other shapes have also been nanomanipulated. Rods, wires and tubes have been investigated extensively. They normally require a series of pushes to reach a target position because they tend to rotate. And sometimes they deform or even break, rather than simply move on the surface. At the LMR we have manipulated Au rods with diameters ~ 10 nm and lengths ~ 70 nm deposited on a SiO_2/Si (1 1 1) substrate modified with MPMDMS (3-mercaptopropylmethyldimethoxysilane) [72]. Xi's group has demonstrated manipulation of Ag rods with diameters ~ 110 nm on a polycarbonate surface [73].

Several groups have manipulated carbon nanotubes (CNTs). Avouris and co-workers at the IBM Yorktown laboratories moved multiwall CNTs on a hydrogen-passivated Si surface by using contact mode AFM [74]. They also showed that the nanotubes can be bent and cut by the AFM tip. The nanoManipulator group in North Carolina reported rolling and sliding of CNTs on a graphite surface [75]. Other work in CNT manipulation include [57, 58, 76, 77], Roschier an co-workers [78] who built a single-electron transistor by manipulating a multiwall CNT so as to connect it to electrodes made by electron-beam lithography. Several of these groups also used the AFM tip to probe the mechanical properties of the CNTs – see for example [50, 57]. Nanowires can also be

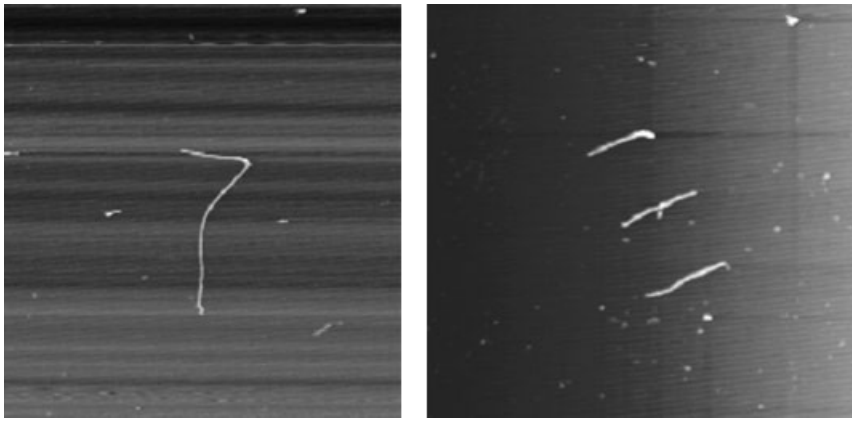


Figure 8.14 Cutting a SnO_2 nanowire and making an array with the resulting three pieces by using the AFM tip.

manipulated by the AFM. Figure 8.14 shows on the left a SnO_2 wire with a diameter of ~10 nm and a length of ~9 μm, and on the right the result of cutting the wire in two spots and then moving the three smaller wires. The manipulation was done at the LMR by using our standard protocols.

For many applications, particles and other nanoscale objects must be linked together to form a single, larger object. Linking may be accomplished by various methods: chemically, by using material deposition, sintering, and "welding". We showed that Au nanoparticles can be connected chemically by using linkers with thiol functional ends [79]. This can be done in two ways: (1) the particles are first functionalized with the di-thiols, then deposited and manipulated against one another to form the target structure, or (2) the manipulation is done first and then the thiol treatment is applied. In either case the result appears to be the same. The resulting assemblies can then be manipulated by using the same protocols, and joined to make larger assemblies. Therefore, we have demonstrated that hierarchical assembly is possible at the nanoscale [80]. Figure 8.15 shows on the left clusters of 2 and 3 particles, which were constructed by manipulating individual particles (initial configuration not shown). On the right is a ring-like structure obtained by moving the clusters on the left.

A different approach is reported in [81]: nanoparticles are manipulated to form a target structure, which is then grown by deposition of additional material. Growth is accomplished essentially by electroless deposition, by immersing the sample in a hydroxilamine seeding solution. Figure 8.16 illustrates the results. On the top left is a "wire" made by manipulating Au particles, and on the top right is a single line scan through the centerline of the structure, showing that the height of the particles is ~8 nm. On the bottom left is the structure after deposition by hydroxilamine seeding, and on the bottom right is a single line scan, which now shows a particle size of ~20 nm. The initial structure looks like a continuous wire

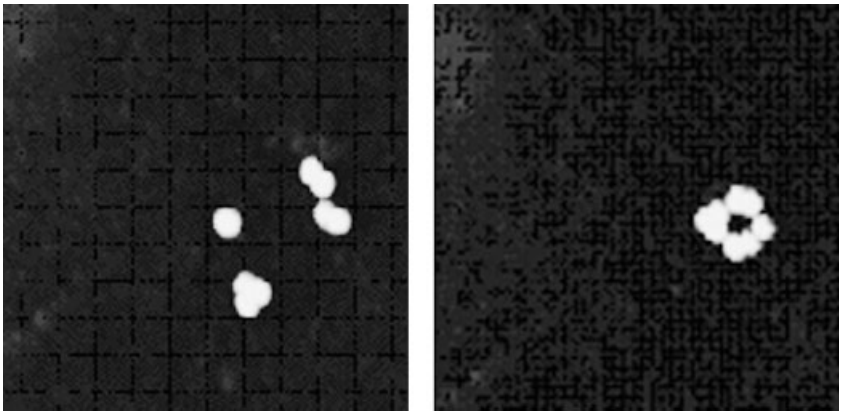


Figure 8.15 Hierarchical assembly. Thiolated 27 nm Au nanoparticles on Si coated with poly-L-lysine, 800 × 800 nm scan size. Reproduced with kind permission from [79].

in the figure but is not mechanically stable; touching it with the tip causes the structure to fall apart. In contrast, the final structure is a solid wire. One disadvantage of this method is that after the seeding the structures can no longer be moved on the substrate surface.

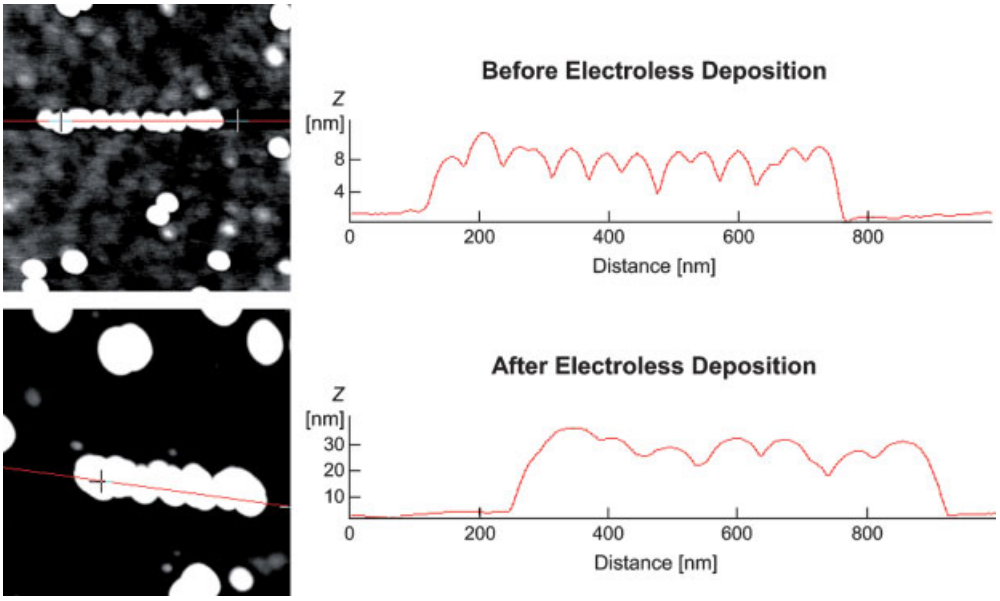


Figure 8.16 Linking nanoparticles by hydroxilamine seeding of a template built by manipulation. 8 nm Au particles on Si modified with aminopropyltrimethoxysilane (APTS). Reproduced with kind permission from [81].

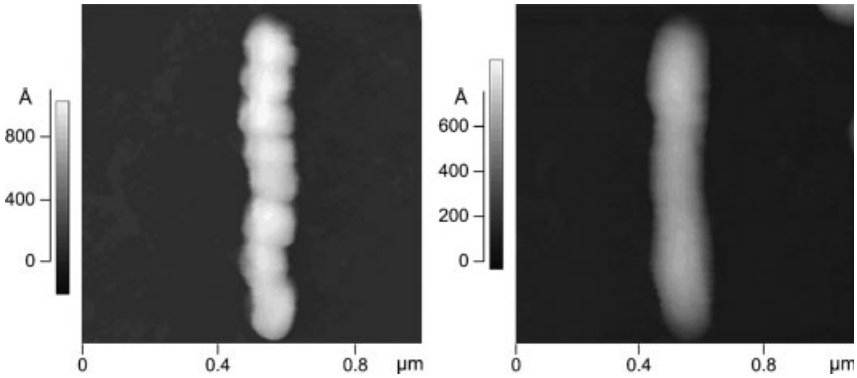


Figure 8.17 Linking nanoparticles by sintering a template built by nanomanipulation. 100 nm latex particles on Si. Reproduced with kind permission from [82].

Finally, we have demonstrated an even simpler approach to particle linking, which is based on sintering a structure after its template is built by nanomanipulation [82]. Figure 8.17 shows on the left a "wire" built by manipulation of latex particles with 100 nm diameter. On the right is the result of heating the sample at $\sim$160° for $\sim$10 min. The particles form a contiguous, solid structure. This process works very well for latex particles, but, unfortunately does not appear to be applicable to Au (and perhaps other metalic) particles.

Another interesting approach to joining nanoobjects is "welding", usually done within an SEM by using the electron beam. Contamination within the SEM chamber usually suffices to generate carbonaceous residues at the beam's target, and this can be used to link objects such as nanotubes [57, 76]. Similar approaches but using an environmental SEM and a field emission SEM are reported, respectively, in [77, 83]. Here they "solder" by inducing deposition of conductive materials with the electron beam in the presence of a source of a precursor.

8.4 Manipulation Systems

8.4.1 Interactive Systems

We discussed briefly in Section 8.3.1 the interactive manipulation capabilities of the LMR software. Ours is an unsophisticated system which provides a set of minimal capabilities a user needs to manipulate nanoobjects with the AFM. Since the beginnings of the LMR we have focused on automation – see the next subsection – and therefore did not invest resources in user interfacing.

Much more sophisticated interfaces have been developed by others. Hollis' group at the IBM Yorktown laboratory (now at Carnegie Mellon University) built an interface to an STM in which the user could drive the tip over the sample by moving a mechanical wrist [84]. The z servo signal is fed back to the wrist so that the user feels the topography of the surface as wrist vertical motions. This force, however, is not (a scaled version of) the actual force between tip and sample.

The nanoManipulator group in North Carolina has developed virtual reality user interfaces, first for STMs and then for AFMs [45, 50, 85, 86]. In the AFM interface a user can either be in imaging or manipulation mode. During imaging, the topographical data collected by the AFM is presented to the user in virtual reality, as a 3-D display. In addition, the user can feel the surface by using a haptic device, as if moving a stylus over a hard surface. Note, however, that the forces felt through the haptic device are not the cantilever-sensed forces, but rather are forces computed by standard virtual reality techniques so as to simulate the feel of a surface that approximates the measured topography of the sample. In the imaging mode, the user haptic input does not control the actual motion of the instrument, but rather the position of a virtual hand over the image of the surface. In contrast, the hand can be used to move the tip over the sample in manipulation mode. As the hand moves in virtual space and the tip moves correspondingly over the sample, the topography data generated by the AFM is used on the fly to compute a planar approximation to the surface. The user feels this approximated surface through the haptic stylus. Although the user does not feel the actual forces sensed by the cantilever, he or she can control the force applied to the sample during manipulation by using a set of knobs.

Sitti and co-workers also implement a virtual reality graphics interface, and add a one degree-of-freedom haptic device [51, 87]. Through a bilateral feedback system based on theoretical models of the forces between tip, sample and particle, the user can drive the tip over the sample by using a mouse, while at the same time feeling with the haptic device the forces experienced by the cantilever.

Xi's group has developed an augmented reality system in which cantilever forces are reflected in a haptic device [88, 89]. They develop a theoretical model for the interaction forces between tip, object and surface, and use it to compute the position of the tip based on the real-time force being measured. The visual display in a small window around the point of manipulation is updated in real-time to reflect the computed particle position. Thus, a user can follow the (computed) motion of the particle in real time during the manipulation.

8.4.2 Automated Systems

The automatic assembly of nanoobject patterns with the AFM consists of planning and executing the motions required for moving a set of objects from a given initial configuration into a goal configuration. The initial state usually corresponds to nanoobjects randomly dispersed on a surface.

As far as we know, there are only two systems today that are capable of building nanoobject patterns *automatically* by AFM nanomanipulation. One is being developed by Xi's group at Michigan State University, and the other at the LMR [73, 90]. The two systems use different planning algorithms and pushing protocols. Xi's system addresses at length the issues that arise in nanorod manipulation, and therefore is more general than ours, which focuses on nanoparticles. On the other hand, the Michigan State system has a more rudimentary drift compensator than ours, and does not compensate for creep or hysteresis, which are important for the manipulation of small objects, with dimensions $\sim$10 nm or less. The manipulation tasks demonstrated in [73] involve objects which are roughly one order of magnitude larger than those we normally manipulate, and the positional errors in the final structures shown in the figures of [73] also appear to be similarly larger than ours.

In the remainder of this section we describe the LMR automatic manipulation systems, from the top down, starting with high-level planning and ending with the system software architecture.

The input to the planner consists of a specification for a goal assembly of nanoparticles, and an initial arrangement that is obtained by imaging a physical sample with a compensated AFM (compensation is discussed below). In an initial step the planner assigns particles to target locations by using the Hungarian algorithm for bipartite matching, which is optimal [91]. It uses direct, straight-line paths if they are collision free, or indirect paths around obstacles computed by the optimal visibility algorithm [92]. Next, the planner computes a sequence of positioning paths, to connect the locations of the tip at the end of a push (determined in the previous step) and at the beginning of another push. This is done by a greedy algorithm, which sequentially selects the shortest paths. It is sub-optimal but performs well in practice. In a general case collisions between particles may arise. The planner handles collisions by exploiting the fact that all particles are assumed identical. It simulates the sequence of operations previously computed, at each step updating the state of the particle arrangement. If a collision is detected, it swaps operations, and does this recursively because solving one collision problem may generate new ones.

The planner just outlined is the second one we write. Our first planner, developed several years ago [93], was more complicated and slower, and did not perform better than the current one. However, we abandoned work on planning at that time not because of planner problems, but rather because we could not implement reliably the primitive operation assumed by the planner, which is simply to move from an initial point P to another goal point Q. The spatial uncertainties associated with the AFM – see Section 8.2.2 – were such that after a few operations we could no longer find the particles and push them without user interaction, and the task could not be completed automatically. We embarked on a research program aimed at compensation of uncertainties, and developed the compensators described briefly below. Details are available in [71, 94, 95].

The drift compensator is based on Kalman filtering, a standard technique in robotics and dynamic systems. We assume a simple (but incorrect) model for the

time evolution of the drift. The model can be used to predict future values of the drift, but these values will become increasingly wrong as time goes by, because the model is not perfect. The Kalman equations provide us with means to estimate the prediction error. When this value exceeds a threshold, drift measurements are scheduled, and the measured and predicted values are combined to produce better estimates, again by using the Kalman equations. A decade-long series of experiments indicates that the drift is accurately approximated as a translation, with a direction and speed that vary slowly. The estimated drift values obtained from the Kalman filter are added as offsets to all the motion commands of the AFM, thus compensating for this translation. Drift measurement techniques require that the AFM tip move on the sample to acquire data. Therefore, manipulation operations must be suspended when a measurement is needed. In contrast, when the filter is in prediction mode the offsets can be calculated very quickly and used to update the coordinates without interrupting the manipulation task.

Creep and hysteresis compensation is achieved through a feedforward scheme. A model for the two phenomena together is constructed as explained below, using a Prandtl-Ishlinskii operator [95]. This operator has the important property of invertibility. The inverse operator is computed and the desired trajectory is fed to the inverse system. The result is the signal required to drive the AFM piezos so as to follow the goal trajectory, assuming that the model is perfect. The model, of course, is not perfect, but experimental results show that it is sufficiently accurate for obtaining very good results – an order of magnitude decrease on the effects of creep and hysteresis has been verified experimentally. Creep is modeled by a linear term plus a superposition of exponentially decaying terms, with different time constants. Hysteresis is modeled by a superposition of operators which are essentially simple hysteresis loops. The piezo extension is the sum of the values of creep and hysteresis obtained from their models, and can be expressed in terms of a Prandtl-Ishlinskii operator. The combined model depends on several parameters, which can be estimated by analyzing the AFM topography signal for a line scan over a few particles. The line should span the entire region in which the manipulations will take place. The parameters are valid as long as the scan size of the AFM is not changed, and can be computed rapidly by running the tip back and forth a few times with the compensator on. Details may be found in [95].

Running both compensators together results in a software-compensated AFM with sufficiently low spatial uncertainties to provide a reliable implementation of the most basic robotic primitive "Move from point P to point Q" on the sample. However, this is not sufficient to reliably push particles between arbitrary points because long pushes tend to be unreliable – see Figure 8.11. Therefore, we break down any long pushing trajectory into smaller segments, currently $\sim$30 nm long. Having a reliable pushing routine, the output of the high-level planner can now be executed also with high reliability.

Now that we have discussed the high-level planner and its primitive commands, let us turn our attention to the software needed to implement the system. We found in the beginnings of the LMR, in 1994, that commercial AFM software was designed for

imaging and not suitable for manipulation. Therefore we designed and implemented a manipulation system, called Probe Control Software (PCS), running on top of the vendor-supplied Application Programming Interface (API) [36]. It was implemented on AutoProbe AFMs (Park Scientific Instruments, which later on became Thermomicroscopes and now Veeco), which to our knowledge were the only instruments sold with an available API. PCS evolved as time went by, and was the workhorse for the interactive manipulation research in our laboratory until recently. Research often moves in unpredictable ways, and we found that the ability to easily modify the software was fundamental to our experimental work. Unfortunately, the API was written for a 16-bit Windows system, which is far from being convenient to program. We concluded that we were spending too much time fighting an inhospitable programming environment and launched a re-write of the whole system, which has been "completed" recently. (We find that research software is in a permanent state of flux.)

The new system is called PyPCS, for Python PCS. It is written in C++ and Python, which is a scripting language that greatly facilitates program development. For example, new modules may be added to the system without the need to recompile and link the whole system. PyPCS has a client-server architecture. The server is written in C++ for 16-bit Windows and runs in the PC that controls the instrument. The client is written in Python, communicates with the server via standard interprocess communication primitives, and may run in the AFM PC or in any computer that is connected to it by Ethernet [96].

We conclude this section with a complete example, including planning and execution in the AFM. Figure 8.18 shows on the left panel the initial random dispersion of nanoparticles on the sample. On the right is the goal configuration (yellow crosses) plus the result of planning, with the pushing paths in red, and the positioning paths between pushes in green. The particles marked with a black

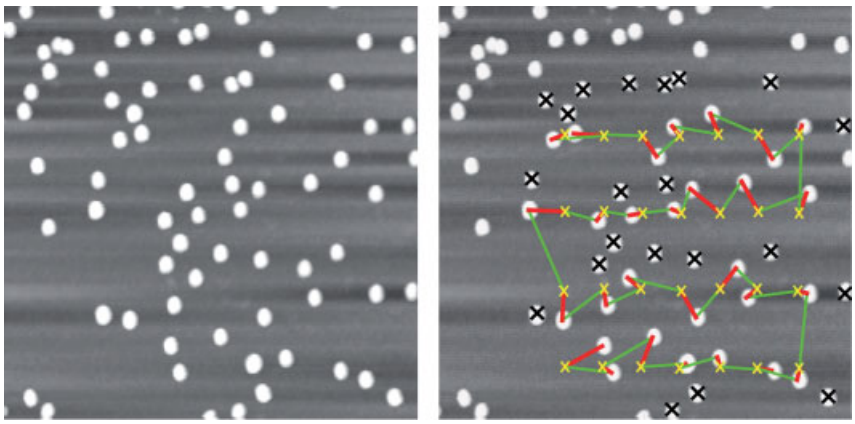


Figure 8.18 Left: Initial state. Right: Goal state (yellow) and planner output superposed on the intial image, showing pushing paths (red) and positioning paths (green). 15 nm Au particles on mica. Reproduced with kind permission from [90].

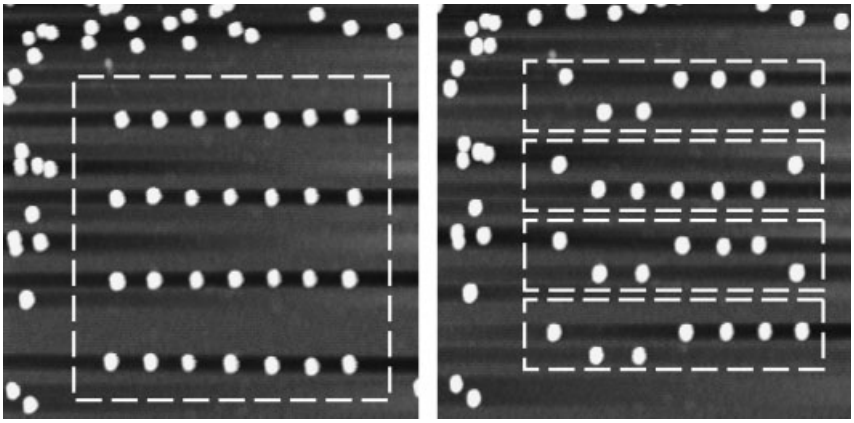


Figure 8.19 Left: Result of the execution of the plan of Figure 8.4.2.8. Right: An additional task, also planned and executed automatically. Reproduced with kind permission from [90].

cross are extraneous and should be removed from the area where the pattern is being built. (This is also done automatically.) Figure 8.19 shows on the left the result of executing the plan. On the right is the result of another similar operation also performed automatically. The pattern on the right of Figure 8.19 represents a different encoding of ASCI characters into nanoparticle positions – compare with Figure 8.13. Here a particle on the top row of each 2-row group signifies a "1" and a particle on the bottom row signifies a "0". The 4 groups of 2 rows read "NANO" in ASCI. This encoding uses twice as many particles as that of Figure 8.13 but has an interesting advantage: editing the stored data amounts simply to pushing particles up or down by a fixed amount, and could be achieved very efficiently by an AFM with a multi-tip array with spacing equal to that of the particle grid – see [97] and the discussion in the next section. The patterns shown in Figure 8.19 were built in a few minutes with the automated system. They would take at least one day of work by a skilled user if they had been built with our interactive system.

8.5 Conclusion and Outlook

Manipulation with the AFM of nanoscale objects with dimensions ~5–100 nm has been under study for over a decade, and is now routinely performed in several laboratories. Nevertheless, some basic questions remain unanswered. For example, how far off center can we hit a particle for it to be pushed reliably? How high above the surface can we strike a particle for it to move? Does the size of a particle matter? Do the shape and size of a tip matter? Why do particles fall off the desired trajectories? What is the force threshold needed to move a particle? Are there preferred directions

of motion? For a given surface, coating, nanoparticle, cantilever and environmental conditions, can we predict which operational parameters, if any, will result in reliable manipulation? And we could go on. In short, we lack a predictive understanding of the manipulation process, especially an understanding grounded on measurable parameters. However, there is enough experimental evidence on certain materials and systems for us to successfully complete non-trivial manipulation tasks, as shown in this chapter.

Nanomanipulation has been used until now for demonstrations or to prototype new devices. It is a useful prototyping tool because it can be used to build devices that cannot be made otherwise, and it greatly facilitates parametric studies. For example, the effect of spatial errors on a given device can be investigated by moving one of its constituent particles and recording the associated functional changes. The complexity of the structures built by nanomanipulation has been severely restricted by the sheer amount of labor and time needed to construct them interactively. Automated systems such as those described in this chapter are beginning to appear, and may significantly impact what can be done by nanomanipulation. It is fair to say, though, that a "killer app" has not yet been found for nanomanipulation. A high value, low volume application seems most appropriate, because, even with automated operation, nanomanipulation with the AFM is a serial and relatively slow process, not very suitable for mass production. Perhaps repair of fabrication masks and other high cost devices will be an important application in the future.

Another technical advance that may have a strong impact on nanomanipulation is the development of instruments with multiple tips. There are currently several research efforts aimed at producing multi-tip arrays, but they tend to be unusable for manipulation because they do not provide individual control of height (z) for each tip – see for example [98]. Furthermore, whereas efficient algorithms for nanolithography with multi-tip arrays are known [97, 99], manipulation tasks are inherently more difficult because of registration problems between tips and particles. The tips are normally arranged in a regular array, while the particles are initially randomly dispersed, and when a tip is positioned near a particle for pushing it, the other tips are unlikely to be in positions where they can push other particles. If this happens, a multi-tip array will be no faster than a single tip.

In summary, much has been learned about nanomanipulation with AFMs, and new automated systems are a breakthrough improvement over their traditional, interactive counterparts, but we still lack a deep, predictive understanding of the manipulation phenomena, as well as a convincing demonstration of economic viability for practical applications. Massively parallel operation by using multi-tip arrays may be he next breakthrough.

Acknowledgment and Disclaimer

The LMR work on nanotechnology was supported in part by the NSF Grants EIA-98-71775, IIS-99-87977, EIA-01-21141, DMI-02-09678 and Cooperative Agreement CCR-01-20778; and the Okawa Foundation.

I would like to thank my LMR faculty colleagues, postdocs and students, too many to mention here, who did much of the LMR work reported in this chapter and from whom I learned much over the last decade. I wish to single out the students who built the probe control software who made possible all of our work on nanomanipulation. They were Cenk Gazen, who started it all, Nick Montoya who extended and maintained PCS for a couple of years, Jon Kelly, who wrote the first version of PyPCS, Babak Mokaberi, who built the drift, creep and hysteresis compensators, Dan Arbuckle, who is the architect of the current version of PyPCS, which integrates the old PCS with Mokaberi's work, and Jaehong Yun, who did the high level planning software and, with Arbuckle, has integrated it into PyPCS.

An exhaustive bibliography on nanomanipulation and related topics is beyond the scope of this Chapter. In many cases I have attempted to cite the pioneering works on specific subjects but I may have failed to acknowledge some of them. I offer my apologies to the colleagues whom I may not have cited, or whose work I may have misinterpreted. There is simply too much research in this area for me to be able to keep up with all of it.

References

1 Binnig, G., Rohrer, H., Gerber, Ch. and Weibel, E. (1982) Surface studies by scanning tunneling microscopy. *Physical Review Letters*, **49**, 57–61.

2 Binnig, G., Quate, C.F. and Gerber, Ch. (1986) Atomic force microscope. *Physical Review Letters*, **56**, 931–933.

3 Dagata, J.A., Schneir, J., Harary, H.H., Evans, C.J., Postek, M.T. and Bennett, J. (1990) Modification of hydrogen-passivated silicon by a scanning tunneling microscope operating in air. *Applied Physics Letters*, **56**, 2001–2003.

4 Snow, E.S. and Campbell, P.M. (1995) AFM fabrication of sub- 10-nanometer metal-oxide devices with in situ control of electrical properties. *Science*, **270**, 1639–1641.

5 Salling, C.T. and Lagally, M.G. (1994) Fabrication of atomic-scale structures on Si(001) surfaces. *Science*, **265**, 502–506.

6 Becker, R.S., Golovchenko, J.A. and Swartzentruber, B.S. (1987) Atomic-scale surface modifications using a tunneling microscope. *Nature*, **325**, 419–421.

7 Mamin, H.J., Guethner, P.H. and Rugar, D. (1990) Atomic emission from a gold scanning-tunneling-microscope tip. *Physical Review Letters*, **65**, 2418–2421.

8 Wiesendanger, R. (1994) *Scanning Probe Microscopy, and Spectroscopy*. Cambridge University Press, Cambridge, UK. Chapter 8.

9 Piner, R.D., Zhu, J., Xu, F., Hong, S. and Mirkin, C.A. (1999) Dip-Pen, Nanolithography. *Science*, **283**, 661–663.

10 Diaz, D.J., Hudson, J.E., Storrier, G.D., Abruna, H.D., Sundararajan, N. and Ober, C.K. (2001) Lithographic applications of redox probe microscopy. *Langmuir*, **17**, 5932–5938.

11 Mesquida, P. and Stemmer, A. (2001) Attaching silica nanoparticles from suspension onto surface charge patterns generated by a conductive atomic force microscope tip. *Advanced Materials*, **13**, 1395–1398.

12 Sun, S. and Legget, G.J. (2002) Generation of nanostructures by scanning near-field photolithography of self-assembled monolayers and wet chemical etching. *Nanoletters*, **2**, 1223–1227.

13 Davis, Z.J., Abadal, G., Hansen, O., Borisé, X., Barniol, N., Pérez-Murano, F. and Boisen, A. (2003) AFM lithography of aluminum for fabrication of nanomechanical systems. *Ultramicroscopy*, **97** (1–4), 467–472.

14 Garno, J.C., Yang, Y., Amro, N.A., Cruchon-Dupeyrat, S., Chen, S. and Liu, G.-Y. (2003) Precise positioning of nanoparticles on surfaces using scanning probe lithography. *Nanoletters*, **3**, 389–395.

15 Takeda, S., Nakamura, C., Miyamoto, C., Nakamura, N., Kageshima, M., Tokumoto, H. and Miyake, J. (2003) Lithographing of biomolecules on a substrate surface using an enzyme-immobilized AFM tip. *Nanoletters*, **3**, 1471–1474.

16 Wouters, D. and Schubert, U.S. (2004) Nanolithography and nanochemistry: probe-related patterning techniques and chemical modification for nanometer-sized devices. *Angewandte Chemie-International Edition*, **43**, 2480–2495.

17 Eigler, D.M. and Schweizer, E.K. (1990) Positioning single atoms with a scanning tunneling microscope. *Nature*, **344**, 524–526.

18 Stroscio, J.A. and Eigler, D.M. (1991) Atomic and molecular manipulation with the scanning tunneling microscope. *Science*, **254**, 1319–1326.

19 Crommie, M.F., Lutz, C.P. and Eigler, D.M. (1993) Confinement of electrons to quantum corrals on a metal surface. *Science*, **262**, 218–220.

20 Lyo, I.-W. and Avouris, Ph. (1991) Field-induced nanometer- to atomic-scale manipulation of silicon surfaces with the STM. *Science*, **253**, 173–176.

21 Uchida, H., Huang, D.H., Yoshinobu, J. and Aono, M. (1993) Single atom manipulation on the Si(111)7×7 surface by the scanning tunneling microscope (STM). *Surface Science*, **287/288** (Part 2), 1056–1061.

22 Bartels, L., Meyer, G. and Rieder, K.-H. (1997) Basic steps of lateral manipulation of single atoms and diatomic clusters with a scanning tunneling microscope tip. *Physical Review Letters*, **79**, 697–700.

23 Stroscio, J.A., Tavazza, F., Crain, J.N., Celotta, R.J. and Chaka, A.M. (2006) Electronically-induced atom motion in engineered $CoCu_n$ nanostructures. *Science*, **313**, 948–951.

24 Jung, T.A., Schlitter, R.R., Gimzewski, J.K., Tang, H. and Joachim, C. (1996) Controlled room-temperature positioning of individual molecules: molecular flexure and motion. *Science*, **271**, 181–184.

25 Cuberes, M.T., Schlittler, R.R. and Gimzewski, J.K. (1996) Room-temperature repositioning of individual C_{60} molecules at Cu steps: operation of a molecular counting device. *Applied Physics Letters*, **69**, 3016–3018.

26 Maruno, S., Inanaga, K. and Isu, T. (1993) Threshold height for movement of molecules on Si(111)-7×7 with a scanning tunneling microscope. *Applied Physics Letters*, **63**, 1339–1341.

27 Beton, P.H., Dunn, A.W. and Moriarty, P. (1995) Manipulation of C_{60} molecules on a Si surface. *Applied Physics Letters*, **67**, 1075–1077.

28 Israelachvili, J.N. (1992) *Intermolecular, and Surface Forces*. 2nd edn., Academic Press, San Diego, CA.

29 Burnham, N.A., Chen, X., Hodges, C.S., Matei, G.A., Thoreson, E.J., Roberts, C.J., Davies, M.C. and Tendler, S.J.B. (2003) Comparison of calibration methods for atomic-force microscopy cantilevers. *Nanotechnology*, **14**, 1–6.

30 Garcia, R. and San Paulo, A. (1999) Attractive and repulsive tip-sample interaction regimes in tapping-mode atomic force microscopy. *Physical Review B-Condensed Matter*, **60**, 4961–4967.

31 Garcia, R. and San Paulo, A. (2000) Amplitude curves and operating regimes in dynamic atomic force microscopy. *Ultramicroscopy*, **82**, 79–83.

32 Sarid, S. (1994) *Scanning Force Microscopy*, Oxford University Press, Oxford, UK.

33 Waser, R. (ed.) (2003) *Nanoelectronics, and Information Technology*, Wiley-VCH, Weinheim, Germany.

34 Meyer, E., Hug, H.J. and Bennewitz, R. (2004) *Scanning Probe Microscopy*, Springer Verlag, Heidelberg, Germany.

35 Villarrubia, J.S. (1994) Morphological estimation of tip geometry for scanned probe microscopy. *Surface Science*, **321**, 287–300.

36 Baur, C., Gazen, B.C., Koel, B., Ramachandran, T.R., Requicha, A.A.G. and Zini, L. (1997) Robotic nanomanipulation with a scanning probe microscope in a networked computing environment. *Journal of Vacuum Science & Technology B*, **15**, 1577–1580.

37 Baur, C., Bugacov, A., Koel, B.E., Madhukar, A., Montoya, N., Ramachandran, T.R., Requicha, A.A.G., Resch, R. and Will, P. (1998) Nanoparticle manipulation by mechanical pushing: underlying phenomena and real-time monitoring. *Nanotechnology*, **9**, 360–364.

38 Bugacov, A., Resch, R., Baur, C., Montoya, N., Woronowicz, K., Papson, A., Koel, B.E., Requicha, A.A.G. and Will, P. (1999) Measuring the tip-sample separation in dynamic force microscopy. *Probe Microscopy*, **1**, 345–354.

39 Requicha, A.A.G., Baur, C., Bugacov, A., Gazen, B.C., Koel, B., Madhukar, A., Ramachandran, T.R., Resch, R. and Will, P. (1998) Nanorobotic assembly of two-dimensional structures. Proceedings IEEE International Conference on Robotics and Automation (ICRA '98), Leuven, Belgium, May 16–21, pp. 3368–3374.

40 Requicha, A.A.G., Meltzer, S., Terán Arce, P.F., Makaliwe, J.H., Sikén, H., Hsieh, S., Lewis, D., Koel, B.E. and Thompson, M.E. (2001) Manipulation of nanoscale components with the AFM: principles and applications. Proceedings 1st IEEE International Conference on Nanotechnology, Maui, HI, October 28–30, pp. 81–86.

41 Resch, R., Bugacov, A., Baur, C., Koel, B.E., Madhukar, A., Requicha, A.A.G. and Will, P. (1998) Manipulation of nanoparticles using dynamic force microscopy: simulation and experiments. *Applied Physics A*, **67**, 265–271.

42 Resch, R., Meltzer, S., Vallant, T., Hoffmann, H., Koel, B.E., Madhukar, A., Requicha, A.A.G. and Will, P. (2001) Immobilizing Au nanoparticles on SiO_2 surfaces using octadecylsiloxane monolayers. *Langmuir*, **17**, 5666–5670.

43 Resch, R., Lewis, D., Meltzer, S., Montoya, N., Koel, B.E., Madhukar, A., Requicha, A.A.G. and Will, P. (2000) Manipulation of gold nanoparticles in liquid environments using scanning force microscopy. *Ultramicroscopy*, **82** (1–4), 135–139.

44 Requicha, A.A.G. (2003) Nanorobots, NEMS and nanoassembly. Proceedings IEEE, Special issue on nanoelectronics and nanoscale processing, Vol. 91, No. 11, November, pp. 1922–1933.

45 Taylor, R.M. II, Chen, J., Okimoto, S., Llopis-Artime, N., Chi, V.L., Brooks, F.P. Jr., Falvo, M., Paulson, S., Thiansathaporn, P., Glick, D., Washburn, S. and Superfine, R. (1997) Pearls found on the way to the ideal interface for scanned-probe microscopes. Proceedings IEEE Visualization '97, Phoenix, AZ, October 19–24, pp. 467–470.

46 Schaefer, D.M., Reifenberger, R., Patil, A. and Andres, R.P. (1995) Fabrication of two-dimensional arrays of nanometer-size clusters with the atomic force microscope. *Applied Physics Letters*, **66**, 1012–1014.

47 Junno, T., Deppert, K., Montelius, L. and Samuelson, L. (1995) Controlled manipulation of nanoparticles with an atomic force microscope. *Applied Physics Letters*, **66**, 3627–3629.

48 Martin, M., Roschier, L., Hakonen, P., Parts, U., Paalanen, M., Schleicher, B. and Kauppinen, E.I. (1998) Manipulation of Ag nanoparticles utilizing noncontact atomic force microscopy. *Applied Physics Letters*, **73**, 1505–1507.

49 Sheehan, P.E. and Lieber, C.M. (1996) Nanotribology and nanofabrication of MoO_3 structures by atomic force microscopy. *Science*, **272**, 1158–1161.

50 Guthold, M., Falvo, M.R., Matthews, W.G., Paulson, S., Washburn, S., Erie, D.A., Superfine, R., Brooks, F.P. Jr. and Taylor, R.M. II. (June 2000) Controlled manipulation of molecular samples with the nanoManipulator. *IEEE/ASME Transactions on Mechatronics*, **5** (2), 189–198.

51 Sitti, M. and Hashimoto, H. (June 2000) Controlled pushing of nanoparticles: modeling and experiments. *IEEE/ASME Transactions on Mechatronics*, **5** (2), 199–211.

52 Li, G., Xi, N., Yu, M. and Fung, W.K. (2003) 3-D nanomanipulation using atomic force microscopy. Proceedings IEEE International Conference on Robotics and Automation (ICRA '03), Taipei, Taiwan, September 14–19, pp. 3642–3647.

53 Theil Hansen, L., Kühle, A., Sørensen, A.H., Bohr, J. and Lindelof, P.E. (1998) A technique for positioning nanoparticles using an atomic force microscope. *Nanotechnology*, **9**, 337–342.

54 Vikramaditya, B. and Nelson, B.J. (1997) Visually guided microassembly using optical microscopes and active vision. Proceedings IEEE International Conference on Robotics and Automation, Albuquerque, NM, April 21–27, pp. 3172–3177.

55 Sato, T., Kameya, T., Miyazaki, H. and Hatamura, Y. (1995) Hand-eye system in the nano manipulation world. Proceedings IEEE International Conference on Robotics and Automation, Nagoya, Japan, May 21–27, pp. 59–66.

56 Miyazaki, H. and Sato, T. (1997) Mechanical assembly of three-dimensional microstructures from fine particles. *Advanced Robotics*, **11**, 169–185.

57 Yu, M.-F., Dyer, M.J., Skidmore, G.D., Rohrs, H.W., Lu, X.-K., Hausman, K.D., von Her, J.R. and Ruoff, R.S. (1999) Three dimensional manipulation of carbon nanotubes under a scanning electron microscope. *Nanotechnology*, **10**, 244–252.

58 Dong, L., Arai, F. and Fukuda, T. (2001) 3D nanorobotic manipulation of multi-walled carbon nanotubes. Proceedings IEEE International Conference on Robotics & Automation, Seoul, S. Korea, May 21–26, pp. 632–637.

59 Fatikow, S., Wich, T., Hülsen, H., Sievers, T. and Jähnisch, M. (2006) Microrobot system for automatic nanohandling inside a scanning electron microscope. Proceedings IEEE International Conference on Robotics & Automation (ICRA '06), Orlando, FL, May 15–19, pp. 1401–1407.

60 Resch, R., Baur, C., Bugacov, A., Koel, B.E., Madhukar, A., Requicha, A.A.G. and Will, P. (1998) Building and manipulating 3-D and linked 2-D structures of nanoparticles using scanning force microscopy. *Langmuir*, **14**, 6613–6616.

61 Decossas, S., Mazen, F., Baron, T., Brémond, G. and Souifi, A. (2003) Atomic force microscopy nanomanipulation of silicon nanocrystals for nanodevice fabrication. *Nanotechnology*, **14**, 1272–1278.

62 Requicha, A.A.G. (1999) Nanoparticle patterns. *J Nanoparticle Res*, **1**, 321–323.

63 Burmeister, F., Schäfle, C., Keilhofer, B., Bechinger, C., Boneberg, J. and Leiderer, P. (1988) From mesoscopic to nanoscopic structures: lithography with colloid monolayers. *Advanced Materials*, **10**, 495–497.

64 Tada, T., Kanayama, T., Koga, K., Seeger, K., Carroll, S.J., Weibel, P. and Palmer, R.E. (1998) Fabrication of size-controlled 10-nm scale Si pillars using metal clusters as formation nuclei. *Microelectronic Eng*, **41/42**, 539–542.

65 Lewis, P.A. and Ahmed, H. (1999) Patterning of silicon nanopillars formed with a colloidal gold etch mask. *Journal of*

Vacuum Science & Technology B, **17**, 3239–3243.

66 Zheng, J., Chen, Z. and Liu, Z. (2000) Atomic force microscopy-based nanolithography on silicon using colloidal Au nanoparticles as a nanooxidation mask. *Langmuir*, **16**, 9673–9676.

67 Chou, S.Y., Krauss, P.R. and Renstrom, P.J. (1996) Imprint lithography with 25-nanometer resolution. *Science*, **272**, 85–87.

68 Junno, T., Carlsson, S.-B., Xu, H., Montelius, L. and Samuelson, L. (1998) Fabrication of quantum devices by Ångström-level manipulation of nanoparticles with an atomic force microscope. *Applied Physics Letters*, **72**, 548–550.

69 Maier, S.A., Brongersma, M.L., Kik, P.G., Meltzer, S., Requicha, A.A.G., Koel, B.E. and Atwater, H.A. (2001) Plasmonics – a route to nanoscale optical devices. *Advanced Materials*, **13**, 1501–1505.

70 Maier, S.A., Kik, P.G., Atwater, H.A., Meltzer, S., Harel, E., Koel, B.E. and Requicha, A.A.G. (2003) Local detection of electromagnetic energy transport below the diffraction limit in metal nanoparticle plasmon waveguides. *Nature Materials*, **2**, 229–232.

71 Mokaberi, B. and Requicha, A.A.G. (2006) Drift compensation for automatic nanomanipulation with scanning probe microscopes. *IEEE Transactions on Automation Science and Engineering*, **3**, 199–207.

72 Hsieh, S., Meltzer, S., Wang, C.R.C., Requicha, A.A.G., Thompson, M.E. and Koel, B.E. (2002) Imaging and manipulation of gold nanorods with an Atomic Force Microscope. *The Journal of Physical Chemistry B*, **106**, 231–234.

73 Chen, H., Xi, N. and Li, G. (2006) CAD-guided automated nanoassembly using atomic force microscopy-based nanorobotics. *IEEE Transactions on Automation Science and Engineering*, **3**, 208–217.

74 Hertel, T., Martel, R. and Avouris, Ph. (1998) Manipulation of individual carbon nanotubes and their interaction with surfaces. *The Journal of Physical Chemistry B*, **102**, 910–915.

75 Falvo, M.R., Taylor, R.H. II, Helser, A., Chi, V., Brooks, F.P. Jr., Washburn, S. and Superfine, R. (1999) Nanometre-scale rolling and sliding of carbon nanotubes. *Nature*, **397**, 236–238.

76 Dong, L.X., Arai, F. and Fukuda, T. (2001) Three-dimensional nanoassembly of multi-walled carbon nanotubes through nanorobotic manipulations by using electron-beam induced deposition. Proceedings 1st IEEE International Conference on Nanotechnology, Maui, HI, October 28–30, pp. 93–98.

77 Dong, L., Arai, F., Nakajima, M., Liu, P. and Fukuda, T. (2003) Nanotube devices fabricated in a nano laboratory. Proceedings IEEE International Conference on Robotics & Automation, Taipei, Taiwan, September 24–29, pp. 3624–3629.

78 Roschier, L., Penttilä, J., Martin, M., Hakonen, P., Paalanen, M., Tapper, U., Kauppinen, E., Journet, C. and Bernier, P. (1999) Single-electron transistor made of multi-walled carbon nanotube using scanning probe manipulation. *Applied Physics Letters*, **75**, 728–730.

79 Resch, R., Baur, C., Bugacov, A., Koel, B.E., Echternach, P.M., Madhukar, A., Montoya, N., Requicha, A.A.G. and Will, P. (1999) Linking and manipulation of gold multi-nanoparticle structures using dithiols and scanning force microscopy. *The Journal of Physical Chemistry. B*, **103**, 3647–3650.

80 Requicha, A.A.G., Resch, R., Montoya, N., Koel, B.E., Madhukar, A. and Will, P. (1999) Towards hierarchical nanoassembly. Proceedings International Conference on Intelligent Robots and Systems (IROS '99), Kyongju, S. Korea, October 17–21, pp. 889–893.

81 Meltzer, S., Resch, R., Koel, B.E., Thompson, M.E., Madhukar, A., Requicha,

A.A.G. and Will, P. (2001) Fabrication of nanostructures by hydroxylamine-seeding of gold nanoparticle templates. *Langmuir*, **17**, 1713–1718.

82 Harel, E., Meltzer, S.E., Requicha, A.A.G., Thompson, M.E. and Koel, B.E. (2005) Fabrication of latex nanostructures by nanomanipulation and thermal processing. *Nanoletters*, **5**, 2624–2629.

83 Madsen, D.N., Mølhave, K., Mateiu, R., Bøggild, P., Rasmussen, A.M., Appel, C.C., Brorson, M. and Jacobsen, C.J.H. (2003) Nanoscale soldering of positioned carbon nanotubes using highly conductive electron beam induced gold deposition. Proceedings IEEE International Conference on Nanotechnology, S. Francisco, CA, August 12–14, pp. 335–338.

84 Hollis, R.L., Salcudean, S. and Abraham, D.W. (1990) Toward a tele-nanorobotic manipulation system with atomic scale force feedback and motion resolution. Proceedings IEEE International Conference on Microelectromechanical Systems, Napa Valley, CA, February 11–14, pp. 115–119.

85 Taylor, R.M. II, Robinett, W., Chi, V.L., Brooks, F.P. Jr., Wright, W.V., Williams, R.S. and Snyder, E.J. (1993) The nanomanipulator: a virtual reality interface for a scanning tunneling microscope. Proceedings ACM SIGGRAPH '93, Anaheim, CA, August 1–6, pp. 127–134.

86 Finch, M., Chi, V.L., Taylor, R.M. II, Falvo, M., Washburn, S. and Superfine, R. (1995) Surface modification tools in a virtual environment interface to a scanning probe microscope. Proceedings ACM Symposium on Interactive 3D Graphics, Monterey, CA, April 9–12, pp. 13–18.

87 Sitti, M. and Hashimoto, H. (1998) Tele-nanorobotics using atomic force microscope. Proceedings IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS '98), Victoria, Canada, October 13–17, pp. 1739–1746.

88 Li, G., Xi, N., Yu, M. and Fung, W.K. (2003) Augmented reality system for real-time nanomanipulation. Proceedings IEEE International Conference on Nanotechnology, S. Francisco, CA, August 12–14, pp. 64–67.

89 Li, G., Xi, N., Yu, M. and Fung, W.K.(June 2004) Development of augmented reality system for AFM-based nanomanipulation. *IEEE/ASME Transactions on Mechatronics*, **9** (2), 358–365.

90 Mokaberi, B., Yun, J., Wang, M. and Requicha, A.A.G. (2007) Automated nanomanipulation with atomic force microscopes. Proceedings IEEE International Conference on Robotics and Automation (ICRA '07), Rome, Italy, April 10–14, pp. 1406–1412.

91 Knuth, D.E. (1993) *The Stanford GraphBase*, The ACM Press, New York, NY.

92 Latombe, J.-C. (1991) *Robot Motion Planning*, Kluwer, Boston, MA.

93 Makaliwe, J.H. and Requicha, A.A.G. (2001) Automatic planning of nanoparticle assembly tasks. Proceedings IEEE International Symposium on Assembly & Task Planning (ISATP '01), Fukuoka, Japan, May 28–30, pp. 288–293.

94 Mokaberi, B. and Requicha, A.A.G. (2004) Towards automatic nanomanipulation: drift compensation in scanning probe microscopy. Proceedings IEEE International Conference on Robotics and Automation (ICRA '04), New Orleans, LA, April 25–30, pp. 416–421.

95 Mokaberi, B. and Requicha, A.A.G. (in press) Compensation of scanner creep and hysteresis for AFM nanomanipulation. *IEEE Transactions on Automation Science & Engineering*. doi:10.1109/TASE.2007.895008.

96 Arbuckle, D.J., Kelly, J. and Requicha, A.A.G. (2006) A high-level nanomanipulation control framework. Proceedings International Advanced Robotics Programme (IARP) Workshop on Micro and Nano Robotics, Paris, France, October 23–24,

97 Requicha, A.A.G. (1999) Massively parallel nanorobotics for lithography and data storage. *International Journal of Robotics Research*, **18**, 344–350.

98 Vettiger, P., Cross, G., Despont, M., Drechsler, U., Dürig, U., Gotsmann, B., Häberle, W., Lantz, M.A., Rothuizen, H.E., Stutz, R. and Binnig, G.K. (March 2002) The millipede – nanotechnology entering data storage. *IEEE Transactions on Nanotechnology*, **1** (1), 39–55.

99 Arbuckle, D.J. and Requicha, A.A.G. (2003) Massively parallel scanning probe nanolithography. Proceedings, 3rd IEEE International Conference on Nanotechnology, San Francisco, CA, August 12–14, pp. 72–74.

9
Harnessing Molecular Biology to the Self-Assembly of Molecular-Scale Electronics

Uri Sivan

9.1
Introduction

Microelectronics and biology provide two distinct paradigms for complex systems. In microelectronics, the information guiding the fabrication process is encoded into computer programs or glass masks and, based on that information, a complex circuit is imprinted in silicon in a series of chemical and physical processes. This top-to-bottom approach is guided by a supervisor whose "wisdom" is external to the circuit being built. Biology adopts an opposite strategy, whereby complex constructs are assembled from molecular-scale building blocks, based on the information encoded into the ingredients. For example, proteins are synthesized from amino acids based on the instructions coded in the genome and other proteins. The assembled objects process further molecules to form larger structures capable of executing elaborate functions, and so on. This autonomous bottom-up strategy allows, in critical bottlenecks, for an exquisite control over the molecular structure in a way which is unmatched by man-made engineering. In other cases it allows for the errors that are so critical for evolution.

The fact that man-made engineering evolved so differently from "nature engineering" deserves a separate discussion that is beyond the scope of this chapter. Here, we will only comment that the perception of nature as a type of engineering is somewhat oversimplifying. While engineering aims at meeting a predefined challenge – namely, to execute a desired function – nature evolved with no aim. Yet, the hope behind biomimetics is that concepts and tools which evolved during several billions years of evolution may find applications in engineering.

Electronics is particularly alien to biology. With the exception of short-range electron hopping in certain proteins, biology relies on ion transport rather than electrons. The electronic conductivity of biomolecules is orders of magnitude too

Nanotechnology. Volume 3: Information Technology I. Edited by Rainer Waser

ISBN: 978-3-527-31738-7

small for implementing them as useful electronic components. For instance, albeit in earlier reports, DNA has been found to be an excellent insulator [1–3]. The foreseen potential of biology in the context of electronics is, therefore, in the assembly process rather than in electronic functionality *per se*. This observation is reflected in the scientific research described below; it concerns the bioassembly of electronic materials to form devices, rather than attempts to use biomolecules as electronic components.

The term "self-assembly" is widely used to describe a variety of processes which include the self-assembly of organic molecules to form uniform monolayers on substrates. This is not the type of self-assembly under consideration in this chapter, whereby the term refers to the construction of an elaborate object, namely, the embedment of a significant amount of information into the object being built. The subject of the intimate relationship between self-assembly, information, and complexity will be revisited in Section 9.4.

The term "complex self-assembly" deserves some introductory remarks. When looking back at nature, one realizes that complex objects are typically assembled in a modular way. Most protein machines, for instance, comprise several subunits, each made of a separate protein. Each such protein is synthesized in the cell from amino acids which are in turn synthesized from atoms. This example is identified in four levels of hierarchy, namely atoms, amino acids, proteins, and machines made of several protein subunits. This hierarchal or modular assembly is an essential ingredient of complex self-assembly, the reason being that none of the modules reflects a global minimal free energy of its elementary constituents. The protein machine, for instance, does not pertain to a minimal free energy of the collection of amino acids making it, and so on.

In many instances the system is guided to a certain configuration by auxiliary molecules (enzymes, chaperones, etc.) which at times consume energy. However, in the cases of interest here, where self-assembly is governed by non-covalent interactions and relatively simple configurations, each step can be driven by a down-hill drift in free energy towards a long-lived metastable state, thus rendering the module amenable for the next assembly step. Clearly, complex electronics cannot be assembled from its elementary building blocks in a single step, and so requires modular assembly.

The next comment concerns the unavoidable errors characterizing self-assembly. In order for molecular recognition to take place, the molecules should effectively explore multiple docking configurations with other parts of the target molecule or with other molecules. The free energy landscape corresponding to the collection of all such configurations should, therefore, facilitate thermally assisted hops between local minima, corresponding to "wrong" configurations, in addition to the desired configuration. Special measures must be devised in order to produce overwhelming discrimination in favor of the desired configuration at finite time experiments. In the absence of such measures, the yield of self-assembly is intrinsically limited by the same fluctuations that facilitate molecular recognition. Over time, biology has evolved sophisticated error suppression and correction tools, and equivalent

methods will have to be developed if the self-assembly of molecular-scale electronics is to be taken seriously.

The effect of errors on modular assembly is of particular importance. As the yield in each assembly step is less than perfect, faulty modules are produced. An uncontrolled modular assembly therefore inevitably produces an exponentially larger fraction of faulty modules as the levels of hierarchy accumulate. One strategy for making useful circuits may thus rely on circuit architectures that are tolerant to faults. One such outstanding example is embodied in the Teramac machine developed at HP laboratories [4]. In the present chapter we adhere to conventional architectures requiring near-perfect circuits. The faulty modules in each step therefore need to be identified and either repaired or eliminated. Within the context of electronic circuits built by biology, the identification of faulty devices and their removal presents a remarkable challenge; that of devising a biomolecular machine that tests non-biological devices for electronic functionality, filters out non-functional devices, and then signals the system to proceed to the next assembly step. Although significant progress has been made towards the isolation of antibodies that sense the electric output presented to them by an electronic device, the discussion of electro-bio feedback loops is deferred to future publications, and focus here is on free-running assembly.

The conjecture behind the experiments described in this chapter may be summarized as follows. Simple functional devices can be assembled efficiently from electronic materials, taking advantage of the remarkable assembly tools provided by molecular biology. The realization of elaborate constructs necessitates hierarchical modular assembly, while the inevitable accumulation of errors with increasing levels of hierarchy requires error suppression and correction mechanisms, as well as biomolecular feedback switches that judge for electronic functionality and feedback to the bioassembly process.

In Section 9.2 the concept is of DNA-templated electronics [1, 5] is introduced, and expanded to include sequence-specific molecular assembly [6, 7] based on the recombinant protein, RecA. This section culminates in the bioassembly of a fully functional field effect transistor made from a carbon nanotube (CNT) [8]. While the topics of Section 9.2 rely on existing biotechnological tools, in Section 9.3 the toolbox is expanded to demonstrate how to isolate antibodies that recognize electronic materials directly [9]. The fact such antibodies can be isolated is encouraging with respect to the prospect of realizing a functional interface between molecular biology and nanoelectronics [10].

As DNA-templated electronics requires long DNA molecules with unique addresses, advantage is then taken of DNA computing to demonstrate the autonomous synthesis of DNA templates having these properties [11]. The synthesis algorithm, as outlined in Section 9.4, relies on the chemical realization of shift registers (SRs), and incorporates an error suppression scheme inspired by the redundancy codes employed in data communication. To the best of the present author's knowledge, these chemical SRs constitute the first embodiment of error suppression codes in chemical synthesis.

9.2
DNA-Templated Electronics

9.2.1
Scaffolds and Metallization

Double-stranded DNA (dsDNA) is chosen in most cases to template the assembly of molecular-scale electronics as well as other constructs of non-biological functionality. dsDNA is mechanically and chemically stable, easy to obtain at any desired sequence, and readily amenable to diverse enzymatic manipulations including restriction, digestion, replication, ligation, and recombination. In the schemes described below, dsDNA doubles as the information-carrying molecule and the physical support for the assembled electronic materials.

The assembly of DNA-templated electronics comprises two steps. First, the biological machinery is employed to construct a DNA scaffold with well-defined molecular addresses. Then, electronic functionality is instilled by the localization of electronic devices at specific addresses along the scaffold and conversion of the DNA template into a conductive network interconnecting devices to each other and to the external world.

An heuristic solution to some of the major challenges faced by molecular electronics, namely, the precise localization of a large number of molecular devices, inter-device wiring, and electrical interface between the molecular and macroscopic worlds, is depicted in Figure 9.1(a–d). The first step involves the definition of macroscopic electrodes on an inert substrate. As the electrodes are macroscopic, this process can be performed using standard photolithographic techniques (Figure 9.1a). The electrodes are provided with an identity by covering each of them

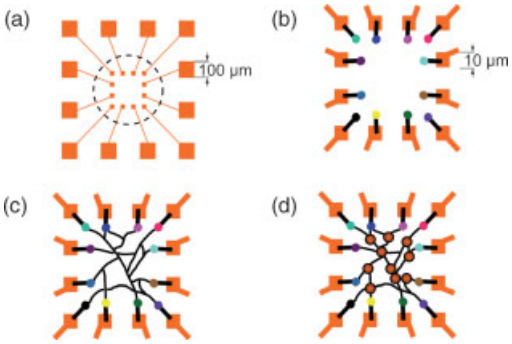


Figure 9.1 Heuristic scheme of a DNA-templated electronic circuit. (a) Gold pads are defined on an inert substrate. Panels (b–d) correspond to the circle of (a) at different stages of circuit construction. (b) Oligonucleotides of different sequences are attached to the different pads. (c) DNA network is constructed and bound to the oligonucleotides on the gold electrodes. (d) Metal clusters or molecular electronic devices are localized on the DNA network. The DNA molecules are finally converted into metallic wires, rendering the construct into a functional electronic circuit. Note that the figures are not to scale; the metallic clusters are nanometer-sized, while the electrode pads are micrometer-sized.

with a monolayer of a different short, single-stranded oligonucleotide using, for example, an ink-jet printer (Figure 9.1b). This step may still involve physical manipulations as the electrodes to be covered are macroscopic. After this step, each electrode is labeled with a monolayer of a unique oligonucleotide sequence and, hence, is able to recognize a specific complementary sequence in solution. In the third step, a network of well-defined connectivity is assembled using DNA hybridization and recombinant processes (see below). The network is then localized on the substrate using, for example hybridization of DNA molecules with the electrode-bound oligonucleotides (Figure 9.1c). The previous steps instill the formerly uniform substrate with well-defined molecular addresses based on distinct sequences of DNA molecules. This allows the subsequent positioning of functional electronic elements at molecularly accurate addresses (Figure 9.1d). At the end of this step, the network should bear functional elements at predesigned sites. However, as DNA molecules have insulating properties, the network should be functionalized (e.g., metallized) in order to render it conductive.

Now, the questions to be asked are how accurate is the topology of the assembled DNA network? Can it be inspected for structural integrity? Was the template assembled properly on the electrodes? Were the devices localized at their planned destinations? Did the devices connect electrically? Are they functional? These are just a few of the questions that must be addressed in any specific attempt to self-assemble molecular-scale electronics.

The experimental procedure used to demonstrate DNA-templated assembly and electrode attachment of a conductive silver wire [1, 5] are depicted in Figure 9.2. First, 12-base oligonucleotides, derivatized with a disulfide group at their 3′ end, were attached to the electrodes through a thiol–gold interaction. Each of the two electrodes was marked with a different oligonucleotide sequence. The electrodes were then bridged by hybridization of a 16 μm-long λ-DNA molecule containing two 12-base-long sticky ends, each of which was complementary to one of the two sequences attached to the gold electrodes.

The inset to Figure 9.3 presents a single DNA molecule bridge as observed by fluorescence microscopy. The measurements on the stretched DNA molecules indicated a resistance higher than the internal resistance of the measurement apparatus ($>10^{13}\,\Omega$). It was therefore concluded that, in order to instill electrical functionality, the DNA bridge must be coated with metal. Albeit contradicting results reported previously in the literature, it is now widely accepted that the intrinsic conductivity of DNA is indeed too small for direct application as a conducting element in a circuit [2].

The three-step silver-coating process (Figure 9.2c–e) was based on the selective localization of silver ions along the DNA molecule through Ag^+/Na^+ ion-exchange [1, 5], and the formation of complexes between the silver and the DNA bases. The silver ion-exchanged DNA was then reduced to form nanometer-sized silver aggregates bound to the DNA skeleton. These aggregates were further "developed" (much as in a standard photographic procedure) by using an acidic solution of hydroquinone and silver ions under low-light conditions [12, 13]. This solution was metastable, and spontaneous metal deposition was normally very slow,

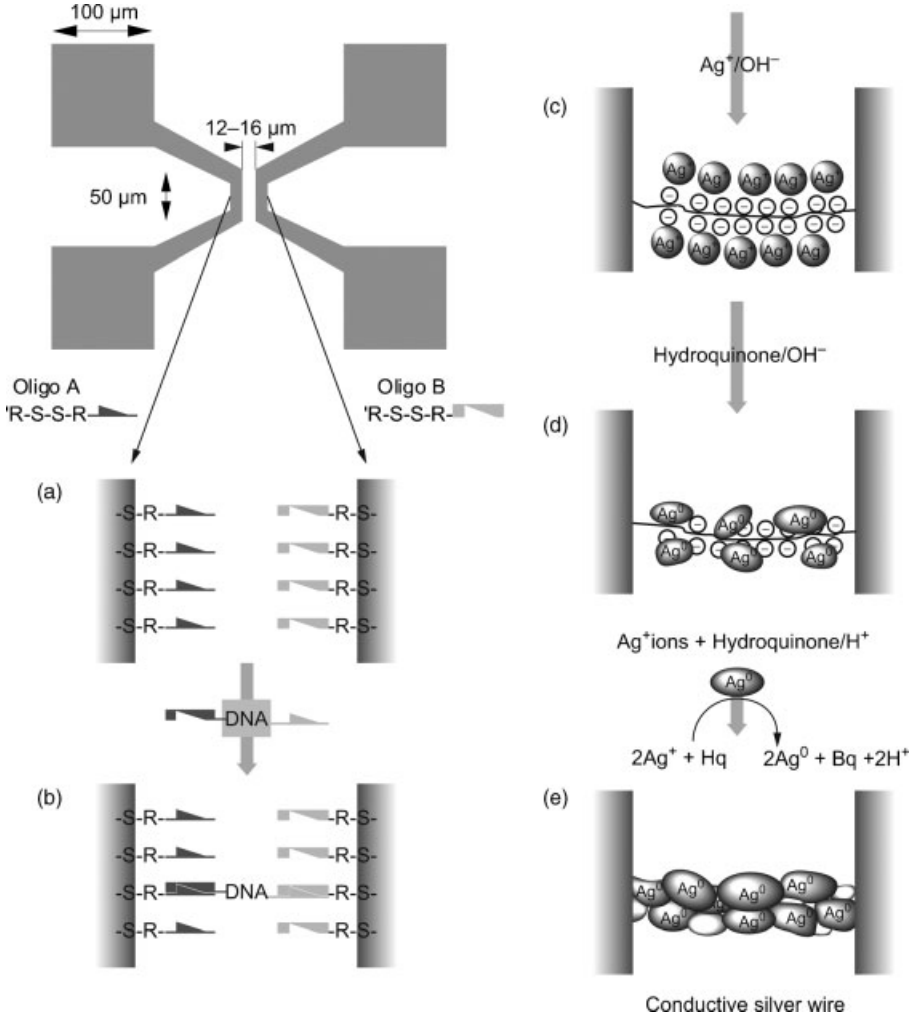


Figure 9.2 A gold pattern, 0.5 × 0.5 mm in size, was defined on a passivated glass using microelectronics techniques. The pattern comprised four bonding pads, each 100 μm in size, connected to two 50 μm-long parallel gold electrodes, 12–16 μm apart. (a) The electrodes were each wetted with a 10^{-4} μL droplet of disulfide-derivatized oligonucleotide solution of a given sequence (Oligos A and B). (b) After rinsing, the structure was covered with 100 μL of a solution of λ-DNA having two sticky ends that are complementary to Oligos A and B. A flow was applied to stretch the λ-DNA molecule between the two electrodes, allowing its hybridization. (c) The DNA bridge was loaded with silver ions by Na^+/Ag^+ ion exchange. (d) The silver ion–DNA complex was reduced using a basic hydroquinone solution to form metallic silver aggregates bound to the DNA skeleton. (e) The DNA templated wire was "developed" using an acidic solution of hydroquinone and silver ions. (Reprinted from Ref. [1]; © *Nature*, 1998.).

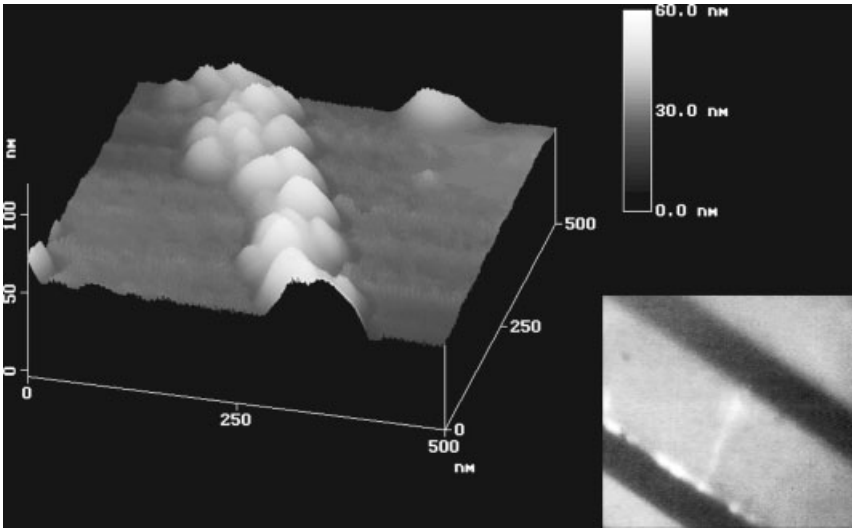


Figure 9.3 Atomic force microscopy (AFM) image of a silver wire connecting two gold electrodes 12 μm apart. Field size = 0.5 μm. Inset: Fluorescently labeled λ-DNA molecule stretched between two gold electrodes (dark strips), 16 μm apart. (Reprinted from Ref. [1]; © *Nature*, 1998.).

except on the silver aggregates attached to the DNA catalyzed the process. Under the experimental conditions, metal deposition therefore occurred only along the DNA skeleton, leaving the passivated substrate practically clean of silver. An atomic force microscopy (AFM) image of a segment of a 100 nm-wide, 12 μm-long silver wire prepared in this way is shown in Figure 9.3.

Since the publication of Ref. [1], the metallization scheme has been improved in two essential ways. First, silver has been replaced with gold [6] in the enhancing step and, after a few hours sintering at 300 °C, excellent wires were obtained. Second, the hydroquinone has been substituted for glutaraldehyde [6, 14] localized on the DNA itself. The confinement of the reducing agent to the DNA molecule suppressed non-specific metal deposition on other objects in the system, leading to much cleaner circuits. A DNA-templated gold wire is depicted in the inset of Figure 9.4, together with its current–voltage (I–V) characteristics.

Other research groups have since extended the scope of the metallization of biomolecules to proteins, amyloid fibrils, protein S-layers, microtubules, actin fibers, and even complete viral particles. Today, the choice of metals includes Pd, Pt, Au, Cu, and Co. An account of biomolecules metallization can be found in Refs. [15–20].

9.2.2
Sequence-Specific Molecular Lithography

In analogy with photolithography in conventional microelectronics, the realization of DNA-templated devices and circuits requires tools for defining circuit architectures.

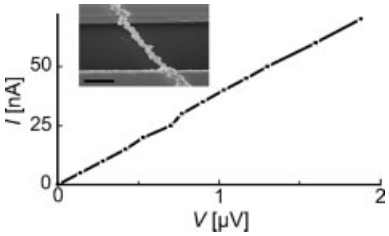


Figure 9.4 Two-terminal current–voltage (I–V) curve of a DNA-templated gold wire. The resistivity of the wire ($1.5 \times 10^{-7}\,\Omega\cdot$m) was only seven-fold higher than that of polycrystalline gold ($2.2 \times 10^{-8}\,\Omega\cdot$m). Inset: Scanning electron microscopy (SEM) image of a typical DNA-templated gold wire stretched between two electrodes deposited by electron-beam lithography. Scale bar = 1 μm. (Reprinted from Ref [6]; © 2002, American Association for the Advancement of Science.).

These include the formation of rich geometries, wire patterning at molecular resolutions, and molecularly accurate device localization. To that end, "sequence-specific molecular lithography" has been developed which enables the elaborate manipulation of dsDNA molecules, including patterning of the metal coating of DNA, the localization of labeled molecular objects at arbitrary addresses on dsDNA, and the generation of molecularly accurate stable DNA junctions [6, 8, 14].

The molecular lithography system developed at Technion utilizes homologous genetic recombination processes carried out by the RecA protein from *Escherichia coli*. The patterning information encoded in the DNA molecules replaces the masks used in conventional photolithography, while the RecA protein serves as the resist. The molecular lithography functions at high resolution over a broad range of length scales, from nanometers to many micrometers.

Homologous genetic recombination is one of several mechanisms that cells use to manipulate their DNA [21]. In this process, two parental DNA molecules which possess some sequence homology cross-over at equivalent sites. The reaction is based on protein-mediated, sequence-specific DNA–DNA interaction. Although RecA is the major protein responsible for this process in *E. coli*, it is also able to carry out the essential steps of the recombination process *in vitro*.

In the present author's procedure, RecA monomers are polymerized on a probe single-stranded DNA (ssDNA) molecule to form a nucleoprotein filament (Figure 9.5, step i). The nucleoprotein filament binds to a substrate molecule at an homologous probe–substrate location (Figure 9.5, step ii). RecA allows the addressing of an arbitrary sequence, from as few as 15 bases [22] to many thousands of bases, by the same standard reaction. This versatility presents an advantage over DNA-binding proteins which are restricted to particular DNA sequences. Moreover, unlike DNA hybridization, sequence-specific recognition can be performed on dsDNA, rather than ssDNA. Being chemically more inert and mechanically more rigid, the former provides a better substrate than the latter. The high efficiency and specificity of the recombination reaction, which evidently is essential for its biological roles, are beneficial in its utilization for molecular lithography.

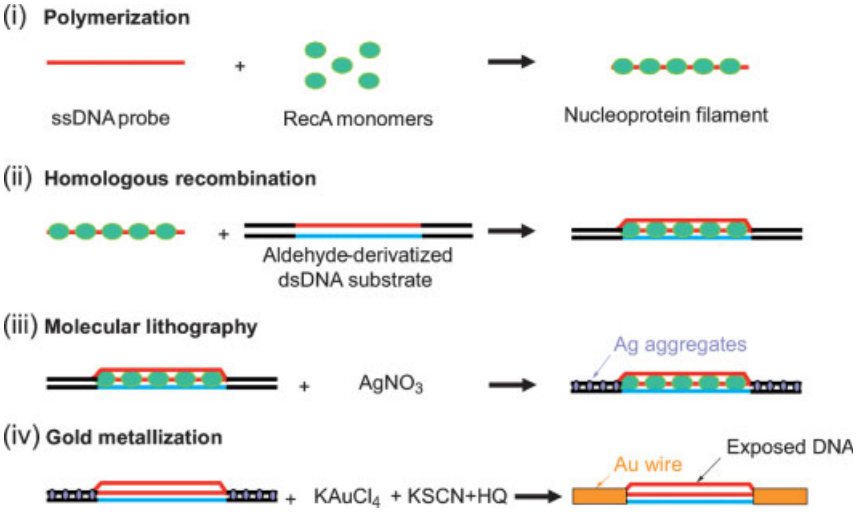


Figure 9.5 Schematics of the homologous recombination reaction and molecular lithography. (i) RecA monomers polymerize on a ssDNA probe molecule to form a nucleoprotein filament. (ii) The nucleoprotein filament binds to an aldehyde-derivatized dsDNA substrate molecule at an homologous sequence. (iii) Incubation in $AgNO_3$ solution results in the formation of silver aggregates along the substrate molecule at regions unprotected by RecA. (iv) The silver aggregates catalyze specific gold deposition on the unprotected regions. A highly conductive gold wire is formed with a gap in the protected segment. (Reprinted from Ref [6]; © 2002, American Association for the Advancement of Science.).

The application of sequence-specific molecular lithography to the definition of a patterned gold wire is outlined in Figures 9.5 and 9.6. Here, the previously described DNA metallization scheme is employed in which DNA-bound glutaraldehyde is used as a localized reducing agent [6], and the RecA is used as a sequence-specific resist. RecA monomers polymerize on a single-stranded probe DNA to form a nucleoprotein filament (Figure 9.5, step i) which locates and binds to a homologous sequence on a dsDNA molecule (Figure 9.5, step ii). Once bound, the RecA in the nucleoprotein filament acts as a sequence-specific resist, physically protecting the aldehyde-derivatized substrate DNA against silver cluster formation in the bound region (Figure 9.5, step iii). Subsequent gold metallization leads to the growth of two extended DNA-templated wires separated by the predesigned gap (Figure 9.5, step iv).

Figure 9.6 depicts images of the products of the various steps leading to a patterned gold-coated λ-DNA. Extensive AFM and scanning electron microscopy (SEM) imaging confirmed that the metallization gap was located where expected. The position and size of the insulating gap could be tailored by choosing the probe's sequence and length. The ability to pattern DNA metallization facilitates modular circuit design, and is therefore valuable for the realization of DNA-templated electronics. Insulating and conducting regions can be defined on the DNA scaffold according to the underlying sequence, thus determining the electrical connectivity in the circuit. In addition, patterning DNA metallization is useful for the integration of molecular objects into a circuit. Such objects can be localized and electrically

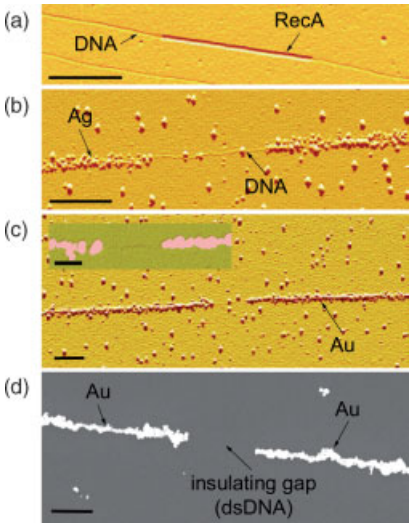


Figure 9.6 Sequence-specific molecular lithography on a single DNA molecule. (a) AFM image of a 2027-base RecA nucleoprotein filament bound to an aldehyde-derivatized λ-DNA substrate molecule. (b) AFM image of the sample after silver deposition. Note the exposed DNA at the gap between the silver-loaded sections. (c) AFM image of the sample after gold metallization. Inset: zoom on the gap. The height of the metallized sections is ~50 nm. (d) SEM image of the wire after gold metallization. All scale bars = 0.5 μm; inset to (c) = 0.25 μm. The variation in the gap length is due mainly to variability in DNA stretching on the solid support. The very low background metallization in the SEM image compared with the AFM images indicates that most of the background is insulating. (Reprinted from Ref [6]; © 2002, American Association for the Advancement of Science.).

contacted within the exposed DNA sequences present in the unmetallized gaps. Further manipulations of DNA templates including the localization of man-made objects at specific addresses along the DNA molecule, the generation of three- and four-armed junctions, and elaborate metallization patterning can be found in Refs. [6, 7, 14, 16].

9.2.3
Self-Assembly of a DNA-Templated Carbon Nanotube Field-Effect Transistor

The superb electronic properties of CNTs [23], their large aspect ratio, and their inertness with respect to the DNA metallization process, make them an ideal choice for the active elements in DNA-templated electronics. The ability to localize molecular objects at any desired address along a dsDNA molecule and to pattern sequence-specifically the DNA metallization (as described above) facilitate the incorporation of CNTs into DNA-templated functional devices, and their wiring. In the assembly of the field-effect transistor (FET), a DNA scaffold molecule provided the address for the precise localization of a semiconducting single-wall carbon nanotube (SWNT), and templated the extended wires contacting it. The localization of the SWNT relied on

homologous recombination by the RecA protein. The assembly of the SWNT–FET, which is shown schematically in Figure 9.7, employed a three-strand homologous recombination reaction between a long dsDNA molecule serving as a scaffold and a short auxiliary ssDNA. The short ssDNA molecule was synthesized so that its sequence was identical to the dsDNA at the designated location of the FET. RecA

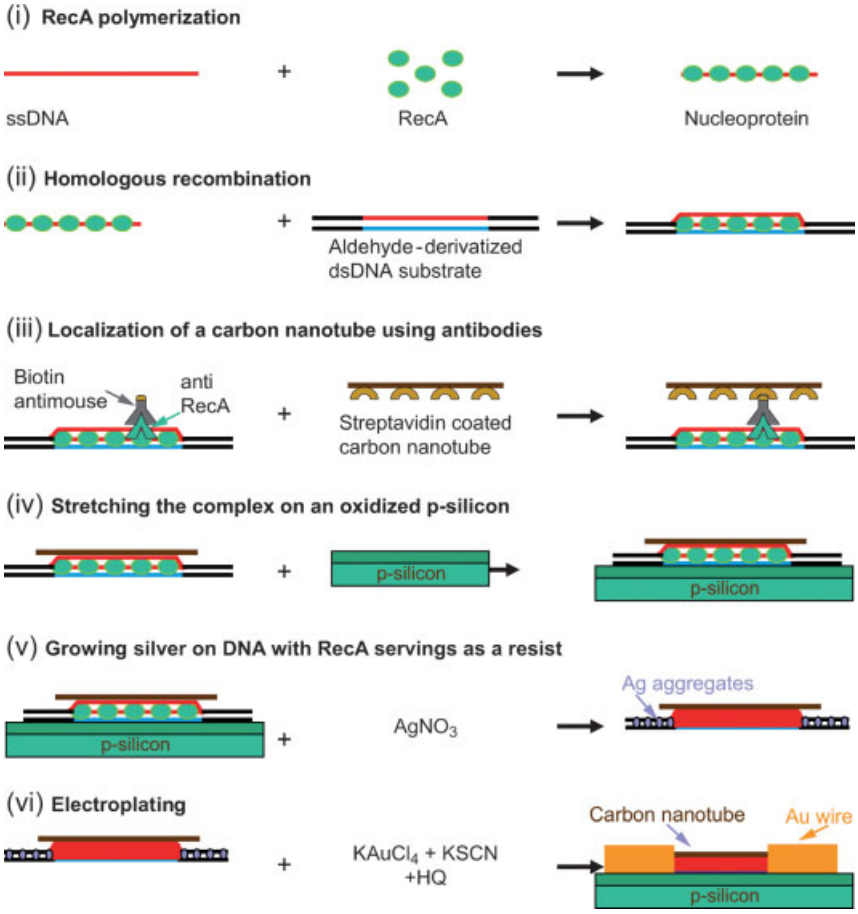


Figure 9.7 Assembly of a DNA-templated FET and wires contacting it. Steps are as follows. (i) RecA monomers polymerize on a ssDNA molecule to form a nucleoprotein filament. (ii) Homologous recombination reaction leads to binding of the nucleoprotein filament at the desired address on an aldehyde-derivatized scaffold dsDNA molecule. (iii) The DNA-bound RecA is used to localize a streptavidin-functionalized single-wall carbon nanotube (SWNT), utilizing a primary antibody to RecA and a biotin-conjugated secondary antibody. (iv) The complex is stretched on an oxidized *p*-type silicon wafer by dipping the substrate in a solution containing the complexes and pulling it out. (v) Incubation in an $AgNO_3$ solution leads to the formation of silver clusters on the segments that are unprotected by RecA. (vi) Electroless gold deposition, using the silver clusters as nucleation centers, results in the formation of two DNA-templated gold wires contacting the SWNT bound at the gap.

proteins were first polymerized on the auxiliary ssDNA molecules to form nucleoprotein filaments (Figure 9.7, step i), which were then mixed with the scaffold dsDNA molecules. The nucleoprotein filament bound to the dsDNA molecule according to the sequence homology between the ssDNA and the designated address on the dsDNA (Figure 9.7, step ii). The RecA later served to localize a SWNT at that address and to protect the covered DNA segment against metallization. A streptavidin-functionalized SWNT was guided to the desired location on the scaffold dsDNA molecule using antibodies to the bound RecA and biotin–streptavidin-specific binding (Figure 9.7, step iii). The SWNTs were solubilized in water by micellization in sodium dodecyl sulfate (SDS) [24] and functionalized with streptavidin by non-specific adsorption [25, 26].

Primary anti-RecA antibodies were reacted with the product of the homologous recombination reaction, and this resulted in specific binding of the antibodies to the RecA nucleoprotein filament. Next, biotin-conjugated secondary antibodies, having high affinity to their primary counterparts, were localized on the primary anti-RecA antibodies. Finally, the streptavidin-coated SWNTs were added, leading to their localization on the RecA via biotin–streptavidin-specific binding (Figure 9.7, step iii). The DNA/SWNT assembly was then stretched on a passivated oxidized silicon wafer. An AFM image of a SWNT bound to a RecA-coated 500-base-long ssDNA localized at the homologous site in the middle of a scaffold λ-DNA molecule is shown in Figure 9.8a. The conducting CNT can be clearly distinguished from the insulating DNA by the use of scanning conductance microscopy [27, 28]. The topographic and conductance images of the same area are depicted in Figure 9.8b and c, respectively. The evident difference between the two images identifies the SWNT on the DNA molecule. It should be noted that the CNT is aligned with the DNA, which is almost always the case due to the stiffness of the SWNT and the stretching process.

Following stretching on the substrate, the scaffold DNA molecule was metallized. The RecA, doubling as a sequence-specific resist, protected the active area of the transistor against metallization. The metallization scheme described above was employed, in which aldehyde residues, acting as reducing agents, were bound to the scaffold DNA molecules by reacting the latter with glutaraldehyde. Highly conductive metallic wires were formed by silver reduction along the exposed parts of the aldehyde-derivatized DNA (Figure 9.7, step v) and subsequent electroless gold plating using the silver clusters as nucleation centers (Figure 9.7, step vi). As the SWNT was longer than the gap dictated by the RecA, the deposited metal covered the ends of the nanotube and contacted it. A SEM image of an individual SWNT contacted by two DNA-templated gold wires is depicted in Figure 9.8d.

The extended DNA-templated gold wires were contacted by electron-beam lithography, and the device was characterized by direct electrical measurements under ambient conditions. The p-type substrate was used to gate the transistor. The electronic characteristics of the device are shown in Figure 9.9a and b. The gating polarity indicated p-type conduction of the SWNT, as is usually the case with semiconducting CNTs in air [29]. The saturation of the drain-source current for negative gate voltages indicated resistance in series with the SWNT; this

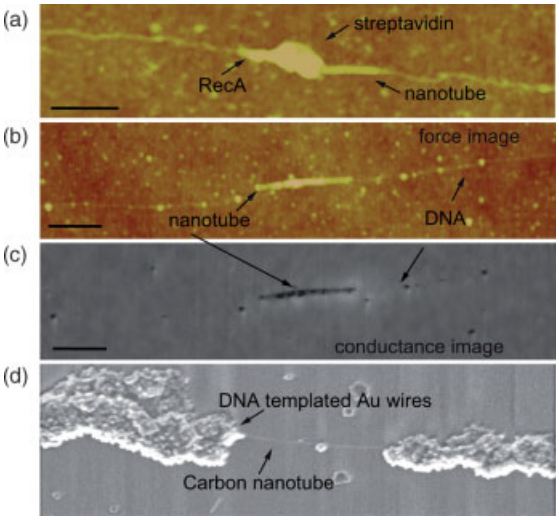


Figure 9.8 Localization of a single-wall carbon nanotube (SWNT) at a specific address on the scaffold dsDNA molecule using RecA. (a) An AFM image of a 500-base-long (~250 nm) RecA nucleoprotein filament localized at a homologous sequence on a λ-DNA scaffold molecule. Scale bar = 200 nm. (b) An AFM image of a streptavidin-coated SWNT bound to a 500-base-long nucleoprotein filament localized on a λ-DNA scaffold molecule. Scale bar = 300 nm. (c) A scanning conductance image of the same region as in (b). The conductive SWNT yields a considerable signal, whereas the insulating DNA is hardly resolved. Scale bar = 300 nm. (d) SEM image of the resulting device. The DNA-templated gold wires and the assembled nanotube are indicated by arrows. The DNA molecule itself is not resolved in this image.

resistance was attributed to the contacts between the gold wires and the SWNT as the four-terminal resistance of the DNA-templated gold wires was typically smaller than 100 Ω. Each of the different devices had somewhat different turn-off voltages.

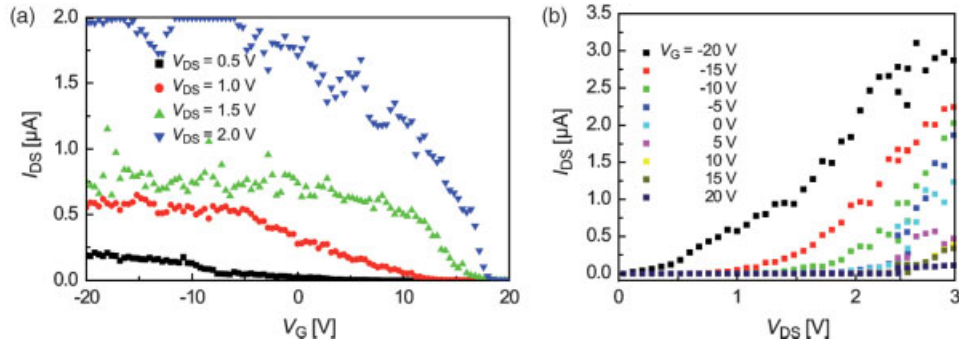


Figure 9.9 Electrical characteristics of a self-assembled *p*-type field effect transistor based on a semiconducting single wall carbon nanotube. (a) Drain-source current versus gate bias applied between the *p*-type substrate and the source electrode. (b) Same versus drain-source voltage for different gate voltages.

9.3
Recognition of Electronic Surfaces by Antibodies

The self-assembly described in Section 9.2 relies on existing biotechnological tools, but in this section the toolbox is expanded to show how to isolate antibody molecules that recognize electronic surfaces directly. As a specific example, the isolation of antibody molecules capable of discriminating between different crystalline facets of a GaAs crystal is reviewed. Beyond the potential application of such antibodies for the direct localization of molecular-scale objects at desired sites on an electronic substrate, the success in isolating these antibodies is encouraging with regards to the prospects of isolating antibodies that can "read" electrical signals presented to them by electronic devices. The latter constitute a critical milestone on the way to *functional* integration between molecular biology and nanoelectronics.

The mammalian immune system offers a vast repertoire of antibody molecules capable of binding, in selective manner, an immense number of molecules presented to the body by invading pathogens such as bacteria, viruses, and parasites. Although this repertoire has evolved to target mostly biomolecules, it may potentially contain selective binders to other targets, or it may be expanded to include such binders. Indeed, the injection into mice of cholesterol and 1,4-dinitrobenzene [30, 31] microscopic crystals, as well as C_{60} conjugated to bovine thyroglobulin [32] have resulted in the generation of antibodies against these materials by the immune system of the injected animal. Here, the scope of the system is expanded, and it is shown that human antibody libraries – specifically, single-chain Fv (scFv; [33]), which are the antibody variable binding domains – contain specific binders, capable of discriminating between different crystalline facets of a GaAs semiconductor crystal, which is an almost flat target and unfamiliar to the immune system. This selectivity is remarkable given the very simple structure of semiconductors compared with biomolecules.

By using phage display technology, the *in-vitro* isolation of scFv that bind GaAs (1 1 1A) facets almost 100-fold better than GaAs (1 0 0). is demonstrated. More generally, this finding implies that antibody molecules may find application in the assembly of nanoelectronics [1, 6, 8], in the production of templates for localizing nanoparticles [34], or for biosensors [35].

The isolation of short peptides that bind inorganic materials has been demonstrated for gold [36, 37], silver [38], silica [39], metal oxides [40, 41], minerals [42], CNTs [43], and various semiconductors [44–46]. Of these reports, Ref. [46] is particularly relevant to the present section, as the authors report the isolation of peptides (by phage display) that bind GaAs (1 0 0) preferentially to GaAs (1 1 1A) and (1 1 1B). However, all assays in Ref. [46] probed the peptides displayed on the phages rather than the free peptides. Indeed, when one of these peptides was later synthesized and applied to GaAs [47], no selectivity was found between the (1 0 0) and (1 1 0) facets (see Figure 5 in Ref. [47] and the following discussion).

As this discrepancy was difficult to comprehend, attention was towards studying the non-specific binding of M13 phages (which carried no peptides or antibodies) to GaAs (1 0 0), GaAs (1 1 1A) and GaAs (1 1 1B). As a consequence, M13

was found to bind preferentially to the (1 0 0) facet through its coat protein (Figure 1S, supplementary material to Ref. [9]). As those phages were identical to the library phages used in Ref. [46], and given the lack of selectivity displayed by the only free peptide tested thus far [47], it seemed that further experiments with free peptides would be needed in order to either confirm or disprove semiconductor facet recognition by short peptides. In contrast to Ref. [46], the present antibodies were also tested and found to be selective towards crystal orientation when detached from the phage.

The 7- and 12-mer peptides used in most *in-vitro* selections of binders to inorganic crystals are typically too short to assume a stable structure. Antibodies on the other hand, display a rigid three-dimensional (3-D) structure which is potentially essential for high-affinity selective binding [30, 31]. Moreover, the recognition site in the latter case involves six amino acid sequences grouped into three complementarity-determining regions (CDR). All together, these CDRs form a large, structured binding site spanning up to 3×3 nm. The critical role of the antibody 3-D structure for the recognition of organic crystal facets is well established [30, 31].

Another hint to the importance of rigidity for facet recognition is provided by the rigid structure characterizing antifreeze peptides that target specific ice facets [48]. It has also been shown that the stable helical structure of a 31-mer peptide catalyzing calcite crystallization is essential for inducing directed crystal growth along a preferred axis [49], possibly due to its differential binding to the various facets. Hence, structure rigidity may turn central to facet recognition by biomolecules, thereby underscoring the importance of antibody libraries as a promising source for selective binders.

Selective binding to specific crystalline facets can be directly utilized for numerous micro- and nanotechnological applications, including the positioning of nanocrystals at a well-defined orientation, governing crystal growth and forcing it to certain directions [49], and positioning nanometer-scale objects at specific sites on a substrate marked by certain crystalline facets. An application of one of these soluble antibodies to the latter task is demonstrated in Figure 9.10. By using conventional photolithography and $H_3PO_4 : H_2O_2 : H_2O$ etching, a long trench has been defined on a GaAs (1 0 0) substrate in the (1 1 0) direction (Figure 9.10a). Due to the slow etching rate of phosphoric acid in the (1 1 1A) direction, the process leads to slanted (1 1 1A) side walls and a flat (1 0 0) trench floor (Figure 9.10a). A SEM image of a cut across the trench, and proving that the slanted walls are indeed tilted in the (1 1 1A) direction (54.7° relative to the (1 0 0) direction), is depicted in Figure 9.10b. When the isolated scFv antibodies are applied to the GaAs substrate they attach themselves selectively to the (1 1 1A) slopes.

In order to image the bound antibody molecules, they were targeted with anti-human secondary antibodies conjugated to a fluorescent dye, Alexa Fluor. As shown in Figure 9.10c, fluorescence is limited solely to the (1 1 1A) slopes with practically no background signal coming from the (1 0 0) surfaces. Control experiments depleted of the scFv fragments exclude possible artifacts such as the natural fluorescence of the (1 1 1A) facets, selective binding of the fluorescent dye, or secondary antibodies to that facet.

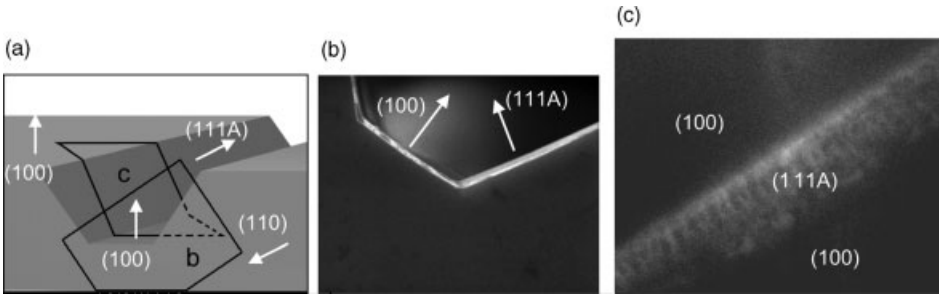


Figure 9.10 (a) Diagrams of the etched trench labeled with the various crystalline facets. The black frames correspond to the views depicted in panels (b) and (c). (b) SEM image of a cut across the trench. (c) Fluorescence image of the trench viewed from the top. Fluorescence is confined to the (1 1 1A) slopes, proving selective binding of the scFv fragments to that facet. Note the negligible binding of antibody molecules to the (1 0 0) facets. (Reprinted from Ref. [9]; © 2006, American Chemical Society.).

The images in Figure 9.10 prove that the selected scFv antibody molecules recognize and bind selectively GaAs (1 1 1A) as opposed to GaAs (1 0 0). As such, they can be used to localize practically any microscopic object on (1 1 1A) surfaces, with negligible attachment to other crystalline facets. The isolation of such binders using phage display technology, and the quantification of their selectivity, is described in the following section.

The Ronit1 scFv antibody phage library [50] used in the present study, is a phagemid library [51] comprising 2×10^9 different human semi-synthetic single-chain Fv fragments, where *in-vivo*-formed CDR loops were shuffled combinatorially onto germline-derived human variable region framework regions of the heavy (V_H) and light (V_L) domains.

To select scFv binders to GaAs (1 1 1A), approximately 10^{11} phages ($\approx$100 copies of each library clone) were applied to the semiconductor crystal (panning step). After washing the unbound phages, the bound units were recovered by rinsing the sample in an alkaline solution. The recovered viruses were then quantified by infecting bacteria and plating dilution series on Petri dishes. The amplified sublibrary was applied again to the target crystal facet, and so on. Typically, three to four panning rounds were required to isolate excellent binders to the target. As is evident from Figure 9.11, the number of bound phages retrieved from the semiconductor grew 300-fold when panning was repeated three times. For comparison, the non-specific binding of identical phages (M13) carrying no scFv fragments remained low throughout the selection process. It was found experimentally that blocking with milk was essential to prevent the non-specific binding of phages to the GaAs targets. Interestingly, as shown in the supplementary material to Ref. [9], in the absence of blocking against non-specific binding (a step missing in Ref. [46]), the non-specific binding of phages through their coat protein to GaAs (1 0 0) was larger than to GaAs (1 1 1A). The data in Figure 9.11 prove the selection of increasingly better binders to GaAs (1 1 1A), but provide no indication of selectivity with respect to GaAs (1 0 0).

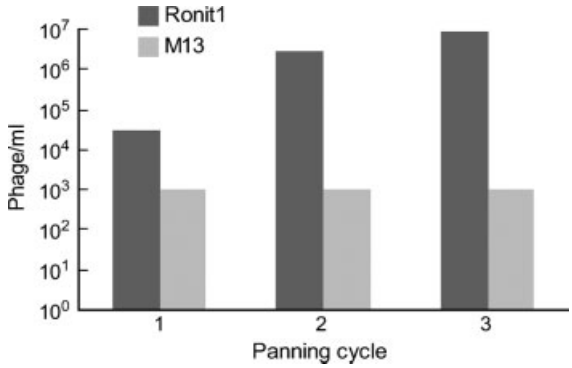


Figure 9.11 Enrichment of anti-GaAs (1 1 1A) phages carrying scFv fragments versus panning cycle. Phage concentration has been deduced by counting colonies of *E. coli* bacteria infected with different dilutions of the phages recovered after each cycle. The monotonic increase in binding of phages carrying scFv (Ronit1) is contrasted with the much weaker, non-specific binding of similar phages lacking the scFv antibody. The value of the latter (1000 phages mL^{-1}) sets an experimental upper limit on their binding; the actual values are likely to be smaller. (Reprinted from Ref. [9]; © 2006, American Chemical Society.).

Indeed, as indicated by the two left-hand columns of Figure 9.12, application of the polyclonal population of binders selected on GaAs (1 1 1A) to GaAs (1 0 0) shows similar binding to the latter crystalline facet. Hence, the process described above produced good, but non-selective, binders.

Preferential binding to a given crystalline facet was achieved by a slight modification of the process. The phages recovered from the first panning on GaAs (1 1 1A)

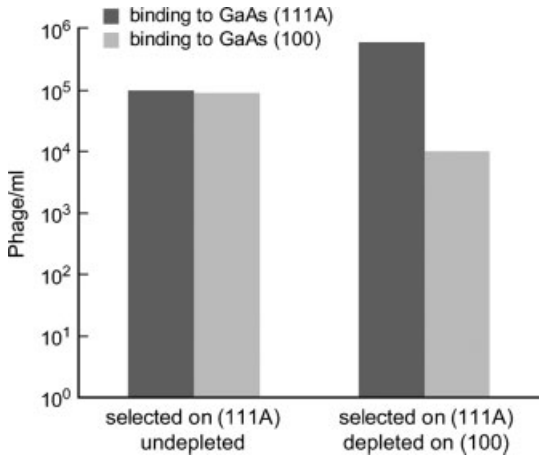


Figure 9.12 Density of recovered binders to GaAs (1 1 1A) after three panning cycles. The two right-hand (or left-hand) columns correspond to selection on GaAs (1 1 1A) with (or without) depletion on GaAs (1 0 0). (Reprinted from Ref. [9]; © 2006, American Chemical Society.).

were amplified in *E. coli* and then applied to GaAs (1 0 0). However, this time the *unbound* phages were collected and applied in a second panning step to GaAs (1 1 1A). As evident from the two right-hand columns of Figure 9.12, the "depletion" step on GaAs (1 0 0) is enriched for specific phage clones that both bind GaAs (1 1 1A) and lack binding to GaAs (1 0 0). On this occasion, binding of the selected phages to the (1 1 1A) facet was almost 100-fold higher than to the (1 0 0) facet. This depletion step, which was crucial to the present case, was missing in Ref. [46].

The polyclonal population of selected phages contains different scFv fragments, each characterized by different affinity and selectivity to the two crystalline facets. In order to correlate specificity with sequence, the binding selectivity of the individual clones was next analyzed. Monoclonal binders were isolated by infecting *E. coli* bacteria with the sublibrary and plating them on solid agar. As each bacterium can be infected by a single phage, all bacteria within a given colony carry DNA coding for the same scFv fragment. Infection of the colony with helper phages resulted in the release of phages displaying the same scFv on their PIII coat proteins. The isolated monoclonal phages were then analyzed with ELISA against GaAs (1 1 1A) and (1 0 0). The sequences of the light (V_L) and heavy (V_H) CDRs of ten monoclonal binders that were identified by the ELISA assay can be found in Ref. [9] and its supplementary material, together with a discussion of their main features.

Figures 9.11 and 9.12 correspond to the scFv fragments displayed on phage particles. For practical applications (such as that demonstrated in Figure 9.10) it is preferable to have soluble monoclonal scFv fragments detached from the phage coat proteins. The results of the ELISA assays of the scFv fragment of Figure 9.10, in its soluble form, are presented in Figure 9.13.

In Figure 9.13, bars 1–6 correspond to the six ELISA assays on GaAs (1 1 1A) and GaAs (1 0 0) pieces, each of 4×4 mm. After washing the substrates, the bound antibodies were reacted with anti-human horseradish peroxidase (HRP), and the binding was quantified by adding tetramethylbenzidine (TMB) as a colorimetric substrate, and reading the resulting optical density (OD) at 450 nm. Bars 7–9

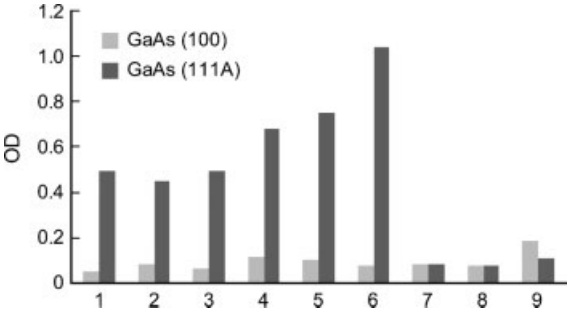


Figure 9.13 Bars 1–6 display the results of six comparative ELISA assays of the scFv molecule (detached from the phage) on GaAs (1 1 1A) and GaAs (1 0 0) substrates. The optical density (OD) reflects the number of bound molecules in arbitrary units. Bars 7–9 display the results of three control experiments (see text), and can be used to estimate the background signal (ca. 0.1 OD) coming from sources, other than selective binding of the scFv to the semiconductor substrates. (Reprinted from Ref. [9]; © 2006, American Chemical Society.).

provided the following controls. Bars 7 quantified the non-specific binding of the secondary anti-human HRP to the ELISA plate in the absence of the EB scFv and semiconductor substrates. Bars 8 corresponded to the non-specific binding of the scFv to the plate, and bars 9 to non-specific binding of the secondary antibodies to the semiconductor substrates.

The background ELISA signal, depicted by bars 7–9, accounts for most of the GaAs (1 0 0) signal in columns 1–6. When subtracting this background from columns 1 to 6, a remarkable preference is found to GaAs (1 1 1A) compared with (1 0 0). Interestingly, the binding of the secondary antibody to GaAs (1 0 0) was almost twice as large compared to its binding to GaAs (1 1 1A), in opposition to the selectivity of the isolated scFv fragments. Overall, the data in Figure 9.13 prove that the isolated scFv preserves its selectivity also when detached from the phage.

Little is known of the interaction between biomolecules and inorganic surfaces, let alone the recognition of such surfaces by antibody molecules. The GaAs surface is modified by surface reconstruction, oxidation, and possibly other chemical reactions. Moreover, it displays atomic steps and possibly surface defects. It is therefore difficult to estimate how much of the underlying crystalline order manifests itself in the recognition process. Unfortunately, as no experimental tools capable of determining these parameters with atomic resolution exist at present, the recognition mechanism is unclear, except for the accumulating indications of the importance of structural rigidity (as discussed in the introduction to this section). The discrimination between the two crystalline facets may reflect the different underlying crystalline structures, they may stem from the different surface chemistries of the two facets, or they may result from global properties such as atom density and different electronegativity. The latter factor has been found to be important for the differential binding of short peptides to different semiconductors [47]. The abundance of positively charged amino acids in the heavy chain of CDR1 and CDR3 and the light chain of CDR1 may indicate an affinity to the exposed gallium atoms. The negatively charged amino acid in CDR3 V_L (missing in anti-gold scFv isolated from the same library) combined with the positively charged CDR3 V_H may match the polar nature of GaAs.

The recognition of man-made materials by antibodies opens new opportunities for a *functional* interface between biology and nanotechnology, far beyond what was has been exercised to date.

9.4
Molecular Shift-Registers and their Use as Autonomous DNA Synthesizers [11]

9.4.1
Molecular Shift-Registers

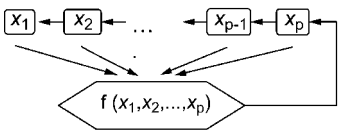
The DNA-templated assembly of elaborate circuits requires distinct dsDNA molecules with non-recurring sequences. For the assembly of periodic structures, such as memories, a segment of non-recurring sequences should be replicated to form a periodic molecule, and the synthesis of such molecules presents a remarkable

challenge to biotechnology. The two existing strategies for generating long molecules, namely PCR [52] and ligase assembly [53], utilize synthetic oligonucleotides which together span (with overlap) the full length of the desired molecule. Hence, when following any of these strategies, the assembly of an N-base long molecule with distinct p-long segments requires $O(N/p)$ oligonucleotides. These approaches therefore quickly become impractical when a rich variety of distinct molecules or addresses along a given molecule are needed for the construction of an elaborate template for molecular electronics [1, 5]. Motivated by the concept of DNA-templated electronics, the present author and colleagues were therefore forced to invent an exponentially more economic synthesis strategy based on the chemical realization of molecular SRs. The dramatic reduction in synthesis effort by SRs is facilitated by exploiting a novel concept in DNA synthesis; a sliding overlapping reading frame. Rather than the fixed frame that directs segment ligation or polymerization in the two schemes listed above or in hairpin-based DNA logic [54, 55] and programmed mutagenesis [56], the SRs utilize a previously synthesized sequence to dictate synthesis of the next bases. The automaton is an example of DNA computing where the result of the computation (tape) is a useful molecule.

An autonomous binary p-shift register (p-SR) is a computing machine with 2^p internal states represented by an array of p cells (Figure 9.14a), each occupying one bit, $x_i\,\{i = 1 \ldots p\}$. In each step a binary function, $f(x_1, x_2, \ldots x_p)$, is computed and its value is inserted into cell p. Simultaneously, x_j is shifted to cell $j-1$; $\{j = 2 \ldots p\}$. On printing x_1 to a tape, a long periodic binary sequence is generated. Electronic SRs are utilized in many applications including secure communication, small signal recovery, and sequence generation [57]. Here, it is shown that molecular SRs can be realized and utilized for the autonomous synthesis of DNA molecules the sequence of which is uniquely determined by a chemical embodiment of the function $f(x_1, x_2, \ldots x_p)$.

Consider a 3-SR with $x_{n+1} = f(x_{n-2}, x_{n-1}, x_n) = x_{n-2} \oplus x_n$ ($\oplus \equiv$ XOR) and an initial setting (seed) $x_1, x_2, x_3 = 001$. Repetitive application of f generates the sequence 001110100111010.... The sequence is periodic with a period seven and any of the seven $L \geq 3$ bit long consecutive strings in a period is different from the rest. In general, it is well known [57] that for any p, a SR can be found a with a linear feedback function [58], $f = \sum_{i=1}^{p} \alpha_i x_i; \alpha_i \in \{0, 1\}$ (the sum is mod 2), that generates a sequence

(a)



(b)

x_1	x_3	f	Rule Strand
0	0	0	$\overline{0}\,1\,\overline{0}\,\overline{0}$
0	1	1	$\overline{0}\,\overline{0}\,1\,\overline{1}$ $\overline{0}\,1\,\overline{1}\,\overline{1}$
1	0	1	$\overline{1}\,\overline{0}\,\overline{0}\,1$ $\overline{1}\,\overline{1}\,\overline{0}\,1$
1	1	0	$\overline{1}\,\overline{0}\,\overline{1}\,\overline{0}$ $\overline{1}\,\overline{1}\,\overline{1}\,\overline{0}$

Figure 9.14 (a) An autonomous binary p-shift register. (b) Truth table and rule strands corresponding to the first example. (Reprinted from Ref. [11]; © 2006, American Physics Society.).

of maximal period $2^p - 1$ bits with no repetition of strings of lengths $L \geq p$ within a period. Such SRs are termed "maximal linear SRs" as they generate all possible permutations of a p-long sequence except the zero string [59]. Functions, f, can always be found such that the number of non-vanishing α_i is smaller than p (in the example above, $\alpha_2 = 0$). Consequently, $2^p - 1$ different addresses can be generated by an exponentially smaller truth table and, hence, as shown below, by an exponentially smaller synthesis effort compared with direct synthesis of all addresses.

We now show how to implement an autonomous molecular SR using DNA. Imagine a DNA molecule for which the Watson–Crick rules are that 1 binds exclusively to its complementary bit, $\bar{1}$, but not to 1, 0 or $\bar{0}$. Similarly, 0 binds to $\bar{0}$ but not to 0, $\bar{1}$, or 1. We translate the function $f(x_1, x_2, x_3) = x_1 \oplus x_3$ to an equivalent truth table (left three columns in Figure 9.14b) and embody it by the mixture of the seven [59] possible four-bit rule strands, $(\bar{x}_1, \bar{x}_2, \bar{x}_3, (\overline{x_1 \oplus x_3}))$, listed in the right-hand column of Figure 9.14b.

The SR sequence is generated by thermally cycling a mixture containing the seven rule strands, a "seed" strand (e.g., the strand 001), and a polymerase. For simplicity, it is assumed that the rule strands are synthesized with ddDNA at their 3′ end and are therefore not elongated in the process. Each cycle comprises annealing, extension, and melting steps. In the first annealing step, some of the first, $\bar{0}\bar{0}\bar{1}\bar{1}$, rule strands bind to seed molecules, leaving an $\bar{1}$ overhang (Figure 9.15, steps i and ii) which is readily copied by the polymerase in the extension step (Figure 9.15, step iii). Next (Figure 9.15, step iv), the temperature is raised to 95 °C and the rule strand dissociates from the elongated seed (tape). In the second annealing step, a $\bar{0}\bar{1}\bar{1}\bar{1}$ rule strand binds to the tape (Figure 9.15, step v), leaving again an $\bar{1}$ overhang which is readily copied by the polymerase (Figure 9.15, step vi). At each additional cycle (Figure 9.15, steps vii–x) some

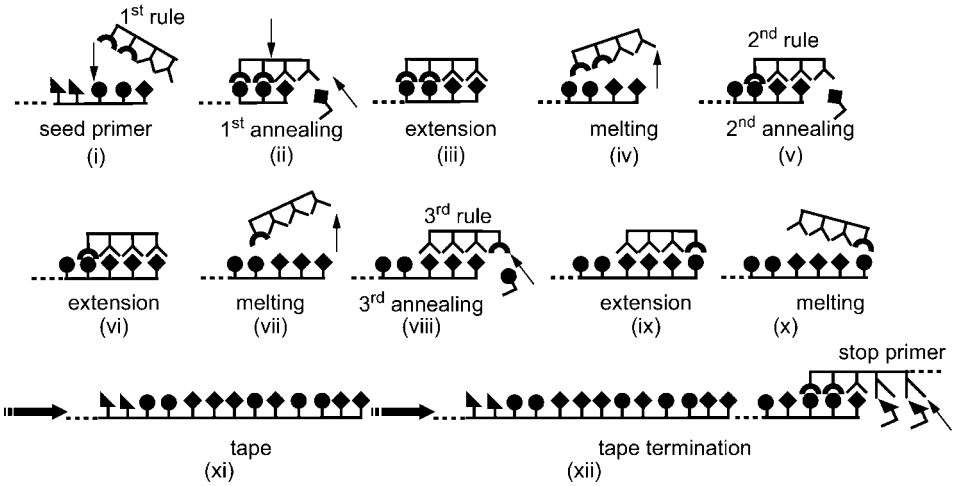


Figure 9.15 The principle of the shift register. ◆ ◠ ● ◠ represent 1, $\bar{1}$, 0, $\bar{0}$, respectively. ◣ ◥ represent sequences other than 0 or 1 and their complementary sequences, respectively. (Reprinted from Ref. [11]; © 2006, American Physics Society.).

of the tape molecules are elongated by one bit according to the rule $x_{n+1} = x_n \oplus x_{n-2}$. Elongation is terminated by addition of excess stop primers that intercept the tape molecules as soon as the latter display a desired tail (001 in the example of Figure 9.15 step xii). The polymerase then copies the stop primer and adds its alien sequence to the tape, which is unrecognizable by any rule strand. As a result, elongation terminates. The 5′ seed and 3′ stop primers tails are later used for PCR amplification of the tape. Elongation is guided by a sliding reading frame where all, except the first, shifted bits from the previous reading frame plus a single new bit provide the current reading frame. The sliding frame is the crux of our concept, as it facilitates exponentially smaller synthesis effort compared with any of the previous, fixed-frame approaches.

It should be noted that rule strands are not consumed during synthesis; rather, they only serve as enzymes to direct the reaction. Thus, synthesis in flow may be envisioned, where the rule strands are attached in synthesis order to subsequent segments of a tube or a column. While the reactants flow through the tube the correct sequence is generated, and this strategy is advantageous to straightforward synthesis in a DNA synthesizer as faulty strands are not recognized (and hence not elongated) by rule strands. Clearly, errors are doomed to be short.

Now, an actual demonstration of the concept may be described. In the first implementation each bit is realized by a sequence of three nucleotides, 5′TGC for "0" and 5′GCT for "1". These sequences were chosen as they minimize errors due to one and two base shifts in the annealing step. The demonstration starts with the three-bit maximal SR as discussed above. Such SR requires seven 4-bit (3-bit rules plus one function bit) strands (Figure 9.14b), but in order to suppress synthesis errors longer, redundant 6-bit rules (5-bit rules plus one function bit) are employed. Error suppression by redundancy is discussed in Section 9.4.2. The seven 6-bit rule strands [60] used in the synthesis comprise, $3'\bar{0}0\bar{1}\bar{1}\bar{1}\bar{0}$, $3'\bar{0}\bar{1}\bar{1}\bar{1}0\bar{1}$, $3'\bar{1}\bar{1}\bar{1}0\bar{1}0$, $3'\bar{1}\bar{1}0\bar{1}0\bar{0}$, $3'\bar{1}0\bar{1}0\bar{0}\bar{1}$, $3'\bar{0}\bar{1}0\bar{0}\bar{1}\bar{1}$, $3'\bar{1}\bar{0}\bar{0}\bar{1}\bar{1}\bar{1}$. The complementary bits, $\bar{0}$ and $\bar{1}$, correspond to 3′ACG and 3′CGA, respectively. The rule strands are synthesized with three nucleotides only (G, C, A) in order to prevent their extension by polymerase ("poor man's ddDNA"). The 2/3 GC content gives [61] $\Delta G \approx 8.5 \div 10.5\, k_B T$ free energy per bit (stacking included) which in a 5-bit realization of a 3-bit SR translates ideally to suppression of the error rate by a factor proportional to $\exp(-3\Delta G/k_B T) \leq \exp(-25.5)$ (see Section 9.4.2). The seed strand comprises a 5′ tail followed by a 5-bit sequence [62], 5′GCATGCGCCCGTCAGGCG00111. The tail is later used to amplify the SR sequence by PCR. The seed, the rule strands, and three nucleotides (dGTP, dCTP, dTTP) are mixed together and subjected to 45 thermal cycles [63], after which a stop primer, $3'\bar{0}1\bar{0}\bar{0}1$GACGTC, is added in 10-fold excess compared with each rule strand. During an additional five to ten cycles the tape molecules are further elongated until in some cycle their last five bits read 01001. At that point a stop primer binds to the tape and its complementary sequence is added to the tape by the polymerase. The elongation now terminates as the sequence added by the stop primer is alien to all rule strands. The absence of dATP guarantees single strand synthesis. The expected synthesized sequences read

$$\underset{\text{seed primer}}{5'\text{GCATGCGCCCGTCAGGCG00111}}\,(0100111)_n \mapsto \underset{\text{complementary to stop primer}}{01001\text{CTGCAG}} \quad \text{with } n = 0, 1, \ldots \tag{9.1}$$

Finally, the elongation products are PCR amplified with two primers, identical to the first 19 nucleotides of the seed (5′GCATGCGCCCGTCAGGCGT) and to the last 19 nucleotides of the stop primer (5′CTGCAGAGCGCAGCAAGCG).

The resulting PCR products, when run against a standard ruler in a polyacrylamide gel, are depicted in Figure 9.16a. Four bands corresponding to Equation 9.1 with $n = 0, 1, 2, 3$ are clearly resolved. Sequencing of the four bands with a primer identical to the first 19 nucleotides on the 5′ end of the seed primer proves the bands identification with the respective n values in Equation 9.1. The high fidelity of the automaton is reflected in the perfect matching of the sequencing with Equation 9.1, and the absence of any unexpected bands.

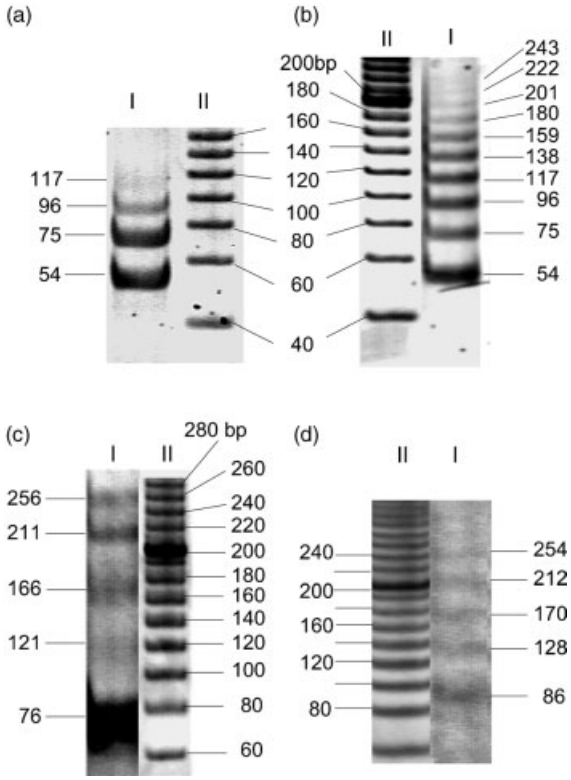


Figure 9.16 (a) Lane I, product after 45 elongation cycles, five cycles with stop primer, and PCR amplification. The four bands correspond to Eq. 9.1 with $n = 0, 1, 2, 3$, namely 54, 75, 96, and 117 base-long sequences. Lane II, ruler. (b) Lane I, same as (a), but with 100 elongation cycles followed by filtering out short sequences (Microcon YM-10; Millipore Corporation, Bedford, MA, USA). Ten bands corresponding to Eq. 9.1 with $n = 0, 1, 2, 3, 4, 5, 6, 7, 8, 9$ are resolved. Lane II, ruler. (c) Four -shift register with 45 bp periodicity realized with 7-bit rule strands; 2 h reaction time at a constant temperature of 72 °C. Lane I, shift register product. The five resolved bands are indicated. Lane II, ruler. (d) Same as (c) for partial 3-shift register with four-letter alphabet. The period comprises 14 bits (42 bp). (Reprinted from Ref. [11]; © 2006, American Physics Society.).

As shown in Figure 9.16b, after 100 elongation cycles it was possible to resolve 10 bands, $n = 0 - 9$, corresponding to 54, 75, 96, 117, 138, 159, 180, 201, 222, and 243 base-long sequences. The automaton thus synthesizes at least 204 bases at a remarkable fidelity. The $n = 9$ sequence comprises 10 periods, each of 21 bases, with exactly one repetition of each 3-bit (or longer) address per period. Direct sequencing of the bands confirmed the results up to $n = 6$. The small material quantities in the higher bands were insufficient for reliable sequencing. As PCR amplification favors shorter sequences, the relative band brightness cannot be taken as a measure for synthesis efficiency of molecules with different n-values.

The synthesis of longer period molecules, as well as of non-binary sequences, is demonstrated in Figure 9.16c and d, and details can be found in Ref. [11]. The synthesis of the last two examples was held in a thermal ratchet mode at a fixed temperature [11].

9.4.2
Error Suppression and Analogy Between Synthesis and Communication Theory

As emphasized in the introduction, errors are intrinsic to molecular assembly. Thus, the invention of error correction and suppression codes is critical for the realization of complex structures. Since the introduction of DNA computing by Adleman [53], the intimate relationship between self-assembly and computation has been slowly revealed. At this point, it may be beneficial to highlight another intriguing link between self-assembly and an engineering concept, this time "communication theory". This link draws an analogy between the synthesis of the long DNA molecule by the SR apparatus and the decoding of a long message transmitted over noisy lines. The addition of a wrong DNA base in synthesis is equivalent in that analogy to the assignment of a wrong value to a bit read in a message. This analogy has far-reaching consequences, as it suggests that some of the powerful strategies developed for suppressing and correcting errors in communication may be adapted to chemical synthesis. One such principle, the addition of degenerate bits to a message, is implemented in the synthesis of DNA molecules by the SRs.

The first stage is to classify any possible synthesis errors. At each annealing step, rule strands other than the correct ones may bind to the tape and affect the SR operation. These events may be divided into two groups: benign, and error. The benign events include all cases where rule strands bind to the tape either with no overhang or with a 2-bit overhang with the correct sequence. In the first case, the particular tape molecule remains idle throughout the cycle, whereas in the latter case it grows by two correct bits. Errors, on the other hand, are generated mostly by annealing of the wrong rule strand, shifted one bit to the right, to form a 2-bit overhang with the wrong sequence. It is easy to verify that, since the maximal register sequence contains all p-bit permutations, an error is manifested in a partial deletion of a period. The same fact guarantees that the tape is always "legal", namely it is available for elongation in the next cycle.

A SR is conveniently represented by a path on a corresponding de Bruijn graph [57], where the nodes depict all distinct internal states and the directed edges connecting them are labeled by the rules, notably by a string comprising the predecessor state plus a function bit. When a p-SR is realized with $p+1$ long rule strands, a maximal linear SR sequence passes exactly once through all nodes, except the zero node. A de Bruijn graph for a 3-SR is depicted in Figure 9.17, with arrows indicating the walk guided by Equation 9.1.

Although an elongation error corresponds to skipping some nodes, synthesis can always proceed as the rule strands recognize all nodes. When a p-SR is realized with rule strands of length $p'+1$; $p'>p$, as is the case here, the sequence passes exactly once through a subset of nodes in the much larger graph corresponding to p'-SRs. In the SR of Figure 9.16a and b, for instance, the 6-bit long rules correspond to a partial walk on de Bruijn graph of order 5 rather than 3. Two types of errors may then occur – a skip to a node in the sequence, or a skip to an alien node. In the first case, synthesis proceeds with partial deletion of the sequence. In the second case, the new node is not recognized by any rule strand and synthesis halts until that node is connected again to the SR sequence by an additional error. In both cases, each additional bit in the rule strands increases the Hamming distance for an error by at least 1 and, hence, suppresses the synthesis error rate by $\sim\exp(-\Delta G/k_BT)$. Optimization of the alphabet minimizes one- and two-base shift errors. Errors other than shifts, including hairpins, require further analysis.

The formation of an unwarranted 2-bit overhang can be minimized with respect to the desired 1-bit overhang by optimizing the temperature. Optimally, the error

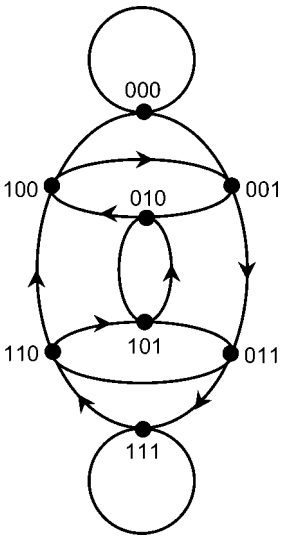


Figure 9.17 de Bruijn graph for a 3-shift register. The maximal path defined by Equation 9.1 corresponds to a walk on the graph (start from node 001 and follow the arrowheads).

rate (the ratio between incorrect and correct annealing) can be reduced in this way to $\approx\exp(-\Delta G/k_BT)$, where ΔG is the corresponding free energy per bit. The error rate may be systematically suppressed by using longer rule strands to generate the same sequence. By using de Bruijn graphs it can optimally be shown that each extra bit can reduce the error rate by an additional factor of $\approx\exp(-\Delta G/k_BT)$. This is the reason for the 6-bit long rule strands used in the realization of the 3-SR. The two extra bits are meant to suppress synthesis errors.

To the best of the present author's knowledge, this is the first incorporation of a redundancy code in chemical synthesis. The analogy drawn between chemical synthesis and transmission of messages over noisy lines suggests further applications of communication theory to chemical synthesis.

9.5 Future Perspectives

In Sections 9.2 to 9.4, a novel concept was outlined, namely the harnessing of the remarkable assembly strategies and tools of molecular biology to the self-assembly of molecular-scale electronics. Central issues such as instilling biomolecules with electrical conductance, molecular lithography for patterning metallization and localizing devices on DNA templates, the direct recognition of electronically relevant man-made objects by biomolecules, and the economic synthesis of DNA molecules characterized by non-recurring sequences have now been resolved to a point where the formidable challenge of complex self-assembly can be faced with confidence. However, harnessing the power of bioassembly presented here to the realization even of simple circuits requires more than mere optimization of the tools developed to date. As argued above, complex self-assembly will require a hierarchical, modular approach and, hence, the development of molecular switches that test for electronic functionality and feed back on the bioassembly process. Such switches will involve a functional interface between molecular biology and electronics, namely the ability of biomolecules to read electronic signals presented to them by the assembled devices and circuits, and then to effect the assembly process based on those findings. Only then can a full merging of biology and electronics be achieved.

Acknowledgments

The concepts and tools described in this chapter have been developed over the past decade by a significant group of researchers at Technion – Israel Institute of Technology. The author is especially grateful to Erez Braun, Kinneret Keren, Yoram Reiter, Arbel Artzi, Stav Zeitzev, Ilya Baskin, and Doron Lipson, whose contributions were immeasurable. Different areas of the research were funded by the Israeli Science Foundation, Bikura, the fifth EU program, the German Israeli DIP, the Rosenbloom family, and the Russell Berrie Nanotechnology Institute.

References

1 Braun, E., Eichen, Y., Sivan, U. and Ben Yoseph, G. (1998) DNA templated assembly and electrode attachment of conducting silver wire. *Nature*, **391**, 775–778.

2 Endres, R.G., Cox, D.L. and Singh, R.R.P. (2004) The quest for high-conductance DNA. *Reviews of Modern Physics*, **76** 195. and references therein.

3 Legrand, O., Côte, D. and Bockelmann, U. (2006) Single molecule study of DNA conductivity in aqueous environment. *Physical Review*, **E73**, 031925. and references therein.

4 Heath, J.R., Kuekes, P.J., Snider, G.S. and Stanley Williams, R. (1998) A defect-tolerant computer architecture: opportunities for nanotechnology. *Science*, **280**, 1716.

5 Eichen, Y., Braun, E., Sivan, U. and Ben Yoseph, G. (1998) Self assembly of nanoelectronics components and circuits using biological templates. *Acta Polymerica*, **49**, 663–670.

6 Keren, K., Krueger, M., Gilad, R., Ben-Yoseph, G., Sivan, U. and Braun, E. (2002) Sequence-specific molecular lithography on single DNA molecules. *Science*, **297**, 72.

7 Keren, K., Berman, R.S. and Braun, E. (2004) Patterned DNA Metallization by sequence-specific localization of a reducing agent. *Nano Letters*, **4** (2), 323–326.

8 Keren, K., Berman, R., Sivan, U. and Braun, E. (2003) DNA-templated carbon-nanotube field-effect transistor. *Science*, **302**, 1380–1382.

9 Artzy-Schnirman, A., Zahavi, E., Yeger, H., Rosenfeld, R., Benhar, I., Reiter, Y. and Sivan, U. (2006) Antibody molecules discriminate between crystalline facets of gallium arsenide semiconductor. *Nano Letters*, **6**, 1870.

10 Brod, E., Nimri, S., Turner, B. and Uri, Sivan (2008) Electrical control over antibody-antigen binding, *Sensors and Actuators B: Chemical*, **128**, 560.

11 Baskin, I., Zaitsev, S., Lipson, D., Gilad, R., Keren, K., Ben-Yoseph, G. and Sivan, U. (2006) A molecular shift register and its utilization for an autonomous DNA synthesis. *Physical Review Letters*, **97**, 208103.

12 Holgate, C.S. *et al.* (1983) Immunogold-silver staining: new method of immunostaining with enhanced sensitivity. *Journal of Histochemistry & Cytochemistry*, **31**, 938–944.

13 Birrell, G.B. *et al.* (1986) Silver-enhanced colloidal gold as a cell surface marker for photoelectron microscopy. *Journal of Histochemistry & Cytochemistry*, **34**, 339–345.

14 Keren, K. (2004) PhD thesis, *Self-assembly of molecular-scale electronics by genetic recombination*, Technion, Haifa Israel.

15 Braun, E. and Sivan, U. (2004) DNA templated electronics, in *Nano-Biotechnology, Concepts, Applications and Perspectives* (eds C.M. Nimeyer and C.A. Mirkin), Wiley-VCH, Weinheim, pp. 244–253.

16 Keren, K., Sivan, U. and Braun, E. (2004) DNA Templated electronics, in *Bio-electronics: From Theory to Applications* (eds I. Willner and E. Katz), Wiley-VCH, Weinheim, pp. 265–284.

17 Braun, E. and Keren, K. (2004) From, DNA to transistors. *Advances in Physics*, **53**, 441–496.

18 Gazit, E. (2007) Use of biomolecular templates for the fabrication of metal nanowires. *FEBS*, **274**, 317–322.

19 Gu, Q. *et al.* (2006) DNA nanowire fabrication. *Nanotechnology*, **17**, R14–R25.

20 Richter, J. (2003) Metallization of DNA. *Physica*, **E16**, 157–173.

21 Cox, M.M. (2000) *Progress in Nucleic Acids Research and Molecular Biology*, **63**, 311–366.

22 Hseih, P., Camerini-Otero, C.S. and Comerini-Otero, D. (1992) The synapsis

event in the homologous pairing of DNAs: RecA recognizes and pairs less than one helical repeat of DNA. *Proceedings of the National Academy of Sciences of the United States of America*, **89**, 6492–6496.

23 Dekker, C. (1999) Carbon nanotubes as molecular quantum wires. *Physics Today*, May, **52**, 22–28.

24 Liu, J. *et al*. (1998) Fullerene pipes. *Science*, **280**, 1253–1256.

25 Balavoine, F. *et al*. (1999) Helical crystallization of proteins on carbon nanotubes: a first step towards the development of new biosensors. *Angewandte Chemie-International Edition*, **38**, 1912–1915.

26 Shim, M., Kam, N.W.S., Chen, R.J., Li, Y. and Dai, H. (2002) Functionalization of carbon nanotubes for biocompatibility and biomolecular recognition. *Nano Letters*, **2** (4), 285–288.

27 Gomez-Navarro, C. *et al*. (2002) Contactless experiments on individual DNA molecules show no evidence for molecular wire behavior. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 8484–8487.

28 Bockrath, M. *et al*. (2002) Scanned conductance microscopy of carbon nanotubes and λ-DNA. *Nano Letters*, **2**, 187–190.

29 Avouris, P. (2002) Molecular electronics with carbon nanotubes. *Accounts of Chemical Research*, **35**, 1026.

30 Perl-Treves, D., Kessler, N., Izhaky, D. and Addadi, L. (1996) Monoclonal antibody recognition of cholesterol monohydrate crystal faces. *Chemistry & Biology*, **3**, 567–577.

31 Bromberg, R., Kessler, N. and Addadi, L. (1998) Antibody recognition of specific crystal faces; 1,4-dinitrobenzene. *Journal of Crystal Growth*, **193**, 656–664.

32 Braden, B.C. *et al*. (2000) X-ray crystal structure of an anti-Buckminsterfullerene antibody Fab fragment: Biomolecular recognition of C60. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 12193–12197.

33 Skerra, A. and Pluckthun, A. (1988) Assembly of a functional immunoglobulin Fv fragment in *Escherichia coli*. *Science*, **240**, 1038–1041.

34 Seeman, N.C. (2003) DNA in a material world. *Nature*, **421**, 427–431.

35 Mirkin, C.A., Letsinger, R.L., Mucic, R.C. and Storhoff, J.J. (1996) A DNA-based method for rationally assembling nanoparticles into macroscopic materials. *Nature*, **382**, 607–609.

36 Brown, S. (1997) Metal recognition by repeating polypeptides. *Nature Biotechnology*, **15**, 269–272.

37 Brown, S., Sarikaya, M. and Johnson, E. (2000) Genetic analysis of crystal growth. *Journal of Molecular Biology*, **299**, 725–732.

38 Naik, R.R., Stringer, S.J., Agarwal, G., Jones, S.E. and Stone, M.O. (2002) Biomimetic synthesis and patterning of silver nanoparticles. *Nature Mater*, **1**, 169–172.

39 Naik, R.R., Brott, L.L., Clarson, S.J. and Stone, M.O. (2002) Silica precipitating peptides isolated from a combinatorial phage display libraries. *Journal of Nanoscience and Nanotechnology*, **2**, 1–6.

40 Kjaergaard, K., Sorensen, J.K., Schembri, M.A. and Klemm, P. (2000) Sequestration of zinc oxide by fimbrial designer chelators. *Applied and Environmental Microbiology*, **66**, 10–14.

41 Brown, S. (1992) Engineering iron oxide adhesion mutants of *Escherichia coli* phage receptor. *Proceedings of the National Academy of Sciences of the United States of America*, **89**, 8651–8655.

42 Gaskin, D.J.H., Starck, K. and Wulfson, E.N. (2000) Identification of inorganic crystal-specific sequences using phage display combinatorial library of short peptides: a feasibility study. *Biotechnology Letters*, **22**, 1211–1216.

43 Wang, S. *et al*. (2003) Peptides with selective affinity for carbon nanotubes. *Nature Mater*, **2**, 196–200.

44 Willett, R.L., Baldwin, K.W., West, K.W. and Pfeiffer, L.N. (2005) Differential

adhesion of amino acids to inorganic surfaces. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 7817–7822.

45 Lee, S.W., Mao, C., Flynn, C.E. and Belcher, A.M. (2002) Ordering quantum dots using genetically engineered viruses. *Science*, **296**, 892–895.

46 Whaley, S.R., English, D.S., Hu, E.L., Barbara, P.F. and Belcher, A.M. (2000) Selection of peptides with semiconducting binding specificity for directed nanocrystal assembly. *Nature*, **405**, 665–668.

47 Goede, K., Busch, P. and Grundmann, M. (2004) Binding specificity of a peptide on semiconductor surfaces. *Nano Letters*, **4**, 2115–2120.

48 Knight, C.A., Cheng, C.C. and DeVries, A.L. (1991) Adsorption of α-helical antifreeze peptides on specific ice crystal surface planes. *Biophysical Journal*, **50**, 409.

49 DeOliviera, D.B. and Laursen, R.A. (1997) Control of calcite crystal morphology by a peptide designed to bind a specific surface. *Journal of the American Chemical Society*, **119**, 10627.

50 Azriel-Rosenfeld, R., Valensi, M. and Benhar, I. (2003) A human synthetic combinatorial library of arrayable single-chain antibodies based on shuffling in vivo formed CDRs into general framework regions. *Journal of Molecular Biology*, **335**, 177–192.

51 Smith, G.P. (1985) Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science*, **228**, 1315–1317.

52 Stemmer, W.P. *et al*. (1995) Single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribonucleotides. *Gene*, **164**, 49–53.

53 Adleman, L.M. (1994) Molecular computation of solutions to combinatorial problems. *Science*, **266**, 1021–1024.

54 Hagiya, M. *et al*. (1997) Towards parallel evaluation and learning of Boolean μ-formulas with molecules, in *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, DNA Based Computers III, Volume **48**, pp. 57–72.

55 Sakamoto, K. *et al*. (2000) Molecular computation by DNA hairpin formation. *Science*, **288**, 1223–1226.

56 Khodor, J. and Gifford, D.K. (2002) Programmed mutagenesis is universal. *Theory of Computing Systems*, **35**, 483–499.

57 Golomb, Solomon W. (1982) *Shift Register Sequences*, Aegean Park Press.

58 Although linear shift registers are discussed here, the automaton should work equally well with non-linear feedback functions.

59 The zero string should be avoided as it maps onto itself by any linear feedback function.

60 A 5-bit degenerate rule is constructed by adding the two preceding bits of the sequence to the 3′ end of the corresponding 3-bit rule.

61 Sugimoto, N., Nakano, S., Yoneyama, M. and Honda, K. (1996) Improved thermodynamic parameters and helix initiation factor to predict stability of DNA duplexes. *Nucleic Acids Research*, **24**, 4501–4505.

62 For brevity, we use notation that mixes bases with bits. Each bit represents three bases.

63 Annealing at 54 °C for 30 s, extension at 72 °C for 1 min, melting at 95 °C for 30 s.

10
Formation of Nanostructures by Self-Assembly

Melanie Homberger, Silvia Karthäuser, Ulrich Simon, and Bert Voigtländer

10.1
Introduction

The increasing demand for high-density electronic devices has triggered – and continues to trigger – the development of new nanofabrication methods. Two conceptually different strategies are applied for the fabrication of nanostructures, namely: (i) the *top-down* strategy; and (ii) the *bottom-up* strategy.

The *top-down* approaches utilize lithographical methods to fabricate nanostructures starting from the bulk materials (see Chapters 5, 6, and 7), whereas in the *bottom-up* approaches nanostructures are built up from atoms, molecules, or nanoscale sub-units. The *top-down* methods enable the generation of a large variety of defined structures, but these are limited by the resolution of current lithography techniques. The *bottom-up* methods offer the opportunity to fabricate structures even in the single-digit nanometer range, but they suffer from the fact that it is still a great challenge to direct the functional sub-units into desired structures. One extreme approach in this context is the utilization of a scanning probe microscope for building up nanostructures atom by atom at low temperatures (see Chapter 9). However, although this approach is ultimate in terms of the size of the nanostructures, it is a very slow and sophisticated method. Compared to this method, processes based on self-organization or self-assembly have the key advantage that they enable the formation of billions of nanostructures with control over size, shape, and composition in a fast and parallel fashion. Due to entropic effects during the formation of nanostructures by self-assembly, defects are expected always to be present, and fault-tolerant architectures are required to cope with this problem. The combination of the self-assembly of atoms, molecules and nanoscale subunits could lead to well-ordered functional nanostructures. For example, inorganic nanostructures, generated by the self-assembly of atoms via epitaxial growth, may serve as templates for the selective adsorption of functional molecules, which themselves display "anchor-points" at which size-selected clusters could be attached, altogether leading to highly ordered functional nanostructures with applications in molecular electronics. One critical

Nanotechnology. Volume 3: Information Technology I. Edited by Rainer Waser

ISBN: 978-3-527-31738-7

factor determining the benefits of this approach for electronic systems will be the surface-selective SAM formation – that is, the selective assembly of functional molecules on special device patterns forming an ordered array. In this context, in the following chapter attention is focused on the formation of nanostructures by self-assembly via epitaxial growth, the self-assembly of molecules, and the formation and self-assembly of nanoscale subunits. Basic physical principles and selected examples will be presented.

10.2 Self-Assembly by Epitaxial Growth

One approach for the fabrication of nanostructures is *epitaxial growth*. Such growth usually occurs under kinetic conditions, so that the sizes can be tuned down to the single-digit nanometer range by choosing appropriate growth conditions. However, size uniformity is the greatest challenge here. If the growth is taking place under (near) equilibrium conditions, then the size distribution of the nanostructures may be narrow, but is provided by the material system and cannot be varied easily. The formation of islands, wires and rods will be presented as examples of nanostructures grown by epitaxy. Subsequently, the growth of nanostructures on template substrates structured by step arrays or underlying dislocation networks will be considered. The combination of self-organized growth with lithography ("hybrid methods") allows the self-assembled nanostructures to be aligned relative to predefined patterns. It is possible that such inorganic nanostructured templates may be used in the future for the selective formation of molecular layers.

10.2.1 Physical Principles of Self-Organized Epitaxial Growth

10.2.1.1 Epitaxial Growth Techniques

The main methods used for semiconductor epitaxial growth are chemical vapor deposition (CVD) [1] and molecular beam epitaxy (MBE) [2, 3]. In CVD, growth gases containing compounds of the elements to be deposited are introduced into the growth chamber. When the gas molecules hit the substrate surface, they decompose (partially) and the gaseous products desorb from the surface. Different chemical reactions taking place at the surface, or even in the gas phase, lead to a quite complex nature of the fundamental processes of epitaxial growth in CVD. Molecular beam epitaxy is conceptually simpler; here, the elements to be deposited are heated in evaporators until they evaporate, whereupon the beam of the atoms hits the surface and the atoms diffuse over the surface and finally bind at surface lattice sites (Figure 10.1).

In spite of the fact that the MBE growth is, in principle, much easier than the CVD growth, there are still many different fundamental processes occurring during epitaxial growth by MBE [4]. Part of these are illustrated schematically in Figure 10.1. Atoms from the molecular beam arrive at the surface of the crystalline substrate (a)

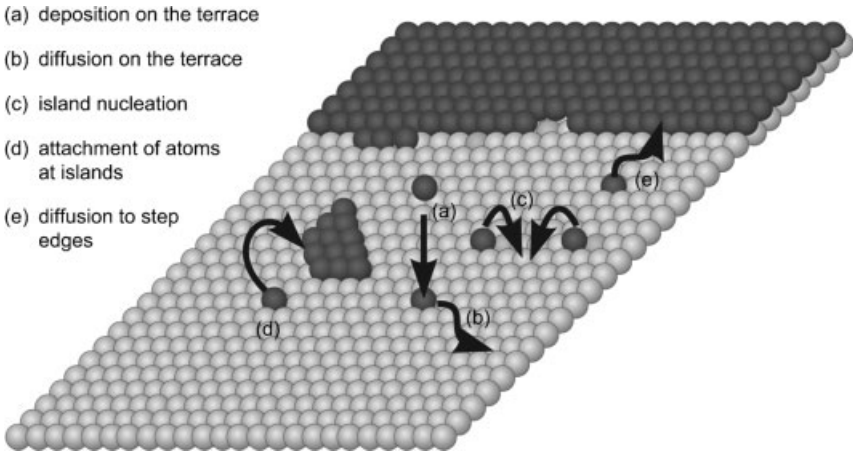


Figure 10.1 Scheme of different fundamental processes occurring during epitaxial growth, leading to a self-organization of two-dimensional islands.

and may diffuse over the surface when the activation energy for diffusion is overcome (b). When two atoms (or sometimes also more than two atoms) meet, they form a nucleus for a stable island (c). Such a nucleus may grow to a stable two-dimensional (2-D) island by attachment of further diffusing adatoms (d). The nucleus for which the probabilities to grow or decay are equal is called the *critical nucleus* [5]. Nuclei which are larger than the critical nucleus are termed stable 2-D islands, while nuclei smaller than the critical nucleus are called subcritical nuclei or *embryos*. Another process is the diffusion and attachment at pre-existing steps if the diffusion length is sufficient (e).

10.2.1.2 Kinetically Limited Growth in Homoepitaxy

In kinetically limited growth the system is governed by energetic barriers such as barriers for the diffusion of adatoms and barriers for incorporation of atoms into the crystal, and additionally by outer conditions such as the growth rate. The 2-D islands (which are one atomic layer high) represent the simplest example of the self-assembled growth of nanostructures. In the following section it will be shown how the density and size of these islands can be controlled by the kinetic parameters temperature and growth rate. First, the deposition temperature influences the island density strongly, as shown by the comparison of Figure 10.2a and b. The island density as function of temperature follows an Arrhenius law: $n \sim \exp(E_{\mathrm{act}}/kT)$, where E_{act} is an effective activation energy consisting of a diffusion energy and binding energy component, having values around 1 eV in the case of semiconductors [5]. The temperature is one important parameter of growth kinetics, and the deposition rate is another. It has been found that the island density (n) scales with the deposition rate (F) in the form of a power law $n \sim F^{\alpha}$, with a scaling exponent α. Combining the temperature and the rate dependence results in the following scaling law: $n \sim F^{\alpha}$

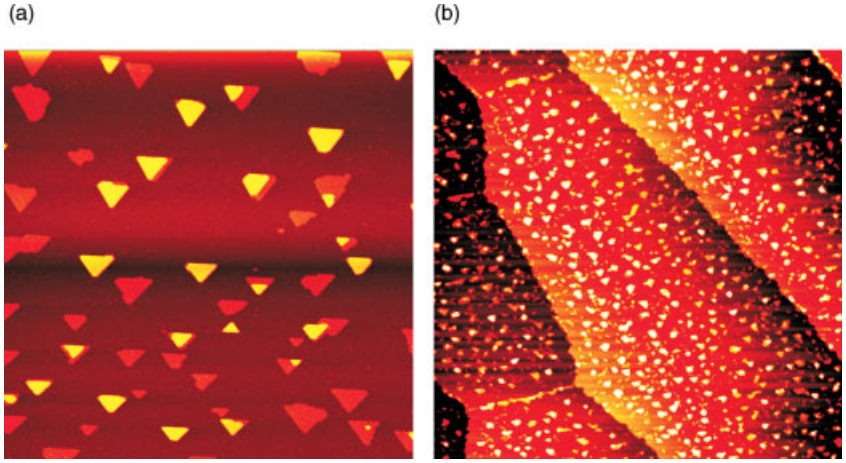


Figure 10.2 Scanning tunneling microscope images after the growth of 0.2 atomic layers of silicon on a Si(1 1 1) surface. The islands have triangular shape due to the symmetry of the substrate, and have a height of one atomic layer (orange) or two atomic layers (yellow). The island density depends on the temperature, as can be seen by comparison of growth at high temperatures of 770 K (a) to growth at a lower temperature 610 K (b). Both images have a size of 350 nm.

$\exp(E_{\mathrm{act}}/kT)$ [5], which shows that the island density can be controlled over a wide range by adjusting the kinetic growth parameters of temperature and growth rate. The average island distance is simply the square root of the inverse of the island density $L = 1/\sqrt{n}$.

Although the nucleation of the islands is a random process, the distribution of the island sizes is centered around a mean value (Figure 10.3). This arises due to a saturation of the island nucleation, as will be explained in the following. During the early stage of growth (nucleation regime), the islands nucleate randomly on the surface and the distance between them decreases. If the distance between the islands is equal to the mean distance that an adatom travels before a nucleation event happens, then the incorporation of adatoms in existing islands becomes a more probable event than the nucleation of new islands; hence, a "capture zone" forms around each island. Adatoms deposited in this capture zone attach to the corresponding island. Without this effect the distribution of island sizes would be even broader. The nucleation of further islands is suppressed beyond a certain coverage (growth regime), and the average island size can be controlled by the deposited amount. The island size distributions for two different temperatures are shown in Figure 10.3, where it can be seen that the peak in the island size distribution scales towards larger sizes with higher temperatures. For very small islands, the surface reconstruction can also modify the island size distribution [4]. In the kinetic growth regime the island density of 2-D islands can be controlled by the kinetic parameters temperature and deposition rate, while the size distribution is quite broad due to the stochastic nature of the nucleation of the islands.

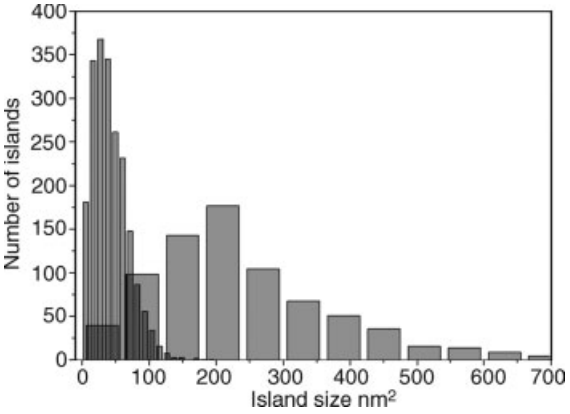


Figure 10.3 Island size distribution for two-dimensional Si islands on Si(1 1 1). The width of the distribution is of the order of the average size of the islands. Two distributions for two different temperatures are displayed. The narrow bins (peak at small island sizes) correspond to deposition at 610 K. The distribution with the wide bins peak at larger island sizes) corresponds to deposition at 710 K.

10.2.1.3 **Thermodynamically Stable Nanostructures**

If nanosized islands were to be thermodynamically stable, their size distribution could be narrow. A thermodynamically stable island size means that the energy (per atom) has a minimum for this stable size. For configurations with larger or smaller islands, the energy (per atom) would be higher, and therefore it only necessary to approach thermodynamic equilibrium in order to obtain a very narrow island size distribution. One way to achieve thermodynamic equilibrium is to heat a sample with different island sizes present and then to wait until equilibrium has established. The equilibrium configuration will be established by material transport between the islands, as the atoms will detach from islands with higher energy and attach to islands with a lower energy (per atom). However, as will be shown below, in the simplest case (considering only a surface or edge energy term) the thermodynamically stable island size is infinitely large. This behavior is not of any use for the formation of nanostructures with a narrow size distribution, and corresponds to the well-known Ostwald ripening. Only if additional terms in the energy are important (e.g., strain energy) will the energy per particle show a minimum for a finite particle size, while a narrow size distribution can be expected under equilibrium conditions.

In order to describe material transport in a system with a variable number of atoms, the chemical potential is used; this is the change of the energy (of an island) when the number of particles changes $\mu = \mathrm{d}E/\mathrm{d}N$. During the equilibration process atoms detach from islands where the chemical potential is highest, and attach to islands with a lower chemical potential. This lowers the total energy of the system, and consequently the material transport between different islands is governed by the chemical potential. A simple example is the chemical potential of quadratic 2-D islands of dimension L (Figure 10.4a). The energy difference between different-sized islands

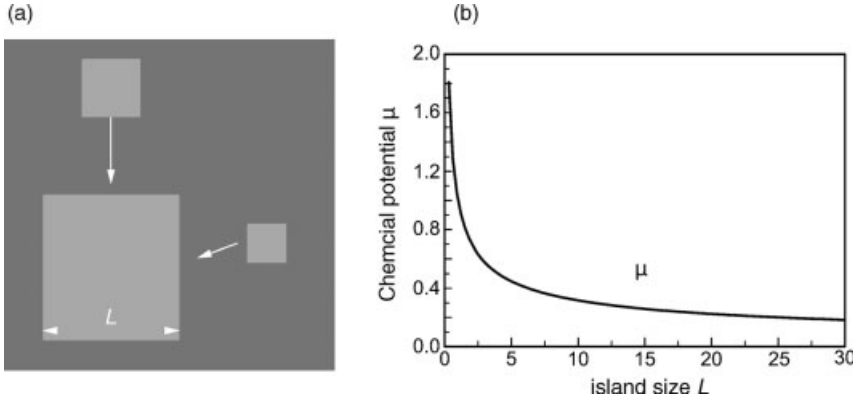


Figure 10.4 (a) Coarsening of a large island at the expense of small islands. (b) The chemical potential of an island.

comes from the edge energy (β is the edge energy per length). The energy of an island is $E = E_{\text{edge}} = 4\,L\beta$. The number of atoms in an island (N) depends on the dimension L, as $N = L^2/\omega$, with ω being the area per atom. The chemical potential is then

$$\mu = \frac{dE}{dN} = \frac{2\omega\beta}{L} \sim \frac{1}{L} \tag{10.1}$$

Since μ is decreasing for larger islands, infinite size islands have the lowest chemical potential (Figure 10.4b), which means that the stable island is infinitely large. In this case, the equilibration does not result in a stable finite island size; equilibration in this model by material transport between islands is also referred to as *coarsening* because it results in the shrinkage of small islands and a growth (coarsening) of large islands (Ostwald ripening).

An infinitely large stable island size is the result for homoepitaxial growth, taking into account only the edge energy. However, the situation becomes different when elastic stress is also taken into account, as it occurs in heteroepitaxy where two different materials grow onto each other. Here, stress is induced by the different lattice constants of the substrate material and the material of the islands. The elastic effect of strained 2-D islands can be approximated by that of a surface-stress domain – that is, the surface stress at the area of the island is different from that at the rest of the surface (Figure 10.5). The strain energy of a quadratic surface-stress domain can be calculated using the elastic theory as $E_{\text{strain}} = 2LC\ln L$ [6]. Adding the step edge energy results in a total energy of a strained island:

$$E = E_{\text{edge}} + E_{\text{strain}} = 2L[2\beta - C'\ln L] \tag{10.2}$$

This results in the following chemical potential:

$$\mu = \omega\left[\frac{2\beta - C'}{L} - \frac{C'}{L}\ln L\right] \tag{10.3}$$

which is illustrated in Figure 10.5. In this case, the chemical potential has a minimum at the size $L_{\min} = \exp(2\beta/C')$, which would mean that during coarsening the islands

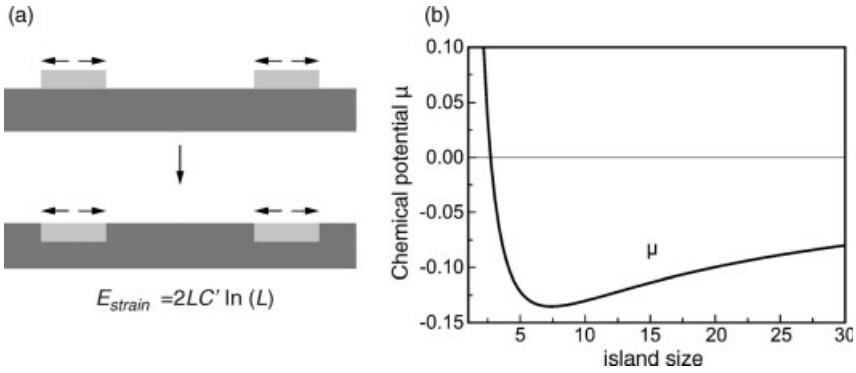


Figure 10.5 (a) The elastic stress induced by two-dimensional islands with a different lattice constant than the substrate can be approximated by surface stress domains. (b) Chemical potential of an island with an energy component due to elastic strain included.

would approach this size. Larger islands would dissolve and smaller islands grow until all islands have the size $L_{\min}$, that is, the lowest chemical potential, and this would result in a very narrow size distribution. Unfortunately, step energies are only very poorly known, so that it is not possible to predict a reliable number for the equilibrium island size. An experimental realization of thermodynamically stable islands has not yet been confirmed, apart from surface reconstructions with a relatively large unit cell.

If the formation of nanostructures in equilibrium is compared to the formation of nanostructures by growth kinetics, the following advantages and disadvantages occur. Nanostructures grown under equilibrium conditions have (under specific conditions) the advantage of a narrow size distribution around the optimum size. However, a disadvantage is that the size is determined by the material parameters (strain energy and step edge energy for instance), and cannot be tuned freely. The size and density of nanostructures formed under kinetic conditions can be tuned easily by variations of the growth parameters such as growth rate and temperature. On the other hand, the size uniformity of the islands grown under kinetic conditions is relatively poor.

10.2.1.4 **Nanostructure Formation in Heteroepitaxial Growth**

Semiconductor nanostructures can be fabricated by self-organization using heteroepitaxial growth, which is the growth of a material B on a substrate of different material A. In heteroepitaxial growth, the lattice constants of the two materials are often different. The lattice mismatch for the two most commonly used material systems, Si/Ge and GaAs/InAs, is 4.2% and 7%, respectively (shown schematically in Figure 10.6a). This lattice mismatch leads to a build-up of elastic stress in the initial 2-D growth in heteroepitaxy. In the case of Ge heteroepitaxy on Si, the Ge is confined to the smaller lattice constant of the Si substrate – that is, the Ge is strained to the Si lattice constant (Figure 10.6b). One way to relax this stress is via the formation of three-dimensional (3-D) Ge islands, in which only the bottom of the islands is

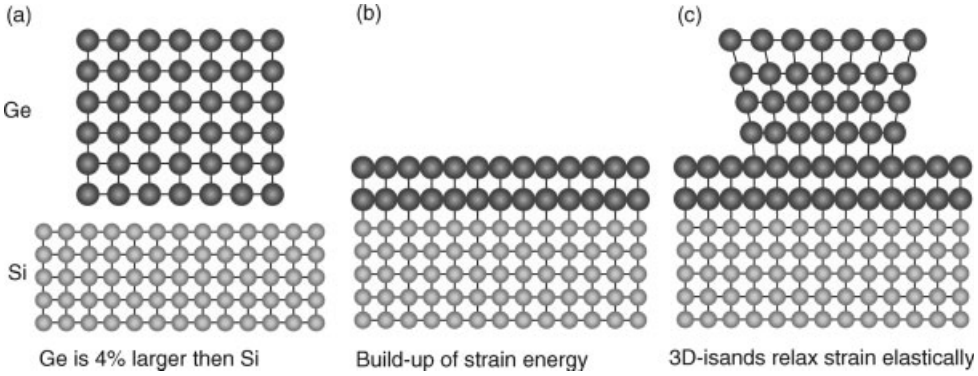


Figure 10.6 (a) Schematic representation of Si and Ge crystals with different lattice constants. (b) Build-up of elastic strain energy during 2-D growth with Ge confined to the Si lattice constant, and (c) elastic relaxation by the formation of 3-D islands (Stranski–Krastanov growth). In the upper part of the 3-D island the lattice constant relaxes towards the Ge bulk constant. The usual form of the 3-D islands is a pyramid, and not like that shown in this schematic.

confined to the substrate lattice constant. In the upper part of the 3-D island the lattice constant can relax to the Ge bulk lattice constant and reduce the stress energy in this way (Figure 10.6c). This growth mode, which is characterized by the formation of a 2-D wetting layer and the subsequent growth of (partially relaxed) 3-D islands, is referred to as the Stranski–Krastanov growth mode, some examples are which are described shown in Section 10.2.2.

The driving force for the formation of self-organized nanoislands in heteroepitaxial growth is the build-up of elastic strain energy in the stressed 2-D layer. As a reaction to this, a partial stress relaxation by the formation of 3-D islands can lower the free energy of the system. The process of island formation close to equilibrium is a trade-off between elastic relaxation by the formation of 3-D islands, which lowers the energy of the system, and an increase of the surface area, which increases the energy.

In a simple model, where the islands are cubes with the length x, the additional surface energy for a film in an island morphology (compared to a strained film) is proportional to the island length squared (x^2). The gained elastic relaxation energy compared to that of a flat film is, in the simplest assumption, proportional to the volume of the island (x^3). For the same total volume in the film, the energy difference between the 3-D island morphology and the flat morphology is

$$\Delta E = E_{\text{surf}} - E_{\text{relax}} = C\gamma x^2 - C'\varepsilon^2 x^3 \tag{10.4}$$

where γ is the surface energy, ε is the lattice mismatch, and C and C' are constants. The contributions of E_{surf}, E_{relax} and the total energy difference between the 3-D island morphology and a flat film are shown in Figure 10.7, as a function of the island size x. For small sizes of the 3-D islands, the 3-D island morphology is unfavorable up until the point where the absolute value of the gained elastic relaxation energy ($\sim x^3$) becomes larger than the cost of the surface energy ($\sim x^2$). For islands larger than a

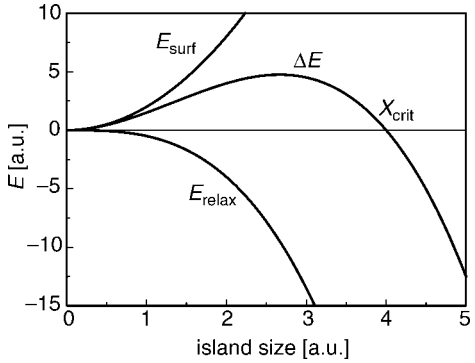


Figure 10.7 Energy difference between a film of flat 2-D morphology and a film morphology consisting of 3-D islands. The total energy difference and the contributions surface energy difference and relaxation energy are plotted.

critical island size, x_{crit}, the formation of 3-D islands is energetically preferred over the 2-D film morphology. While this simple model shows the basic driving forces for the 2-D to 3-D transition, it contains several simplifications. For example, in this simple model the island morphology is assumed as being cuboid, which does not correspond to the experimentally observed island shapes. Further, the simple model contains only energetic considerations of two final states. Kinetic effects, such as the required material transport necessary during the 2-D and 3-D transition are not considered.

Apart from the formation of 3-D islands, there is another process which can partially relax the stress of a strained 2-D layer, namely the introduction of misfit dislocations. This corresponds to the removal of one lattice plane of a compressively strained 2-D layer. If a lattice plane is removed in regular distances in the 2-D layer, then a misfit dislocation network forms. Depending on the growth parameters of temperature and growth rate, the self-organized growth can either be close to equilibrium or in the kinetically limited regime. At close to equilibrium (i.e., at high growth temperatures or low deposition rates), the occurring morphology (strained layer, 3-D islands, or a film with dislocations) is determined only by the energies of the particular configurations, and the morphology with the lowest energy will be formed. If the growth is kinetically limited, then the activation barriers are important. For instance, an initially flat strained layer can transform to a morphology with 3-D islands or to a film with dislocations. Yet, what actually happens depends on the kinetics of the growth process – that is, on the activation energy for the formation of 3-D islands compared to the activation energy for the introduction of misfit dislocations.

10.2.2 Semiconductor Nanoislands and Nanowires

10.2.2.1 Stranski–Krastanov Growth of Nanoislands

Stranski–Krastanov growth occurs, for example, in InGaAs/GaAs growth [7]. An example of InAs nanoislands grown on a GaAs substrate is shown in the transmission

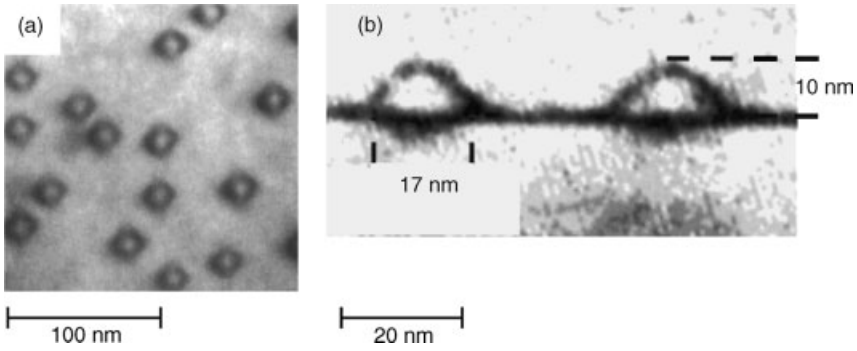


Figure 10.8 InAs nanoislands grown on a GaAs surface. (a) As imaged by plan-view transmission electron microscopy (TEM); (b) cross-sectional view with TEM [7].

electron microscopy (TEM) image in Figure 10.8. The GaAs islands were grown by MBE at a growth temperature of 775 K, and the density of the islands was 4.5×10^{10} cm^{-2}, with an average lateral size of 17.5 ± 0.5 nm. The challenges in the growth ofthese semiconductor islands are to grow islands of desired size and density, and with a high size uniformity. As in the case of the 2-D islands, a higher growth temperature generally leads to the formation of larger islands, while a higher growth rate leads to the formation of smaller islands. The size of the islands increases with coverage; often, the density of the islands saturates during an early stage of the growth. These are general trends which may depend on the material system and the particular deposition technique. In some cases (self-limiting growth), the size of the islands saturates while the density increases with coverage, and this type of growth mode leads to a high size uniformity of the islands. The size uniformity achieved in self-assembled growth of semiconductor islands may be as small as a small percent. The confinement of charge carriers in all three directions gives rise to atomic-like energy levels. Quantum dot lasers operating at room temperature have now been realized [8]. The islands grown on a flat substrate are usually not ordered laterally due to the random nature of the nucleation process. In the following section, it will be shown how nucleation at specific sites can be achieved.

10.2.2.2 Lateral Positioning of Nanoislands by Growth on Templates

An example of ordered nucleation at a prestructured substrate is shown in Figure 10.9a [9], where Ge islands nucleate above dislocation lines. However, when a SiGe film is grown on a Si(0 0 1) substrate, dislocations form at the interface between the SiGe film and the substrate. The driving force for the formation of dislocations is the relief of elastic strain, which arises due to the different lattice constants between the Si substrate and a Ge/Si film on this substrate. During annealing, the dislocations form a relatively regular network, due to a repulsive elastic interaction between the dislocations. The preferred nucleation of Ge islands above the dislocation lines (Figure 10.9a) can be explained by local stress relaxation above the dislocation lines providing a lattice constant closer to the Ge one. The nucleation does not occur

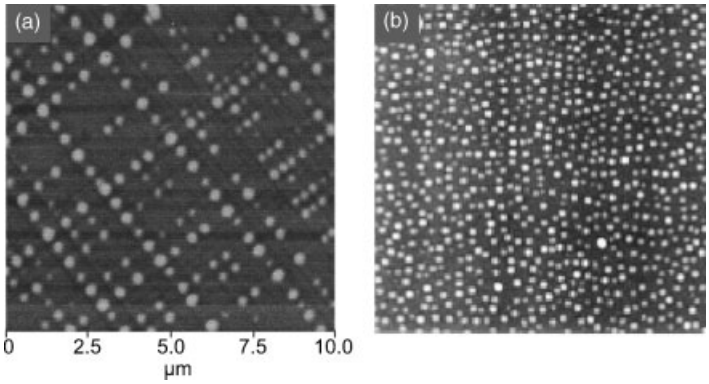


Figure 10.9 (a) Ordered nucleation of Ge islands on a template which is pre-structured by an underlying network of dislocations. (b) Germanium islands grown on a substrate without dislocations [9]. Image sizes 7 μm.

randomly at the surface, but rather occurs simultaneously at sites which have the same structure. This leads to a more narrow size distribution than that for the growth on unstructured Si(0 0 1) substrates (Figure 10.9b).

10.2.2.3 **Silicide Nanowires**

If the crystal structure of the deposited material is different from that of the substrate, then effects related to the anisotropic match of both crystal structures may appear. If the overlayer material has a crystal structure which is closely lattice-matched to the substrate along one major crystallographic axis, but has a significant lattice-mismatch along the perpendicular axis, this should allow unrestricted growth of the epitaxial crystal in the first direction but limit the width in the other direction. Such a strategy has been applied to grow silicide nanowires [10]. Here, the substrate is a Si(1 0 0) surface (Si has diamond crystal structure), and by deposition of Er and subsequent annealing, $ErSi_2$-oriented crystallites with a hexagonal AlB_2-type crystal structure were formed on the Si substrate. The [0 0 0 1] axis of the $ErSi_2$ was oriented along a $[\bar{1}\,1\,0]$ axis of the Si(0 0 1) substrate, and the $[1\,1\,\bar{2}\,0]$ of the $ErSi_2$ was oriented along the perpendicular [1 1 0] axis, with lattice mismatches of $+6.5\%$ and -1.3%, respectively; this almost satisfies the proposed growth conditions for nanowires. $ErSi_2$ nanowires grown on the Si(1 0 0) surface are shown in Figure 10.10. The $ErSi_2$ nanowires align along one of the two perpendicular $<1\,1\,0>$ Si directions, which are the small mismatch directions. In these directions the crystal can grow without much build-up of stress, while the width of the $ErSi_2$ nanowire is $\sim$4 nm, the height $\sim$0.8 nm, and the length is several hundred nanometers. Such self-assembled arrays of nanowires may also be used as conductors for defect-tolerant nanocircuits, or as a template for further nanofabrication.

10.2.2.4 **Monolayer-Thick Wires at Step Edges**

Monolayer-high surface steps can be used to fabricate Ge nanowires using step-flow growth. Here, pre-existing step edges on the Si(1 1 1) surface are used as templates for

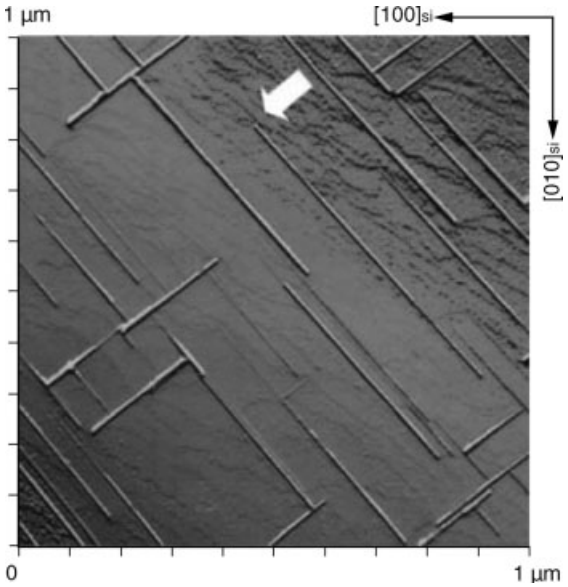


Figure 10.10 Scanning tunneling microscopy (STM) topograph showing $ErSi_2$ nanowires grown on a flat Si(0 0 1) substrate. The long direction of the nanowires is the one with the low lattice mismatch (1.3%), while the lattice mismatch in the perpendicular direction is 6.5%.

the growth of 2-D Ge wires at the step edges. When the diffusion of the deposited atoms is sufficient to reach the step edges, the deposited atoms are incorporated exclusively at the step edges, and the growth proceeds by a homogeneous advancement of the steps (step flow growth mode [4]. If small amounts of Ge are deposited, then the steps will advance only a few nanometers and narrow Ge wires can be grown.

One key issue for the controlled fabrication of nanostructures consisting of different materials is a method of characterization which can distinguish between the different materials on the nanoscale. If the surface is terminated with a monolayer of Bi, it is possible to distinguish between Si and Ge [11]. Figure 10.11a shows a scanning tunneling microscopy (STM) image after repeated alternating deposition of 0.15 atomic layers of Ge and Si, respectively. Due to the step-flow growth, the Ge and Si wires are formed at the advancing step edge. Whilst both elements can be easily distinguished by the apparent heights in the STM images, it transpired that the height measured by the STM was higher in areas consisting of Ge (red stripes) than in areas consisting of Si (yellow stripes). The apparent height of the Ge areas was $\sim$0.1 nm higher than that of the Si wires (Figure 10.11b), and the cross-section of a 3.3 nm-wide Ge nanowire was seen to contain only approximately 20 atoms (Figure 10.11c). The apparent height difference arises due to an atomic layer of Bi which is deposited initially and always floats on top of the growing layer. The different widths of the wires can easily be achieved by depositing different amounts of Ge and Si.

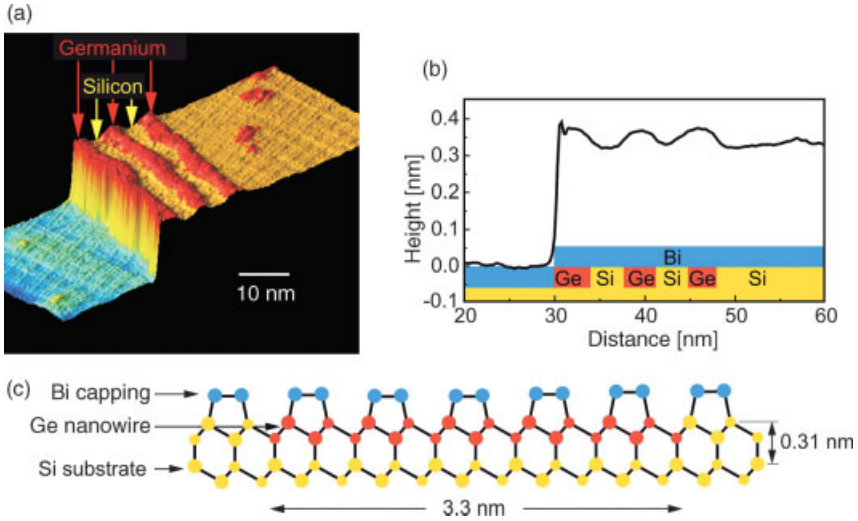


Figure 10.11 (a) STM image of 2-D Ge/Si nanowires grown by step-flow at a pre-existing step edge on a Si(1 1 1) substrate. The Si wires (yellow) and Ge wires (red) can be distinguished by different apparent heights. (b) A cross-section across the nanowires. (c) The atomic structure of a Ge wire on the Si substrate capped by Bi. The cross-section of the Ge wire contains only approximately 20 Ge atoms [11].

10.2.3
Hybrid Methods: The Combination of Lithography and Self-Organized Growth

In hybrid methods, self-organization is combined with lithographic patterning to form nanostructures on a smaller scale than are accessible by lithography. Most importantly, the hybrid methods provide a direct contact of nanostructures formed by self-organization to mesoscopic lithographically patterned structures. The self-organized growth of Ge islands in oxide holes is shown in Figure 10.12a–d. The starting surface is a silicon substrate with a thin oxide layer at the surface, and electron lithography is used to remove the oxide and form holes of a diameter of 0.5 μm where the bare Si surface is exposed [12]. The self-organized growth of Ge leads to the formation of Ge islands which may be smaller than the size scale of the electron beam lithography (EBL). The gas-phase growth of Ge is selective; that is, Ge will only grow on Si areas (inside the holes in the oxide), and not on the oxide itself. Figure 10.12 illustrates the nucleation of Ge islands in the holes in the oxide for different growth temperatures. At lower temperatures, the island density is so large that several islands nucleate in one oxide hole. However, if the temperature is increased, ultimately only one Ge island is able to nucleate in each oxide hole, and the size of the Ge island is smaller than the lithographically defined oxide hole. However, as seen in Figure 10.12d, the position of the Ge island inside the oxide hole is not defined but is rather randomly distributed within the oxide hole. Due to the fact that

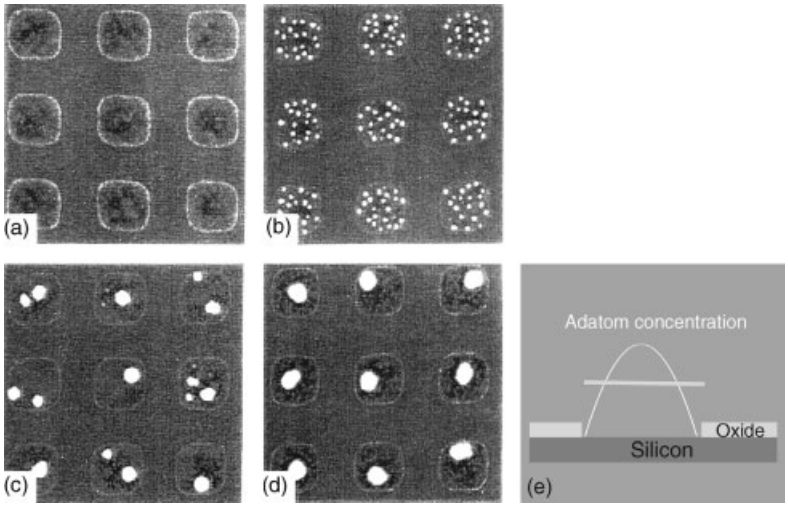


Figure 10.12 (a–d) Growth of Ge islands inside holes on an oxidized Si substrate [12]. (e) Adatom density in an oxide hole for those cases where the hole edges are sinks of adatoms (parabolic line), or for the case when the edges are not sinks for adatoms (horizontal line).

the Ge does not grow on the oxide, the edges of the oxide hole cannot serve as sinks for deposited Ge atoms, and therefore the Ge adatom concentration is homogeneous across the hole and nucleation of the Ge island is random within the oxide hole. If the edges of the hole were to serve as sinks for Ge atoms (e.g., if the edges of the hole were to consist of Si), then the adatom density would have a maximum at the center of the hole and the nucleation of Ge islands would occur preferentially at the center of the oxide holes (Figure 10.12e).

10.2.4
Inorganic Nanostructures as Templates for Molecular Layers

Several of the nanostructures discussed here can potentially be used as templates for the selective formation of molecular structures onto specific areas of the inorganic nanostructures generated by the self-assembly of atoms via epitaxial growth. The importance of the inorganic substrate for the formation of molecular layers, which is discussed in detail in the following section, is manifold. The role of the inorganic nanostructured template for the molecular self-assembly may be to steer the adsorption process kinetically, and to direct the molecules towards predefined adsorption sites. The first steps in this direction have been taken recently. Initially, special substrate surfaces were selected in accordance with their ability to adsorb molecules. For example, substrates with only weak adsorption properties are useful for molecular assemblies with weak intermolecular interactions, because such substrates allow for the necessary reorganization of molecules. In addition, it is

necessary that the interatomic distances at the substrate surface correspond to the dimensions of molecular structure elements. One example of such an substrate, which allows for a weak adsorption of polycyclic aromatic compounds is the Ag/Si(1 1 1)-$\sqrt{3}\times\sqrt{3}$R30° surface (Figure 10.13a). This surface is described by the honeycomb-chain-trimer model, in which each surface Si atom is bound to one Ag atom. This structure is derived from a Si(1 1 1) bulk termination by removing the top half of the first bilayer of Si atoms, forming trimers from the remaining Si atoms, and then adding a full monolayer of Ag atoms in positions slightly distorted from the regular triangular lattice (Figure 10.13a). An STM image of this structure is shown in Figure 10.13b. In this empty states image (+1.6 V sample bias), the bright protrusions correspond to the center of three Ag atoms (Ag trimers indicated by A in Figure 10.13b), and the minima in this image, indicated by B, correspond to the Si trimers). In the following, for simplicity, this surface is represented by a hexagonal network also indicated in Figure 10.13a. A supramolecular 2-D honeycomb network,

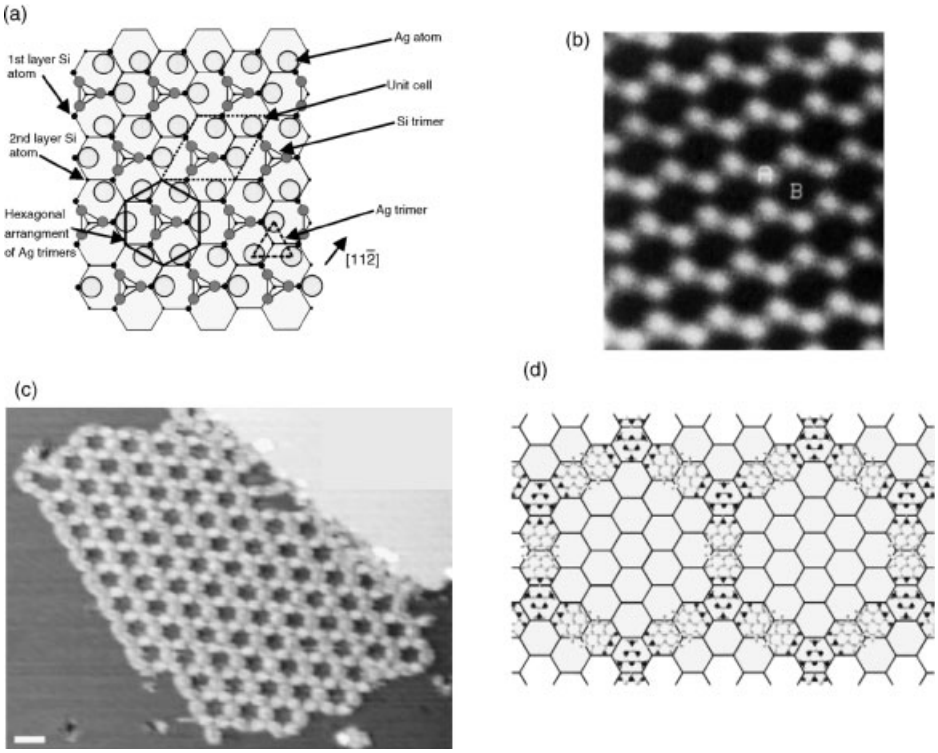


Figure 10.13 (a) Schematic showing the honeycomb-chain-trimer model for the Ag/Si (1 1 1)-$\sqrt{3}\times\sqrt{3}$R30° reconstruction [14]. STM image (empty states, +1.6 V sample bias) of the Ag/Si(1 1 1)-$\sqrt{3}\times\sqrt{3}$R30° substrate surface. The bright protrusions correspond to the center of three Ag atoms (image size 3 nm) [15]. (c) STM image of the hexagonal molecular network (see also Section 10.3.3) [13]. Scale bar = 3 nm. (d) Schematic diagram showing the registry of the molecular network with the underlying Ag/Si (1 1 1)-$\sqrt{3}\times\sqrt{3}$R30° surface reconstruction shown as hexagons.

with a larger periodicity (five times that of the Ag/Si(1 1 1)-$\sqrt{3} \times \sqrt{3}$R30° lattice constant; see Figure 10.13d) has been created by the assembly of two types of molecule on the Ag-terminated silicon surface [13]. This hexagonal molecular network is shown in Figure 10.13c, and is discussed in detail in Section 10.3.3. The registry of the molecules with respect to the underlying silver-terminated Si surface has been determined, and is shown schematically in Figure 10.13d. The calculated melamine–melamine separation has a near-commensurability with the surface lattice, showing the importance of the underlying inorganic template for the formation of the supramolecular structure.

In the future, the selective bonding of molecular species to inorganic template structures, which would enable site direction, will also represent a major challenge for the successful combination of inorganic templates and molecular structures.

10.3 Molecular Self-Assembly

Self-assembly is a bottom-up technique that uses the self-organization capabilities of molecular building blocks – that is, the ability to rearrange continuously until a complete ordered monolayer of molecules is formed – to assemble desired nanostructures. As a result of the self-assembly process, the molecular constituents form an ordered structure with a minimum global energy on well-defined, atomically flat surfaces. The term "molecular self-assembly" is reserved for the adsorption of molecular constituents onto surfaces and the spontaneous organization into regular arrangements. If only non-covalent interactions are used to direct the molecular constituents into the resulting surface pattern, these structures are termed "supramolecular" (supramolecular chemistry = the chemistry of the intermolecular non-covalent bond [16]). On the other hand, the term "self-assembled monolayer" (SAM) is reserved, according to Whitesides [17], for a 2-D film with the thickness of one molecule that is attached to a solid surface through covalent bonds.

The surface properties of metals, metal oxides or semiconductors can be changed in a desired way by the adsorption of SAMs onto these materials. Therefrom, a number of useful applications result, such as: (i) the modification of adhesion and wetting control [18]; (ii) an increase in corrosion resistance [19]; or (iii) the development of heterogeneous chiral catalysts [20]. By exploiting the chemical properties of the organic molecules used, additional functionalities can be created, and consequently the development of chemical sensors [21] and chemical force microscopy [22], the site-selective adsorption of nanoscale subunits (see also Section 10.4.2), or the fabrication of electronic devices [23, 24], is possible. Additionally, SAMs themselves are nanostructures with nanoscale dimensions useful in nanolithography [25]. Supramolecular surface patterns on the other hand can be used to create nanocavities, to provide well-defined reaction spaces, and they may also control host–guest chemistry or steer heterogeneous catalysis [26]. Further details of the present state of molecular self-assembly on planar substrates are provided in a series of reviews [20, 26–30].

10.3.1
Attaching Molecules to Surfaces

Bare surfaces of metals and metal oxides tend to adsorb organic materials because the adsorbates lower the free energy of the interface between the respective material and the ambient environment. The character of the chemical bond between the adsorbed molecules and the metal surface determines the interfacial electronic contact and the strength of the geometric fixation. Two main groups of links between molecules and solids can be distinguished: (i) covalent bonds, which result from the overlap of partially occupied orbitals of interacting atoms; and (ii) non-covalent bonds, which are based on the electrical properties of the interacting atoms or molecules.

Planar molecules with extended π-systems have been found to physisorb onto surfaces, such as highly oriented pyrolytic graphite (HOPG), Au(1 1 1), Cu(1 1 0), in a flat-lying geometry. This allows functional groups at the molecular periphery to approach each other easily and to build up intermolecular interactions, predominantly comprising hydrogen bonds and metal–ligand interactions. If the molecules are sufficiently mobile to diffuse on the surface, then the intermolecular interactions will guide the adsorbed molecules into 2-D supramolecular systems. Then, by adjusting the molecular backbone size and the position or number of the functional "recognition groups", complex supramolecular nanostructures can be designed [31].

Covalent bonds are established, if there is a significant overlap of the electron densities of the molecules and the metal, and this will result in a strong electronic and structural coupling. The spontaneous formation of SAMs on substrates through covalent bonds requires organic molecules with a chemical functionality or "headgroup" and a specific affinity for a selected substrate. There exists a number of headgroups, which bind to specific substrates forming directed covalent links. One frequently used covalent link is the bond between a thiol group on the molecular site and a noble metal substrate. Here, gold is favorable due to its proper non-oxidizing surface, although thiol or selenol bonds are also possible to Ag, Pt, Cu, Hg, Ge, Ni, and even semiconductor surfaces. The reason for the great success of the S–Au bond is its good stability at ambient temperature, and the ease of reorganization to form an ordered array. Both are elementary requirements for the building up of a self-assembled monolayer.

Besides the prominent thiolates, other functional molecules, such as alcohols (ROH) or acids, have been demonstrated to form organized monolayers on metals or metal oxide surfaces, such as Al_2O_3, TiO_2, ZrO_2, or HfO_2. SAMs of alkylchlorosilanes ($RSiCl_3^-$) and other silane derivatives require hydroxylated surfaces as substrates for their formation. The driving force for this self-assembly is the *in-situ* formation of polysiloxane, which is connected to surface silanol groups (−SiOH) via robust Si−O−Si bridges [32]. Substrates on which these monolayers have successfully been prepared include silicon oxide, aluminum oxide, quartz, glass, and mica.

During the past few years, significant advances have been made by coupling alkenes and alkines onto Si and Si−H surfaces. The covalent coupling of vinyl compounds on H-terminated silicon yields very stable Si−C covalent bonds [33], and recently a method for the direct assembly of aryl groups on silicon and gallium

arsenide using aryl diazonium salts has also been developed. There is a spontaneous ejection of N_2 and direct carbon–silicon formation [34], but the C−Si bonds are so strong that a facile reconstruction in order to form a highly ordered SAM is implausible.

10.3.1.1 Preparation of Substrates

For the deposition of a SAM, a 2-D film with a thickness of one molecule, a high-quality surface with a very low surface roughness is required. Depending on any further use of the SAM, the quality of the surface must be adapted. For example, if the SAMs are applied as etch resists, protection layers, chemical sensors or model surfaces for biological studies, then polycrystalline films will mostly suffice as substrates. In contrast, if the properties of the SAMs themselves are to be studied in detail, such as their organization, structure or electronic properties, then oriented single crystalline surfaces are required as substrates.

Planar substrates for SAMs are either thin films or single crystals of metals, semiconductors, or metal oxides. Thin films can be grown on silicon wafers, glass, single crystals or mica by CVD, physical vapor deposition (PVD), electrodeposition, or electroless deposition. Metal films on glass or silicon are polycrystalline and composed of grains that can range in size from 10 to 1000 nm.

As pseudo "single crystals", thin films of metals on freshly cleaved mica are commonly used. Gold films grow epitaxially with a strongly oriented (1 1 1) texture on the (1 0 0) surface of mica. The films are usually prepared by thermal evaporation of gold at rates of 0.1–0.2 nm s^{-1} onto a heated (400–650 °C) sample [35]. By using an optimized two-step process, a surface roughness down to 0.4 nm over areas of 5×5 μm can be achieved [36]. Surfaces with almost comparable roughness can be created by a method known as "template stripping" [37]. Here, a glass slide or a silicon wafer is glued to the exposed surface of a gold film on mica, and subsequently the gold film is peeled from the mica to expose the surface that had been in direct contact with the mica. Typically, these methods result ultimately in surface roughnesses of 1 nm over areas of 200×200 nm^2. For fundamental studies of SAMs by ultra-high vacuum (UHV) methods, single-crystal metal substrates provide the highest quality with respect to surface roughness, orientation, and cleanliness. These substrates result in densely packed SAMs of the highest order.

10.3.1.2 Preparation of Self-Assembled Monolayers

In principle, there are two possibilities of preparing SAMs, namely deposition from solution, and deposition from the vapor phase. For deposition from solution, a clean, freshly prepared substrate is immersed into a highly diluted solution of the corresponding organic molecules. After only a few minutes of immersion, a dense molecular monolayer is built; however, to ensure that the film reaches equilibrium the substrates are kept in solution for several hours to allow reorganization (Figure 10.14). In particular, the structure of the adsorbate determines the highest achievable density of the SAM on a given surface, or whether a SAM can be formed at all. The other parameters, such as solvent, temperature, concentration and immersion time, should be chosen adequately to achieve the best possible result. The

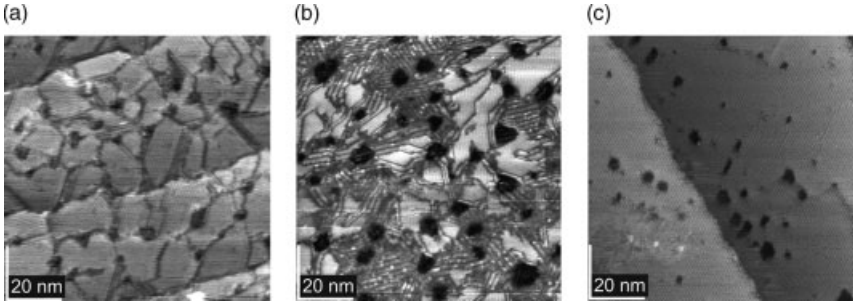


Figure 10.14 (a) Dodecanethiol SAM grown from solution. (b) SAM grown from solution. A 6.5-h annealing step at 78 °C in solution leads to a partial desorption of the dodecanethiol molecules, and results in the striped lying-down phase of alkanethiols [38]. (c) SAM grown from vapor phase. The domains extend over the whole gold terrace.

advantages of this method are the simplicity of the equipment and the ease of preparation.

In the case of alkanethiols on Au(1 1 1), the annealing procedure at elevated temperatures to increase the quality of the films has been studied extensively. Annealing the SAMs in a diluted solution of their molecules for short periods at 80 °C often results in a reduction in the number of vacancy islands, and an enlargement of the domain sizes due to *Ostwald ripening*. This behavior is explained by an intralayer diffusion of monovacancies towards larger holes, which grow at the expense of smaller holes. Furthermore, some vacancy islands diffuse towards the gold step edges and annihilate there, which explains the decrease in area occupied by the vacancy islands. In addition, the conformational defects in the SAMs decrease, and this will result in a higher order.

In the case of gas-phase deposition, UHV systems with base pressures in the range of 10^{-5} to 10^{-7} mbar are used. The amount of deposited molecules is controlled by the pressure, the temperature and the time. Vapor deposition has the advantage that absolutely clean surfaces can be used, a good control of the amount of deposited molecules is possible, and the SAM can be transferred to an analyzing tool without breaking the vacuum. By applying this method, submonolayers and highly ordered monolayers of extreme size can be created (Figure 10.14).

10.3.1.3 **Preparation of Mixed Self-Assembled Monolayers**

Mixed SAMs – that is, SAMs built up from different organic molecules and showing a well-defined structure – can be created in several ways; however, the two most widely used approaches will be described here.

The first method is coadsorption from solutions containing mixtures of selected organic molecules, and results in mixtures of molecular structures (Figure 10.15). This process allows the formation of SAMs with widely varying compositions and physical properties [39, 40].

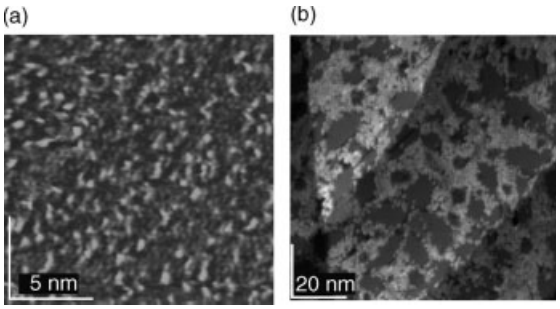


Figure 10.15 Scanning tunneling microscopy images of mixed monolayers. (a) By coadsorption of 11-mercaptoundecanoloctanethiol (=1:3) from solution [40]. (b) By insertion of a biphenylbutanethiol derivate into a closely packed SAM of dodecanethiol on Au(1 1 1). The film shows separate domains of the biphenylbutanethiol derivate (which appear higher) and dodecanethiol [43].

The second method is a two-step deposition process which begins with a full coverage monolayer of one organic species, which is used as the host matrix [41–43]. In a second step, the substrate covered with this host-matrix is immersed into the solution of an organic molecule of interest. Insertion of this "guest-molecule" takes place preferentially at defect sites such as pin holes or domain boundaries in the host matrix. The rate-determining step is the replacement of host molecules by guest-molecules. Depending on the immersion time, domains of inserted molecules, bundles or even single guest molecules can be identified in the resulting mixed monolayer (Figure 10.15). A well-ordered surrounding matrix can be used as the reference system for the analysis of the structural and electrical properties of the inserted molecules. This matrix-isolation method, in combination with scanning probe microscopy (SPM) techniques, is suitable for investigating series of organic molecules in order to determine new physical properties.

10.3.2 Structure of Self-Assembled Monolayers

The structure of SAMs is widely studied using spectroscopic methods including optical ellipsometry, reflectance absorption infrared spectroscopy (RAIRS), X-ray photoelectron spectroscopy (XPS), low-energy electron diffraction (LEED), and high-resolution electron energy loss spectroscopy (HREELS). In particular, the development of near edge X-ray absorption fine structure spectroscopy (NEXAFS) has led to new insights into the structure of SAMs. In addition, an increased understanding of the SAM structures has been achieved by the development and intense use of high-resolution topographic methods such as SPM.

During the self-assembly of organic molecules on planar substrates, complex hierarchical structures are formed involving multiple energy scales and multiple degrees of freedom (Figure 10.16). The geometric arrangement of organic molecules

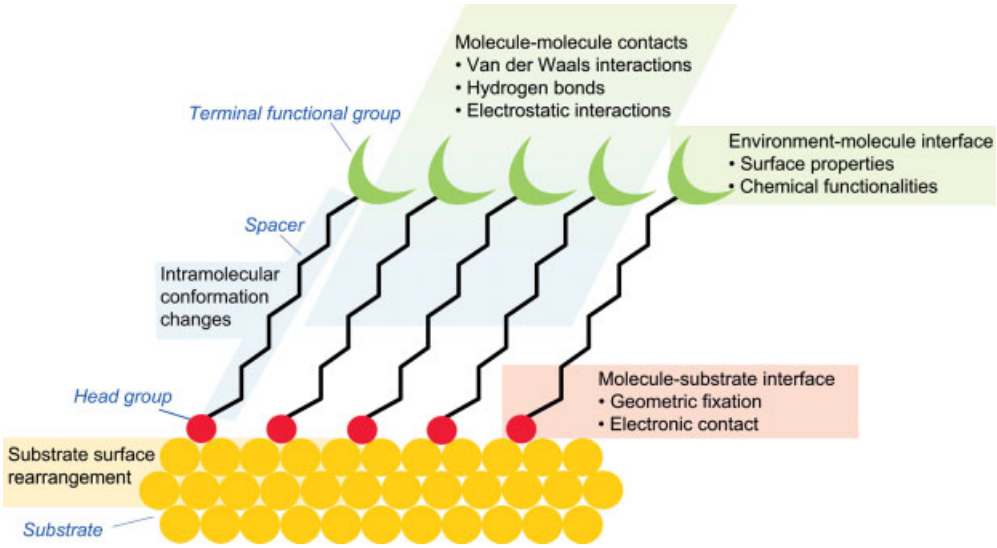


Figure 10.16 A schematic diagram of a SAM, with the characteristic features highlighted.

on a surface is determined in a first level of organization by the footprint of the molecule, the nearest-neighbor distances between the metal atoms at the surface, and the chemical bond formation of the molecules with the surface. The resulting 2-D density of the molecules on the surface may not correspond to the density that the same molecules can attain in crystalline form. In order to minimize the free energy of the organic layer, the molecules perform intramolecular conformation changes such as bond stretches, angle bends, or torsions, which in turn maximize the lateral interactions (e.g., van der Waals interactions, hydrogen bonds, or electrostatic interactions) in a second level of organization. The surface rearrangement of the substrate corresponds to a third level of organization. The balance of these forces determines the specific molecular arrangement, while the driving force is the minimization of the global energy.

10.3.2.1 Organothiols on Metals

The most studied – and probably best understood – SAM is the full-coverage phase of alkanethiols (R−SH) on Au(1 1 1) surfaces. The adsorbing species on the gold surface is the thiolate (RS^-), while the hydrogen atoms are desorbed in form of H_2 molecules with the gold surface acting as catalyst. The thus-formed Au–S bond that anchors the SAM is a strong homolytic bond with a strength on the order of approximately 200 kJ mol^{-1}. The alkanethiols are stabilized by van der Waals interactions between adjacent molecules. These dipole–dipole interactions are proportional to the alkyl chain length (~4.0 kJ mol^{-1} of stabilization to the SAM for each methylene group), and are responsible for the degree of order in the SAM.

A number of studies of alkanethiolate monolayers on gold have shown that the formed structure is commensurate with the sulfur atoms occupying every sixth

hollow site on the Au(1 1 1) surface. The symmetry of the alkanethiolates is hexagonal with a $(\sqrt{3}\times\sqrt{3})$ R30° structure relative to the underlying Au(1 1 1) substrate, a S−S spacing of 0.4995 nm, and a calculated area per molecule of 0.216 nm^2. The alkanethiols are tilted ∼30° off the surface normal, and the hydrocarbon backbones are in all-*trans* configuration. Additionally, the alkanethiolates on Au(1 1 1) surfaces exhibit a c(4×2) superlattice which is characterized by a systematic arrangement of molecules showing a distinct height difference (Figure 10.17) [44]. The height differences in STM images are believed to be due to different conformations of the molecules.

Highly ordered SAMs can easily be built up from alkanethiols, although their structure is affected directly by the addition of any sterically demanding top-end group. The size and the chemical properties (e.g., high polarity) of additionally introduced surface functionalities may reduce the monolayer order.

10.3.2.2 **Carboxylates on Copper**

Compared to the extensive studies of organothiols on gold surfaces, very few investigations have been undertaken to study the self-assembly process of carboxylic acids on metal surfaces. The carboxyl group is known to be an anchoring group for the chemical bonding to metal surfaces. During the adsorption process of simple carboxylic acids the acid group is deprotonated into the carboxylate functionality, resulting in an upright adsorption configuration onto copper or nickel surfaces, as are observed for formic, acetic, and thiophene carboxylic acids [20]. The oxygen atoms in

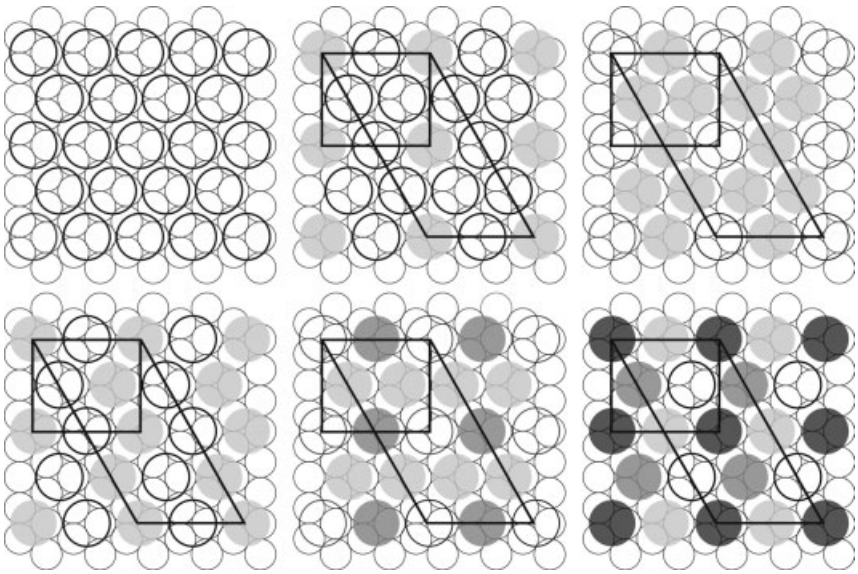


Figure 10.17 Schematic diagram of different phases for the superstructure of alkanethiols on Au(1 1 1) with the $(3\times2\sqrt{3})$ and the c(4×2) superlattice unit cell outlined [44].

the carboxylate group are equidistant to the surface and form a rigid adsorption geometry.

More recently, tartaric acid adsorbed onto copper or nickel surfaces [20] has attracted interest because of a possible use in chiral technology. It is the aim of this technology to establish enantioselective catalytic methods to produce pure enantiomeric forms of materials such as pharmaceuticals and flavors. One way to create heterogeneous chiral catalysts is to adsorb chiral organic molecules at metal surfaces in order to introduce asymmetry. Tartaric acid has two chiral centers, and is therefore of potential interest as chiral modifier; indeed, recently it has been used successfully to stereodirect hydrogenation reactions with a yield of >90% of one enantiomer.

The self-assembly process of *R,R*-tartaric acid on Cu(1 1 0) under varying coverage and temperature conditions leads to a variety of different structures. Due to the two carboxylic acid functionalities, *R,R*-tartaric acid can adsorb in the monotartrate, the bitartrate, or the dimer form. It can be seen from Figure 10.18 that the bitartrate and the dimer–monomer assembly have no symmetry elements and create a chiral surface which is non-superimposable on its mirror image. This is a result of the inherent chirality of the *R,R*-tartaric acid molecules and their two-point bonding at the surface, which uniquely dictates the position of all its functional groups. Subsequently, the intermolecular interactions control the placement of the neighboring molecules. Due to the chirality of the adsorbates the lateral interactions are anisotropic and lead to organized chiral structures.

10.3.3
Supramolecular Nanostructures

Highly ordered 2-D supramolecular nanostructures can be created by using the MBE technique, or at the solid–liquid interface from solution [31] by tuning the molecular backbone size and controlling the supramolecular binding. This approach is based, in principle, on the concepts of supramolecular chemistry directing 3-D structures [16], but in the case of 2-D structures the influence of the substrate must also be taken into account [26].

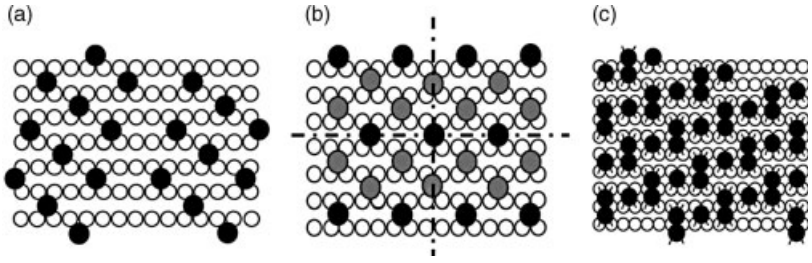


Figure 10.18 Adsorbate templates on Cu(1 1 0) surfaces created by (a) bitartrate (organizational chirality), (b) monotartrate (two symmetry planes), and (c) dimer–monomer assembly (organizational chirality) [20].

A supramolecular 2-D honeycomb network has been created by the assembly of two types of molecule on a silver-terminated silicon surface (this was discussed earlier, in Section 10.2.5) [13]. In the first step, a submonolayer of perylene tetra-carboxylic di-imide (PTCDI) was deposited by sublimation under UHV. Subsequently, melamine (1,3,5-triazine-2,4,6-triamine) was deposited while the sample was annealed at ~100 °C. The substrate allows a free diffusion of the molecules, and this makes formation of the supramolecular network possible. Furthermore, the compatibility of the molecular geometries results in three hydrogen bonds per melamine–PTCDI pair, an intentionally strong heteromolecular hydrogen bonding. In the 2-D honeycomb network the melamine molecules form three-fold connection sites, while the linear PTCDI molecules are used as one-dimensional linkers (Figure 10.19). This ordered array of pores can serve as traps for the co-location of several large molecules. By subliming C_{60} onto the hexagonal network, heptameric C_{60} clusters with a compact hexagonal arrangement are formed within the pores and are clearly stabilized by the PTCDI–melamine network.

The pronounced effect of the substrate in the self-assembly process of 2-D supramolecular structures becomes obvious, when prochiral molecules are used. Upon adsorption, these molecules lose their freedom of rotation and, in consequence, their symmetry and become chiral. Recently, the thermally induced switching of such prochiral molecules between different enantiomeric forms on the surface has been studied [45]. The molecule under investigation was a multiple-substituted phenylenethynylene oligomer (Figure 10.20a). The molecules align into rows and

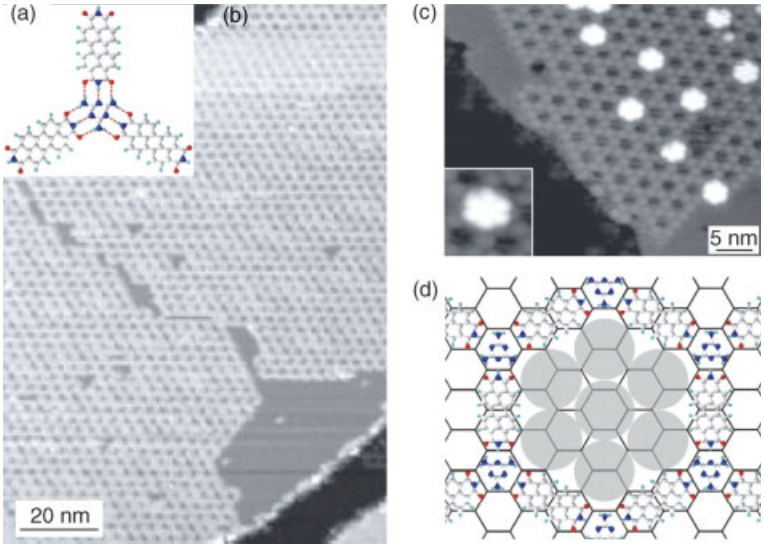


Figure 10.19 (a) Trigonal motif built by perylene tetracarboxylic di-imide and melamine. (b) STM image of a large-area PTCDI–melamine network. (c) STM image of C_{60} heptamers trapped in the pores of the PTCDI–melamine network. (d) Schematic diagram of C_{60} heptamer [13].

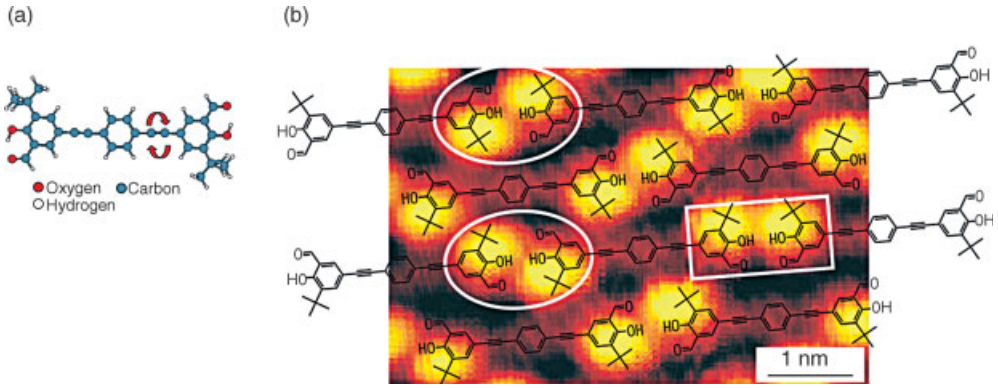


Figure 10.20 (a) The chemical structure of the investigated molecule. (b) A schematic model of the brick-wall adsorption structure superimposed on an STM image [45].

form a brick-wall structure on the Au(1 1 1) surface. The bright protrusions in the STM image (Figure 10.20b) may be attributed to *tert*-butyl groups at both ends of the molecules; the positions of these groups with respect to the molecular backbone identifies the enantiomeric form of the molecule. Two chiral enantiomers (LL and RR) and one achiral meso-form (LR/RL) of this molecule exist. The molecules are not completely stereochemically fixed by the substrate, but change between different surface conformers. An intermolecular *trans* configuration of the headgroups of two adjacent molecules has been found to exhibit the lowest potential energy ($\Delta \sim 4\,\mathrm{kJ\,mol^{-1}}$).

Supramolecular structures built by strong metal–ligand interactions are of high stability, and the incorporated metal centers offer additional functionalities. A system which forms a variety of 2-D surface-supported networks is based on iron (Fe) and aromatic dicarboxylic acids in different relative concentrations on copper surfaces. Mononuclear metal–carboxylate clusters are obtained from one Fe center per four tricarboxylic acid (TCA) molecules on Cu(1 0 0) surfaces [46]. The ($Fe(TPA)_4$) complexes form large, highly ordered arrays which are thought to be stabilized by substrate templating and weak hydrogen bonds between neighboring complexes (Figure 10.21a). From this a perfect arrangement of the Fe ions results, which cannot be achieved by using top-down methods.

A completely different network is obtained when two Fe atoms per three dicarboxylic acid (DCA) molecules are deposited onto the Cu(1 0 0) surface. The resulting array can be described as a ladder structure forming a regular array of nanocavities [26] (Figure 10.21b). The ladders are formed by metal–ligand interactions, while the connections between the ladders are formed by hydrogen bonds. If one Fe atom is deposited per linker molecule, a fully interconnected metal–ligand 2-D network results [47] (Figure 10.21c). By using DCAs of different length as linker molecules between the Fe centers, the size of the resulting nanocavities in the network can be tuned.

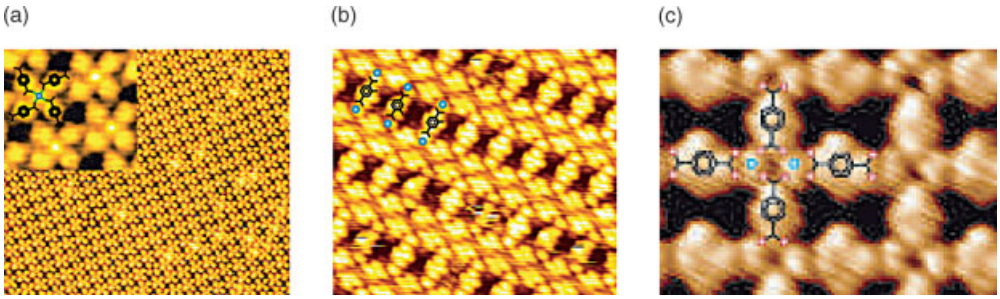


Figure 10.21 Supramolecular assembly of Fe-carboxylate coordination systems on Cu(100) substrates. STM images and schematic models: (a) mononuclear complexes [46]; (b) ladder structure [26]; (c) coordination network [47].

10.3.4 Applications of Self-Assembled Monolayers

10.3.4.1 Surface Modifications

One potential application for SAMs is the modification of surfaces in order to change the surface properties for special applications. For example, polar surfaces can be created by the adsorption of SAMs with terminal groups such as cyano (C≡N). These polar surfaces are useful for the investigation of dipole–dipole interactions in surface adhesion. On the other hand, SAMs with terminal OH groups can vary wetting behaviors, and are used in investigations to study the importance of H-bonding in surface phenomena. Additionally, surface OH and COOH groups – and especially acid chlorides – are very useful groups for chemical transformations. For example, reacting the acid chloride with a carboxylic acid-terminated thiol provides the corresponding thioester. The control of surface reactions opens up the way to chemical sensors [21], and is the basis of chemical force microscopy [22].

10.3.4.2 Adsorption of Nanocomponents

Mixed SAMs containing two or more constituent molecules can be used as test systems to study the interactions of surfaces with bioorganic nanocomponents (proteins, carbohydrates, antibodies). Usually, the SAM contains alkanethiols with a surface terminal group of interest (e.g., suitable for hydrophobic or hydrophilic interactions) and an alkanethiol with a reactive site for linking to a biological ligand. SAMs make it possible to generate surfaces with anchored biomolecules that remain biologically active and in their native conformations.

Additionally, it is possible to use the specific chemical binding properties of SAM surface groups to direct nanocomponents into desired structures. This approach has been used extensively to form selected assemblies of nanoparticles, and opens up a pathway to a variety of different structures (this subject is discussed in detail in Section 10.4.2). Another example of the fabrication of desired structures due to the binding properties of SAMs is the directed assembly of carbon nanotubes (CNTs). Recently, a method was developed which is based on the observation that CNTs are

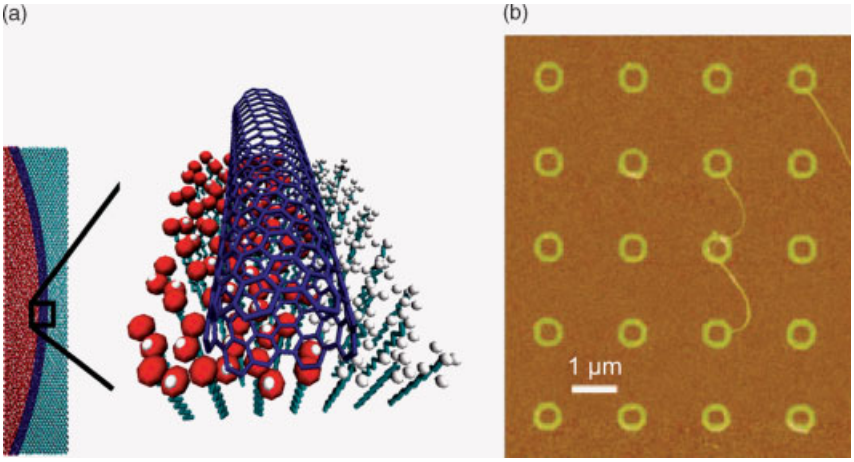


Figure 10.22 (a) Schematic diagram of the directed assembly process. (b) AFM tapping mode topographic image of carbon nanotubes assembled into rings [48].

strongly attracted to COOH-terminated SAMs or to the boundary between COOH- and CH_3-terminated SAMs. By using nanopatterned "affinity templates", desired structures of CNTs can be formed (Figure 10.22). Useful methods for the generation of appropriate templates include dip-pen nanolithography (see Chapter 8) and micro- or nano-contact printing [48].

10.3.4.3 Steps to Nanoelectronic Devices

Molecular electronics requires several structural elements such as wires, diodes, switches, and transistors in order to build up nanodevices. In the studies conducted by Weiss and colleagues [24], conjugated oligophenylene ethynylenes (OPE) have been investigated, which possess potentially interesting features, including negative differential resistance (NDR) (increased resistance with increasing driving voltage), bistable conductance states, and controlled switching under an applied electric field. Single OPEs have been studied in a 2-D isolation matrix of host SAMs of dodecanethiolate on a gold electrode. As a result, series of surface images (Figure 10.23) showed the conductance switching due to conformational changes of the OPE molecules with a low rate, if the surrounding matrix was well ordered. Conversely, when the surrounding matrix was poorly ordered, the inserted molecules switched more often [24]. The switching of OPE molecules can only be observed in arrays of small bundles of molecules. Therefore, it is assumed that the forming and breaking off of hydrogen bonds between adjacent molecules – and the consequent twisting of the molecule, which prevents conjugation of the π-orbitals of the molecular backbone – is responsible for the two conduction states. Such a device, constituted by a switching molecule attached to a bottom electrode and a conductive tip, represents a simple form of a memory.

According to Avriam and Ratner in 1974 [49], the working principle of a molecular diode should be based on two separated electron–donor and electron–acceptor

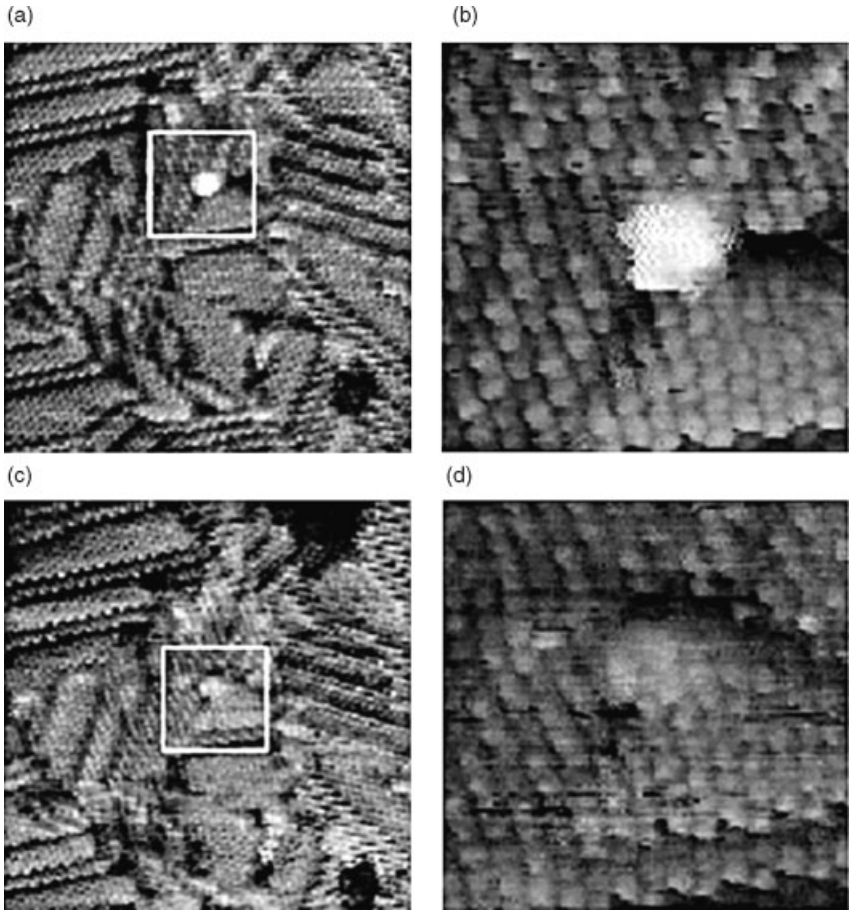


Figure 10.23 Topographic STM images of a molecular switch (OPE) inserted in a dodecanethiol SAM [24]. (a) A 20×20 nm image of the molecule in the ON state; (b) a 5×5 nm image of the same area; (c) 20×20 nm; and (d) 5×5 nm images of the same area, with the molecule in the OFF state.

π-systems (see Chapter 24). However, in recent years asymmetric electron transmission through symmetrical molecules has been also observed by SPM investigations on the molecular level [50]. Diode behavior is possible for symmetric molecules, if they are connected asymmetrically to the electrodes – that is, with two different molecule–electrode spacings, or if asymmetric electrodes are used. A further development of this idea leads to the assumption, that engineering the frontier orbitals of the molecule in the asymmetric junction should make it possible to control the orientation of the diode – that is, whether an alignment of the cathode to the LUMO of the molecule or of the anode to the HOMO is achieved at lower bias. If, additionally, the frontier orbitals of the molecule can be changed reversibly by an electrical pulse, then an optical pulse

or a chemical reaction a transistor would result. An example of such a molecular transistor device based on a supramolecular assembly is given by Rabe and coworkers [51], who used a hexa-peri-hexabenzocoronene (HBC) derivative with six electron-accepting anthraquinones (AQs) symmetrically attached to the HBC, and has the function of an electron donor. The resulting HBC–AQ_6 molecules (Figure 10.24a) were investigated at the HOPG/solution interface, where they form monolayers with an ordered structure. The identification of the conjugated HBC cores and the attached AQ molecules, as well as the recording of the current–voltage (I–V) curves through HBC cores, AQs and alkyl chains was possible by using STM/scanning tunneling spectroscopy (STS).

In a next step, the frontier orbitals of the HBC–AQ_6 molecules were intentionally changed in order to vary the electron transmission properties of the HBC cores. This was achieved by the addition of 9,10-dimethoxyanthracene (DMA) to the solution. DMA is an electron donor which is known to build a charge-transfer complex with AQ. It is remarkable that two different I–V curves through the HBC core are observed, depending on whether or not charge-transfer complexes are coadsorbed next to HBC.

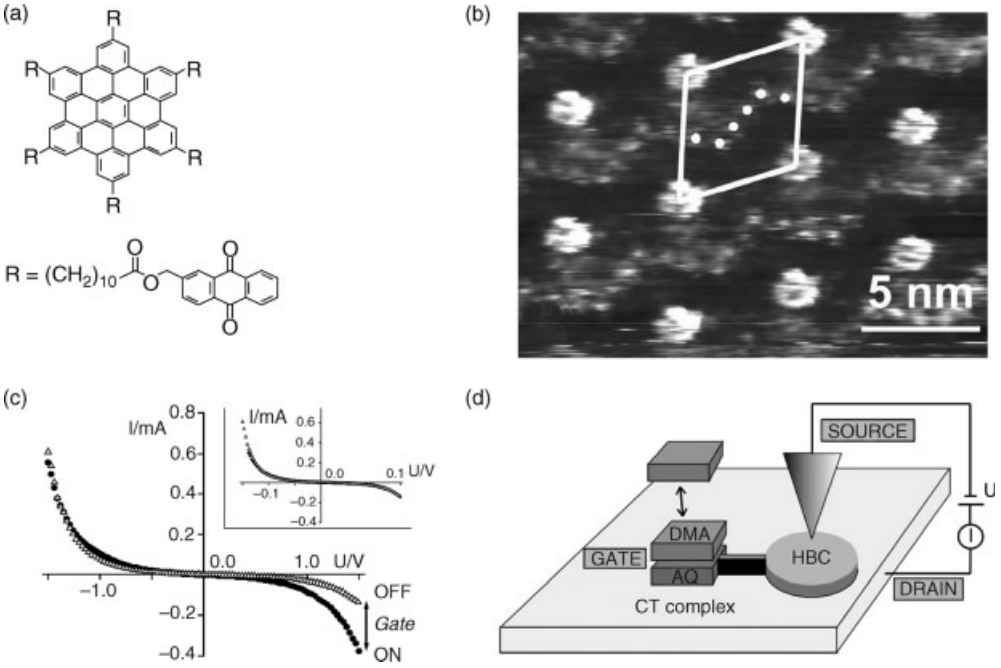


Figure 10.24 (a) Chemical formula of hexa-peri-hexabenzocoronene (HBC) decorated with six anthraquinone (AQ) functions. (b) STM current image of HBC–AQ_6 molecules with coadsorbed charge-transfer (CT) complexes. (c) Current–voltage (I–V) relationships through HBC cores in domains where the charge-transfer complexes are adsorbed, or where no charge-transfer complexes were present. (d) Schematic of a prototypical single-molecule chemical field effect transistor (CFET) [51].

This set-up can be regarded as a "single-molecule chemical field-effect transistor", as the change in the I–V relationship results from the chemical formation/solution of a charge-transfer complex (= gate) which alters the electron transmittance through the covalently attached HBC (= channel) (Figure 10.24). Despite the fact, that the gates cannot be addressed selectively and the device structure changes simultaneously with the electron transmission properties, this approach is a major step towards monomolecular electronics with a complete transistor integrated into one molecule.

10.4 Preparation and Self-Assembly of Metal Nanoparticles

In addition to the previously discussed two nanofabrication methods of epitaxial growth and molecular self-assembly, metal nanoparticles also play an important role in the context of nanofabrication. The extraordinary size-dependent electronic, magnetic and optical properties [52] of nanoparticles have triggered many fascinating ideas for potential applications in breakthrough future technologies, including sensors, medical diagnostics, catalysis, and nanoelectronics. Therefore the preparation and the question of how to assemble metal nanoparticles remain objectives of great interest. The major challenges that are still to be overcome in this context are, on the one hand, the preparation of (ideally) monodisperse metal nanoparticles with simultaneous control over size, shape and composition; and on the other hand, the controlled assembly. Much effort has been expended in attempts to prepare metal nanoparticles of different sizes and shapes, to assemble them into three, two or even one dimensions, and to study and to understand their physical properties. The results of these investigations form the fundamental knowledge for potential applications in nanotechnology. Some of these synthetic routes and self-assembly patterns are outlined in the following paragraphs.

10.4.1 Preparation of Metal Nanoparticles

Generally, metal nanoparticles are prepared by the reduction of a soluble metal salt via suitable reducing agents, or via electrochemically [53–55] or physically assisted methods (e.g., thermolysis [56], sonochemistry [57], photochemistry [58]), or directly via the decomposition of labile zero-valent organometallic complexes. In all cases the synthesis must be performed in the presence of surfactants, which form SAMs on the nanoparticles surfaces (see also Section 10.3) and thus stabilize the formed nanoparticles. The stabilizing effects of the surfactants refer to: (i) steric effects, meaning stabilization due to the required space of the ligand shell; and (ii) electrostatic effects, implying stabilization due to coulombic repulsion between the particles. Furthermore, the surfactants influence the size, shape, and the physical properties and assembly patterns of the nanoparticles. In recent years, many excellent reviews have been produced providing detailed overviews on the preparation techniques, properties and surfactant influences [59–63]. In this context it should be mentioned that,

because of the protecting SAM on the surface, ligand-stabilized nanoparticles are also sometimes denoted as monolayer-protected clusters (MPCs). However, in comparison to SAMs on planar surfaces, the structures of the SAMs on the nanoparticles surfaces differ greatly due to the surface curvature [64].

Important – and already more or less standardized – examples of the preparation of non-stoichiometrically composed metal nanoparticles via the reduction of a metal salt with a suitable reducing agent include the route of Turkevich *et al.* [65] and that of Brust *et al.* [66, 67]. Turkevich and colleagues were the first to introduce a standardized method for the preparation of gold nanoparticles with diameters ranging from 14.5 ± 1.4 nm to 24 ± 2.9 nm, via the reduction of $HAuCl_4$ with sodium citrate in water. Thereby, the nanoparticle size can be controlled by variation of the ratio $HAuCl_4$/sodium citrate. This route is often applied due to the fact that citrate-stabilized gold nanoparticles can simply be surface-modified because of a weak electrostatically bound, and thus easily exchangeable, citrate ligand. Brust *et al.* utilized sodium borohydride as a reducing agent, and took advantage of the high binding affinity of thiols to gold; this enabled the preparation of relatively stable nanoparticles that could be precipitated, redissolved, analyzed chromatographically, and further surface-modified without any apparent change in properties. This high stability represents an important property in terms of controlling nanoparticle assembly.

A recently published report described the surfactant-free synthesis of gold nanoparticles [68]. This approach is especially interesting in terms of the preparation of small gold nanoparticles with narrow dispersity, protected by ligands carrying functional groups that are typically not stable towards reducing agents. In this method, a solution of $HAuCl_4$ in diethyleneglycol dimethyl ether (diglyme) is reduced by a solution of sodium naphthalenide in diglyme to yield weakly solvent-molecule-protected gold nanoparticles. In the first step, these formed nanoparticles are further stabilized and functionalized simply by the addition of various ligands (1-dodecanethiol, dodecaneamine, oleylamine and triphenylphosphine sulfide). The size of the nanoparticles can be tuned within the range of 1.9 to 5.2 nm, with dispersities of 15–20% depending on the volume of the added reduction solution and the time between addition of the reduction solution and the ligand molecule solution.

Magnetic nanoparticles, such as cobalt or iron nanoparticles, are typically prepared via the decomposition of a zero-valent organometallic precursor, for example carbonyl metal complexes. One example is the synthesis of monodisperse ($\pm$ one atomic layer) Co and Fe nanoparticles with sizes of approximately 6 nm via thermal decomposition of the respective carbonyl compounds ($Fe(CO)_5$, $Co_2(CO)_9$) under an inert atmosphere [69]. In this way the nanoparticle size can be controlled by adjusting the temperature and the metal precursor:surfactant ratio. For example, higher temperatures and higher metal precursor:surfactant ratios produce larger nanoparticles. An additional control parameter is the ratio of the surfactants tributyl phosphine and oleic acid, both of which bind to the nanoparticle surface. Tributyl phosphine binds weakly, allowing rapid growth, while oleic acid binds tightly and favors slow growth to produce smaller particles. As might be expected, iron

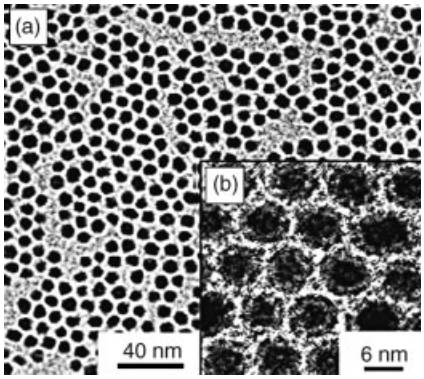


Figure 10.25 (a) TEM image of an ensemble of 6-nm iron nanoparticles. (b) At higher magnification, the surface oxide layer is clearly visible. (Illustration reprinted from Ref. [69], with kind permission.).

nanoparticles show great sensitivity towards oxidation in air; even a short contact of the nanoparticle surface with air resulted in the formation of an Fe_3O_4 layer with a thickness of approximately 2 nm (Figure 10.25).

The thermal decomposition of metal carbonyl complexes for the preparation of nanoparticles or nanostructured materials can also be achieved by treatment with ultrasound. Treatment of a liquid with ultrasound causes the formation, growth and implosive collapse of bubbles in the liquid, and this in turn generates a localized hot-spot [70]. As an example, amorphous Fe/Co nanoparticles are prepared by the sonolysis of $Fe(CO)_5$ and $Co(NO)(CO)_3$ in decanediphenylmethane at 293–300 K under an argon atmosphere, to produce pyrophoric amorphous Fe/Co alloy nanoparticles. Annealing of these particles in an argon atmosphere at 600 °C leads to growth of the Fe/Co particles, and this finally yields air-stable nanocrystalline Fe/Co particles due to carbon coating on the surface [70].

While the above-mentioned examples were all non-stoichiometrically composed gold nanoparticles, one famous example of a stoichiometrically composed gold nanoparticle – and thus great control over size-dependent properties – is the so-called *Schmid cluster* $[Au_{55}(PPh_3)_{12}Cl_6]$, which was introduced in 1981 [71]. The cluster is prepared by the reduction of $Au(PPh_3)Cl$ with *in-situ*-formed B_2H_6 in warm benzene. The relevance of this cluster refers to its quantum size behavior and to the fact that it can be regarded as a prototype of a metallic quantum dot [72, 73]. The defined stoichiometric composition of $Au_{55}(PPh_3)_{12}Cl_6$ is based on the so-called "full-shell cluster principle", whereby the cluster is seen as a cut-out of the metal lattice of the bulk metal. This implies that the cluster consists of a metal nucleus surrounded by shells of close-packed metal atoms, so that each shell has $10n^2 + 2$ atoms (n = number of shells) [59, 74]. Further examples in this context are $[Pt_{309}phen^*_{36}O_{30}]$ (four-shell cluster) and $[Pd_{561}phen_{36}O_{200}]$ (five-shell cluster) (phen* = bathophenanthroline; phen = 1,10-phenanthroline) [75–77].

10.4.2 Assembly of Metal Nanoparticles

As mentioned above, the unique physical properties of metal nanoparticles with diameters of between one and several tens of nanometers make them promising building blocks for the construction of functional nanostructures. Furthermore, it was found that assemblies of nanoparticles show physical properties that are situated between those of an isolated cluster and the bulk material; this in turn would lead to a new class of materials, the properties of which are affected by the nanostructure itself. Arrays of nanoparticles exhibit delocalized electron states that depend on the strength of the electronic coupling between the neighboring nanoparticles, whereby the electronic coupling depends on the particle size, including the particle size distribution, the particle spacing, the packing symmetry and the nature and the covering density of the stabilizing surfactant [78, 79]. Thus, major efforts are under way to organize nanoparticles into one to three dimensions in order to investigate electronic, magnetic and optical coupling phenomena within such assemblies, and even to utilize these coupling effects for the set-up of novel nanoelectronic, diagnostic, or nanomechanical devices [80].

The assembly principle to achieve large ordered arrangements of nanoparticles is the self-organization of ligand-protected nanoparticles due to weak ionic or van der Waals interactions, or due to strong covalent bond formation via respective functional ligand-protected nanoparticles. Some examples of building up 1-D to 3-D nanoparticle arrangements via self-organization are presented in the following sections.

10.4.2.1 Three-Dimensional Assemblies

The easiest achievable construction scheme is the self-assembly into three dimensions which, when occurring spontaneously, is in principle *crystallization*. For example, $Au_{55}(PPh_3)_{12}Cl_6$ crystallizes from dichloromethane solution if the solvent is evaporated, yielding small hexagonal microcrystals. Such a microcrystal obtained in this way is shown in Figure 10.26 [81].

The self-assembly of FePt nanocubes (see also Section 10.3.1) during controlled evaporation of the solvent from a hexane dispersion leads to (1 0 0) textured arrays. This assembly is energetically favored, as it gives the maximum van der Waals interaction energy arising from face-to-face interactions in a short distance of the cube assembly. The interparticle distance is approximately 4–5 nm, which is close to the simple thickness of the surfactant layer (2–2.5 nm, the length of oleate or oleylamine). Interestingly, thermal annealing induces an internal particles structure change and transforms the nanocube assembly from superparamagnetic to ferromagnetic. The study cited here is an example of the influence of particle shape on the assembly scheme.

Co, Ni, and Fe nanoparticles self-assemble from solution to form close-packed nanoparticle arrays on a variety of substrates if the dispersing solvent is evaporated. A TEM image of such a hexagonal superlattice with rows of 8-nm mt-fcc Co nanoparticles aligning to form facets is shown in Figure 10.27. Such 3-D magnetic

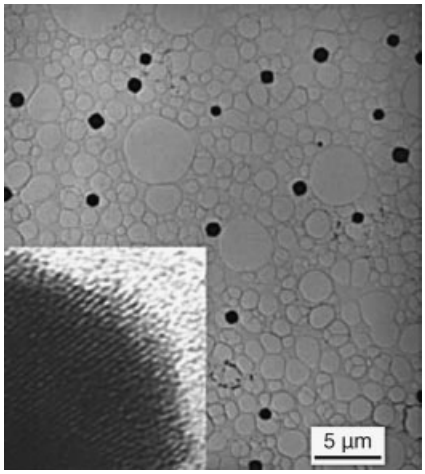


Figure 10.26 TEM image of typical microcrystals of $Au_{55}(PPh_3)_{12}Cl_6$ formed on the grid from dichloromethane solution. The inset shows a high-resolution TEM image of a $Au_{55}(PPh_3)_{12}Cl_6$ microcrystal. (Illustration reprinted from Ref. [81], with kind permission.).

nanoparticle arrays provide important models by which collective interparticle interactions may be explored [69].

Covalently linked gold nanoparticles can be obtained by using bifunctional ligands that are able to interconnect nanoparticles. Brust *et al*. were the first to describe the formation of gold nanoparticle networks by using dithiol molecules as ligands [82]. Their method involved the preparation of gold nanoparticles in a two-phase liquid–liquid system [83], whereby dithiols rather than monothiols were used. The use of dithiols leads directly to the formation of an insoluble precipitate of dithiol crosslinked clusters. The existence of self-assembled nanoparticles is deduced from TEM images (Figure 10.28).

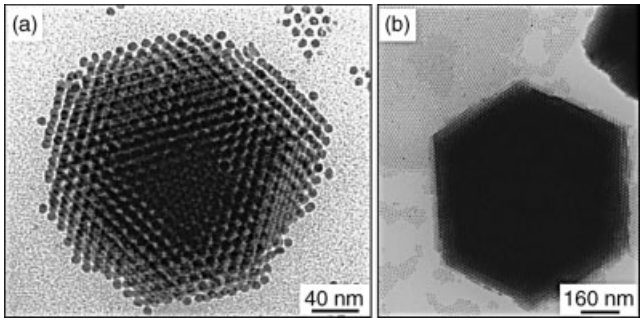


Figure 10.27 (a) TEM image of a hexagonal superlattice grown from 8-nm mt-fcc Co NPs. (b) Lower-magnification image of large hexagons formed in the sample. (Illustration reprinted from Ref. [69], with kind permission.).

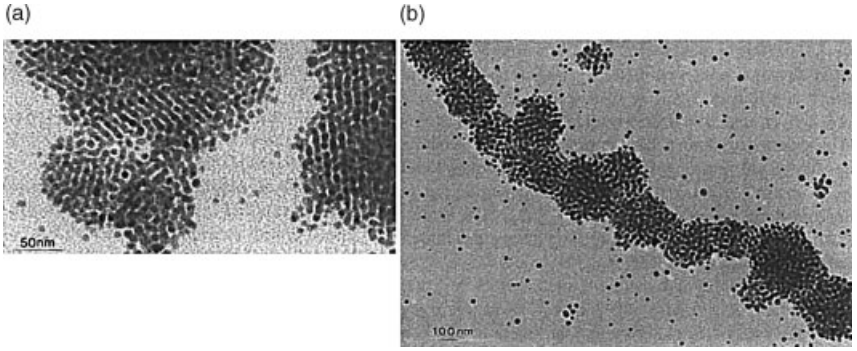


Figure 10.28 (a) TEM image of 8-nm gold nanoparticles crosslinked with 1,9-nonanedithiol, showing the parallel alignment of adjacent particles. (b) TEM image of 8-nm gold nanoparticles crosslinked with 1,9-nonanedithiol, showing a self-assembled string of "superclusters", which is a typical feature of these preparations. (Illustration reprinted from Ref. [82], with kind permission.).

The influence of the ligand, the type, and the linker length on charge-transport properties in metal nanoparticle assemblies has been studied for a variety of cases. For example, the insertion of bifunctional amines into the 3-D arrangement of $Pd_{561}phen_{36}O_{200}$ clusters yields an increased interparticle spacing compared to the closed sphere packing obtained from solution. The increased interparticle spacing is reflected in an increase of the activation energy of the electron transport through the material [84]. The influence of ligand type was recently discussed for the case of Au_{55}-cluster arrangements by Simon and Schmid [85]. When comparing the charge-transport properties of networks of Au_{55}-clusters interconnected by either weak ionic interaction or by covalent bond formation between bifunctional ligands, it transpired that both types showed characteristically different charge-transport properties. The non-covalently interconnected cluster systems showed a continuous increase of activation energy for the charge transport with increasing interparticle distance, whereas the covalently linked cluster systems showed a decrease in activation energy, to significantly lower values [85].

10.4.2.2 Two-Dimensional Assemblies: The Formation of Monolayers

In general, the assembly of nanoparticles into two dimensions is achieved by binding the metal nanoparticles onto a substrate surface. Thus, in order to direct the assembly into a defined pattern, the key point is the existence of functional groups on the substrate surface which enable a specific interaction between the nanoparticle and the surface. These interactions may be either weak electrostatic forces or weaker van der Waals forces; alternatively, covalent bonds may be formed between the monolayer-protected nanoparticle and the substrate surface. The weak interactions have the advantage that they allow a reasonable mobility on the surface, and so enable ordering due to self-organization. Covalent bond formation has the advantage of building

more stable and durable arrays. However, in order to enable electrostatic or covalent interactions, the substrate surface must be adequately modified (see also Section 10.2). The formation of 2-D gold nanoparticle arrays is presented in the following examples.

One example of a 2-D assembly due to electrostatic interactions is the assembly of hydrophilic $Au_{55}(Ph_2PC_6H_4SO_3H)_{12}Cl_6$-clusters on a poly-(ethyleneimine)-coated TEM-grid, as reported by Schmid and coworkers [86]. The driving force for the deposition of colloids on the surface is the acid–base interaction between the $-SO_3H$ group of the ligand and the NH-group of the imine. This procedure leads to a close packing which can be visualized by using TEM (Figure 10.29).

A recently published example for the controlled assembly of gold nanoparticles into two dimensions, and which utilizes electrostatic forces, is the assembly of citrate-protected gold nanoparticles on carbon surfaces [87]. The key point here is that the carbon surfaces are electrochemically modified with primary amines (*n*-hexylamine, tetraethylene glycol diamine). This electrochemical surface modification method allows the number of amine functionalities on the surface to be controlled, and in turn allows control to be exerted on the density of the nanoparticle assembly.

Another example, which is within the context with the electrical characterization of 2-D arrays of gold nanoparticles, is the electrostatic adsorption of 15-nm gold nanoparticles on 3-aminopropyltrimethoxysilane (APTS)-modified silicon substrates. Thereby, the particles are deposited from an aqueous solution (pH 5) to yield a densely packed monolayer of gold colloids with an average density of

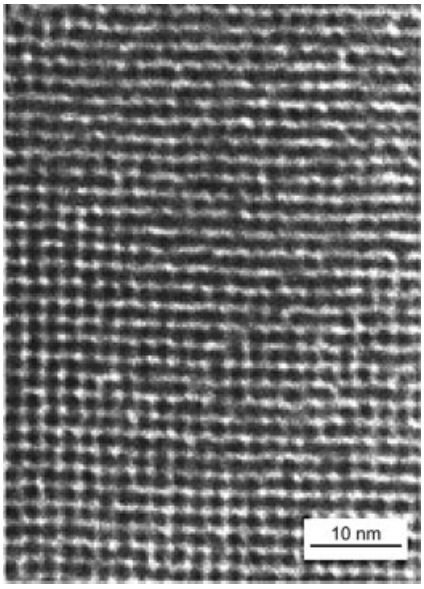


Figure 10.29 $Au_{55}(PPh_3)_{12}Cl_6$-clusters fixed on a PEI-coated grid, as imaged with TEM. (Illustration reprinted from Ref. [86], with kind permission.).

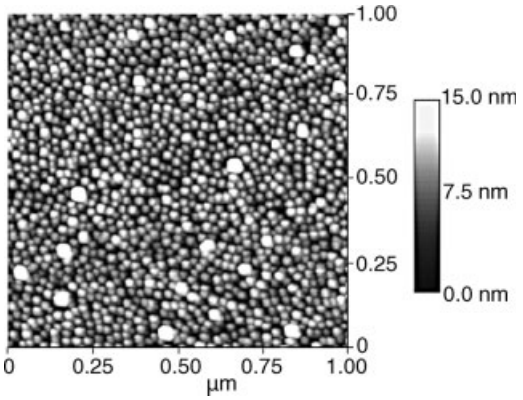


Figure 10.30 AFM image of densely packed citrate stabilized gold nanoparticles on amino-functionalized silicon surfaces. (Illustration reprinted from Ref. [88], with kind permission.).

approximately 1500 particles per μm^2 (Figure 10.30) [88]. The average interparticle spacing is adjusted by the thickness of the citrate shell.

SAMs based on the covalent attachment of nanoparticles were presented by Schmid and coworkers, who described the formation of 2-D arrangements of ligand-stabilized gold clusters and gold colloids on various inorganic conducting and insulating surfaces [89]. For this purpose, oxidized silicon as well as quartz glass surfaces were treated with (3-mercaptopropyl) trimethoxysilane to generate monolayers of the SH-functionalized silane. When dipped into an aqueous solution of 13-nm gold colloids, stable covalent S–Au bonds were formed, thus fixing the colloids in a highly disordered arrangement. The coating of the surface was visualized using AFM.

Microcontact printing (μCP) is also used to create nanoparticle surface assemblies based on covalent forces, and with predefined positions of the nanoparticles within these arrays. A recent example of μCP use was the chemically directed assembly of monolayer-protected gold nanoparticles on lithographically generated patterns [90]. Here, gold surfaces patterned with mercaptohexadecanoic acid (MHA) were prepared using μCP and dip-pen nanolithography. A diamine molecule was used to link the mercaptoundecanoid acid-coated gold nanoparticles onto the MHA-defined patterns. Sonication was then used to remove the non-specifically absorbed nanoparticles, and the nanoparticle assembly was proven by using AFM, with height increases of 2–6 nm with respect to the preformed MHA SAM (Figure 10.31).

10.4.2.3 **One-Dimensional Assemblies**

The 1-D assembly of nanoparticles remains a major challenge, as 1-D (or at least quasi-1-D) assemblies require appropriate nanoparticle surface modification, appropriate templates, or special techniques (including special substrate modifications, e.g., AFM-based methods).

One recently published example is the spontaneous quasi-1-D arrangement of spherical Au nanoparticles protected by a liquid crystal ligand (the 4′-(12-mercapto-

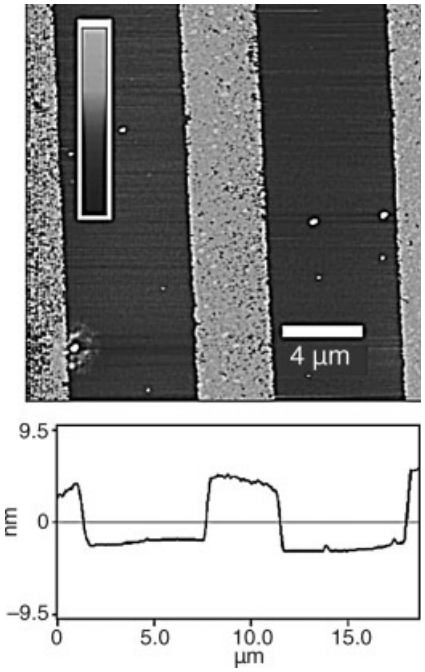


Figure 10.31 Contact mode AFM height image (top image) and average height profile (bottom image) of gold nanoparticles chemically directed onto mercaptohexadecanoic acid (MHA) features patterned by microcontact printing (µCP). The lines show an average height of 6.5 nm (the average height profile is shown at the bottom of the figure), and a 6 nm increase over the feature height after nanoparticle assembly. The inset shows the height scale bar of 20 nm). (Illustration reprinted from Ref. [90], with kind permission.).

dodecyloxy)biphenyl-4-carbonitrile). Thereby, gold nanoparticles protected by a ligand consisting of a liquid crystal mesogen unit and an alkane thiol unit spontaneously ordered themselves just by a simple thermal treatment, without the sue of any templates. The length of the arrangement was 1 to 60 nm, and the inter-array distance was approximately 7 nm (Figure 10.32) [91].

Another approach that utilizes covalent forces is the chemically directed assembly of gold nanoparticles on a thiol-patterned silicon surface [92]. The patterning of the silicon surface is predefined using AFM, and achieved via *chemical force microscopy* [22]. The process is as follows. A silicon surface is modified with a monolayer of 3,5-dimethoxy-α,α-dimethylbenzyloxycarbonyl (DZZ)-protected thiol (DZZ is a photocleavable group typically used in organic synthesis and lithography). It transpired that the application of a voltage between an AFM tip and a selected location on the monolayer yielded deprotection (Figure 10.33), and consequently it was possible to "write" a thiol-end-group pattern into the given monolayer. Following treatment with a 10-nm citrate-stabilized gold nanoparticle solution, the particles were chemisorbed via thiol–gold bond formation. In this way, single lines of gold nanoparticles can be produced, even with a defined particle separation (Figure 10.34) [92].

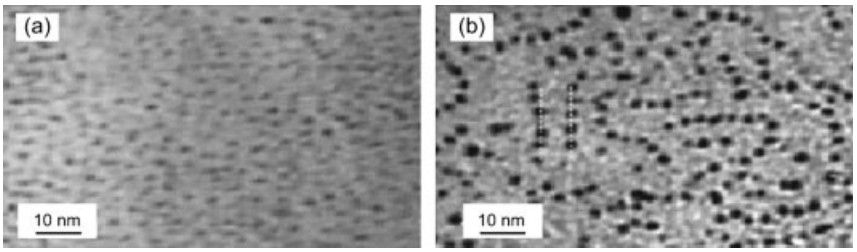


Figure 10.32 TEM images of Au nanoparticles with liquid crystalline ligands, (a) before and (b) after thermal treatment. (Illustration reprinted from Ref. [91], with kind permission.).

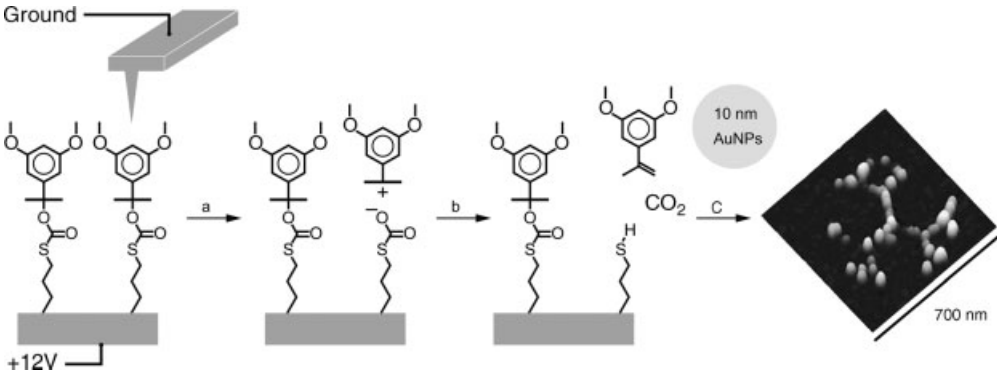


Figure 10.33 (Step a) Electrically stimulated bond cleavage. (Step b) Elimination of carbocation and release of carbon dioxide to produce surface-bound thiol. (Step c) Directed self-assembly of gold nanoparticles into a dendrimer pattern. (Illustration reprinted from Ref. [92], with kind permission.).

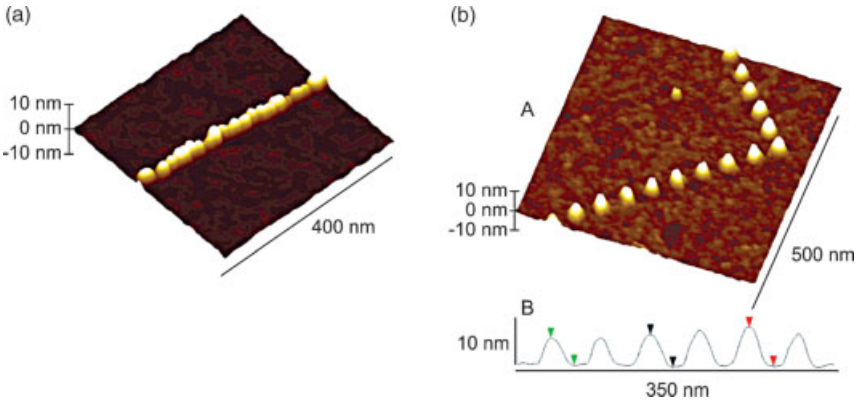


Figure 10.34 (a) A line of gold nanoparticles, of one nanoparticle width. (b) (A) Gold nanoparticles patterned with a 50-nm spacing between individual particles. (B) Cross-sectional analysis of the pattern. (Illustration reprinted from Ref. [92], with kind permission.).

10.5 Conclusions

In this chapter, we have discussed the basic principles and selected examples of the formation of nanostructures via the self-assembly of atoms by epitaxial growth, of molecules, and of metal clusters. The examples presented have provided an impressive demonstration that these methods represent powerful tools for the controlled preparation of highly ordered nanostructures. In future, a combination of the three methods should represent the next key step towards building up well-defined functional nanostructures with suitable properties for applications in molecular electronics. Moreover, besides the applications of self-assembled structures discussed here, such nanoscale structures are of growing interest in the development of new sensors and catalysts. Inorganic substrates with epitaxially grown nanostructures display suitable starting points for the site-selective attachment of molecules, and such molecules may serve as intelligent adhesives for the binding of nanoscale subunits. The future goal of fabrication of complex functional nanoscale structures will be achieved only by employing a hierarchical self-assembly approach, using different scales and materials.

References

1 Vescan, L. (1995) *Handbook of Thin Film Process Technology*, D.A. Glocker and S.I. Shah (Eds.), IOP, Bristol.

2 Kasper, E. (1988) *Silicon Molecular Beam Epitaxy*, Vol. **1–2**, CRC Press.

3 Herman, M.A., Richter, W. and Sitter, H. (2004) *Epitaxy – Physical Principles and Technical Implementation*, Springer.

4 Voigtländer, B. (2001) *Surface Science Reports*, **43**, 127.

5 Venables, J.A. (1994) *Surface Science*, **299/300**, 798.

6 Jesson, D.E., Voigtländer, B. and Kästner, M. (2000) *Physical Review Letters*, **84**, 330.

7 Shchukin, V.A., Lendentsov, A.A. and Bimberg, D. (2003) *Epitaxy of Nanostructures*, Springer, Heidelberg.

8 Heinrichsdorff, F., Ribbat, Ch., Grundmann, M. and Bimberg, D. (2000) *Applied Physics Letters*, **76**, 556–558.

9 Shiryaev, S.Yu., Jensen, F., Lundsgaard Hansen, J., Wulff Petersen, J. and Nylandsted Larsen, A. (1997) *Physical Review Letters*, **78**, 503.

10 Chen, Y., Ohlberg, D.A.A., Medeiros-Ribeiro, G., Chang, Y.A. and Williams, R.S. (2000) *Applied Physics Letters*, **76**, 4004.

11 Kawamura, M., Paul, N., Cherepanov, V. and Voigtländer, B. (2003) *Physical Review Letters*, **91**, 096102.

12 Kim, E.S., Usami, N. and Shiraki, Y. (1998) *Applied Physics Letters*, **72**, 1617.

13 Theobald, J.A., Oxtoby, N.S., Phillips, M.A., Champness, N.R. and Beton, P.H. (2003) *Nature*, **424**, 1029–1031.

14 Butcher, M.J., Nolan, J.W., Hunt, M.R.C., Beton, P.H., Dunsch, L., Kuran, P., Georgi, P. and Dennis, T.J.S. (2001) *Physical Review B-Condensed Matter*, **64**, 195401.

15 Wan, K.J., Lin, X.F. and Nogami, J. (1992) *Physical Review B-Condensed Matter*, **45**, 9509.

16 Lehn, J.M. (1995) *Supramolecular Chemistry*, Wiley-VCH Weinheim, Germany.

17 Prime, K.L. and Whiteside, G.M. (1991) *Science*, **252**, 1164–1167.

18 Arte, S.V., Liedberg, B. and Allara, D.L. (1995) *Langmuir*, **11**, 3882–3893.

19 Scherer, J., Vogt, M.R., Magnussen, O.M. and Behm, R.J. (1997) *Langmuir*, **13**, 7045–7051.

20 Barlow, S.M. and Raval, R. (2003) *Surface Science Reports*, **50**, 201–341.

21 Rickert, J., Weiss, T. and Göppel, W. (1996) *Sensors and Actuators B: Chemical*, **31**, 45–50.

22 Schönherr, H. and Vansso, G.J. (2006) Chemical force microscopy, in *Scanning Probe Microscopies Beyond Imaging* (ed. P. Samori), Wiley-VCH Weinheim, Germany.

23 Joachim, C., Gimzewski, J.K. and Aviram, A. (2000) *Nature*, **408**, 541–548.

24 Donhauser, Z., Mantooth, B., Kelly, K., Bumm, L., Monnell, J., Stapleton, J., Price, D., Rawlett, A., Allara, D., Tour, J. and Weiss, P. (2001) *Science*, **292**, 2303–2307.

25 Gölzhäuser, A., Geyer, W., Stadler, V., Eck, W., Grunze, M., Edinger, K., Weimann, Th. and Hinze, P. (2000) *Journal of Vacuum Science & Technology*, **B18**, 3414–3418.

26 Barth, J.V., Costantini, G. and Kern, K. (2005) *Nature*, **437**, 671–679.

27 Blinov, L.M. (1988) *Soviet Physics Uspekhi*, **31**, 623–644.

28 Metzger, R.M. (2003) *Chemical Reviews*, **103**, 3803–3834.

29 Schreiber, F. (2000) *Progress in Surface Science*, **65**, 151–256.

30 Yang, G. and Liu, G. (2003) *Journal of Physical Chemistry. B*, **107**, 8746–8759.

31 De Feyter, S. and De Schryver, F.C. (2005) *Journal of Physical Chemistry. B*, **109**, 4290–4302.

32 Sugimura, H., Hanji, T., Hayashi, K. and Takai, O. (2002) *Ultramicroscopy*, **91**, 221–226.

33 Buriak, J.M. (2002) *Chemical Reviews*, **102**, 1271–1308.

34 Kosynkin, D.V. and Tour, J.M. (2001) *Organic Letters*, **3**, 993–995.

35 DeRose, J., Thundat, T., Nagahara, L.A. and Lindsay, S.M. (1991) *Surface Science*, **256**, 102–108.

36 Lüssem, B., Karthäuser, S., Haselier, H. and Waser, R. (2005) *Applied Surface Science*, **249**, 197–202.

37 Wagner, P., Hegner, M., Güntherodt, H.J. and Semenza, G. (1995) *Langmuir*, **11**, 3867–3875.

38 Müller-Meskamp, L., Lüssem, B., Karthäuser, S. and Waser, R. (2005) *Journal of Physical Chemistry. B*, **109**, 11424–11426.

39 Bain, C.D. and Whitesides, G.M. (1989) *Journal of the American Chemical Society*, **111**, 7164–7175.

40 Li, L., Cheng, S. and Jiang, S. (2003) *Langmuir*, **19**, 3266–3271.

41 Bumm, L.A., Arnold, J., Cygan, M.T., Dunbar, T.D., Burgin, T.P., Jones, L., Allara, D.L., Tour, J.M. and Weiss, P.S. (1996) *Science*, **271**, 1705–1707.

42 Moth-Poulsen, K., Patrone, L., Stuhr-Hansen, N., Christensen, J.B., Bourgoin, J.-P. and Bjornholm, T. (2005) *Nano Letters*, **5**, 783–785.

43 Lüssem, B., Müller-Meskamp, L., Karthäuser, S., Homberger, M., Simon, U. and Waser, R. (2006) *Langmuir*, **22**, 3021–3027.

44 Lüssem, B., Müller-Meskamp, L., Karthäuser, S. and Waser, R. (2005) *Langmuir*, **21**, 5256–5258.

45 Weigelt, S., Busse, C., Petersen, L., Rauls, E., Hammer, B., Gothelf, K., Besenbacher, F. and Linderoth, T.R. (2006) *Nature Mater*, **5**, 112–117.

46 Dmitriev, A., Spillmann, H., Lingenfelder, M., Lin, N., Barth, J.V. and Kern, K. (2004) *Langmuir*, **20**, 4799–4801.

47 Stephanow, S. *et al.* (2004) *Nature Mater*, **3**, 229–233.

48 Wang, Y., Maspoch, D., Zou, S., Schatz, G.C., Smalley, R.E. and Mirkin, C.A. (2006) *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 2026–2031.

49 Avriam, A. and Ratner, M. (1974) *Chemical Physics Letters*, **29**, 277–283.

50 Stabel, A., Herwig, P., Müllen, K. and Rabe, J.P. (1995) *Angewandte Chemie-International Edition*, **34**, 1609–1611.

51 Jäckel, F., Watson, M.D., Müllen, K. and Rabe, J.P. (2004) *Physical Review Letters*, **92**, 188303-1–188303-4.

52 Halperin, W.P. (1986) *Reviews of Modern Physics*, **58**, 533–606.

53 Yu, Y., Chang, S., Lee, C. and Wang, C.R.C. (1997) *Journal of Physical Chemistry. B*, **101**, 6661–6664.

54 Mohamed, M.B., Ismail, K.Z., Link, S. and El-Sayed, M.A. (1998) *Journal of Physical Chemistry. B*, **102**, 9370–9374.

55 Ma, H., Yin, B., Wang, S., Jiao, Y., Pan, W., Huang, S., Chen, S. and Meng, F. (2004) *ChemPhysChem*, **5**, 68–75.

56 Nakamoto, M., Kahiwagi, Y. and Yamamoto, M. (2005) *Inorganica Chimica Acta*, **358**, 4229–4236.

57 Okitsu, K., Mizukoshi, Y., Bandow, H., Maeda, Y., Yamamote, T. and Nagata, Y. (1996) *Ultrasonics Sonochemistry*, **3**, S249–S251.

58 Mallick, M., J. Witcomb, M. and Scurrell, M. (2004) *Journal of Materials Science*, **39**, 4459–4463.

59 Schmid, G. (2004) *Nanoparticles – From Theory to Applications*, Wiley-VCH, Weinheim.

60 Richards, R. and Bönnemann, H. (2005) *Nanofabrication Towards Biomedical Applications: Techniques, Tools, Applications, and Impact* (eds C.S.S.R. Kumar, J. Hormes, C. Leuschner), Wiley-VCH, Weinheim.

61 Daniel, M. and Astruc, D. (2004) *Chemical Reviews*, **104**, 293–346.

62 Burda, C., Chen, X., Narayanan, R. and El-Sayed, M.A. (2005) *Chemical Reviews*, **105**, 1025–1102.

63 Pileni, M. (2003) *Nature Mater*, **2**, 145–150.

64 Love, J.C., Estroff, L.A., Kriebel, J.K., Nuzzo, R.G. and Whitesides, G.M. (2005) *Chemical Reviews*, **105**, 1103–1169.

65 Turkevich, J., Stevenson, P.C. and Hiller, J. (1951) *Discuss Faraday Soc*, **11**, 55–75.

66 Brust, M., Walker, M., Bethell, D., Schiffrin, D.J. and Whyman, R. (1994) *Journal of the Chemical Society: Chemical Communications*, **7**, 801–802.

67 Brust, M., Fink, J., Bethell, D. , Schiffrin, D.J. and Kiely, C. (1995) *Journal of the Chemical Society: Chemical Communications*, **24**, 1655–1656.

68 Schulz-Dobrick, M., Srathy, K.V. and Jansen, M. (2005) *Journal of the American Chemical Society*, **127**, 12816–12817.

69 Murray, C.B., Sun, S., Doyle, H. and Betley, T. (2003) *MRS Bulletin*, **26**, 985–991.

70 Li, Q., Li, H., Pol, V.G., Bruckental, I., Koltypin, Y., Calderon-Moreno, J., Nowik, I. and Gedanken, A. (2003) *New Journal of Chemistry*, **27**, 1194–1199.

71 Schmid, G., Boese, R., Pfeil, R., Bandermann, F., Meyer, S., Calis, G.H.M. and van der Velden, J.W.A. (1989) *Chemische Berichte*, **114**, 3634–3642.

72 Schmid, G. (1998) *Journal of The Chemical Society-Dalton Transactions*, **7**, 1077–1082.

73 Simon, U., Schön, G. and Schmid, G. (1993) *Angewandte Chemie*, **105**, 264–267. (b) Simon, U., Schön, G. and Schmid, G. (1993) *Angewandte Chemie-International Edition*, **32**, 250–254.

74 Schmid, G., Lehnert, A., Kreibig, U., Damczyk, Z.A. and Belouschek, P. (1990) *Zeitschrift Fur Naturforschung Section B-A Journal of Chemical Sciences*, **45b**, 989–994.

75 Schmid, G., Morun, B. and Malm, J. (1989) *Angewandte Chemie*, **101**, 772–773. (b) Schmid, G., Morun, B. and Malm, J. (1989) *Angewandte Chemie-International Edition*, **28**, 778–780.

76 Vargaftik, M.N., Zagorodnikov, V.P., Stolyarov, I.P., Moiseev, I.I., Likholobov, V.A., Kochubey, D.I., Chuvilin, A.L., Zaikovsky, V.I., Zamaraev, K.I. and Timofeeva, G.I. (1985) *Journal of the Chemical Society: Chemical Communications*, **14**, 937–939.

77 Moiseev, I.I., Vargaftik, M.N., Chernysheva, T.V., Stromnova, T.A., Gekhman, A.E., Tsirkov, G.A. and Makhlina, A.M. (1996) *Journal of Molecular Catalysis A-Chemical*, **108**, 77–85.

78 Remacle, F. and Levine, R.D. (2001) *ChemPhysChem*, **2**, 20–36.

79 Mote, L., Courty, A., Ngo, A., Sisiecki, I. and Pileni, M.,(Eds.), (2005) *Self-Organization of Inorganic Nanocrystals, in Nanocrystals Forming Mesoscopic Structures*,

Wiley-VCH, Weinheim. (b) Willner, I. and Katz, E. (2004) *Angewandte Chemie*, **116**, 6166–6235.

80 Willner, I. and Katz, E. (2004) *Angewandte Chemie-International Edition*, **43**, 6042–6108.

81 Schmid, G., Pugin, R., Sawitowski, T., Simon, U. and Marler, B. (1999) *Chemical Communications*, **14**, 1303–1304.

82 Brust, M., Bethell, D., Schiffrin, D.J. and Kiely, C.J. (1995) *Advanced Materials*, **7**, 795–797.

83 Brust, M., Walker, M., Bethell, D., Schiffrin, D.J. and Whyman, R. (1994) *Journal of the Chemical Society: Chemical Communications*, **7**, 801–802.

84 Simon, U., Flesch, R., Wiggers, H., Schön, G. and Schmid, G. (1998) *Journal of Materials Chemistry*, **8**, 517.

85 Schmid, G. and Simon, U. (2005) *Chemical Communications*, 697–710. (b) Schmid, G., Bäumle, M. and Beyer, N. (2000) *Angewandte Chemie*, **112**, 187–189.

86 Schmid, G., Bäumle, M. and Beyer, N. (2000) *Angewandte Chemie-International Edition*, **39**, 181–183.

87 Downard, A.J., Tan, E.S.Q. and C. Yu, S.S. (2006) *New Journal of Chemistry*, **30**, 1283–1288.

88 Koplin, E., Niemeyer, C.M. and Simon, U. (2006) *Journal of Materials Chemistry*, **16**, 1338–1344.

89 Schmid, G., Peschel, S. and Sawitowski, T. (1997) *Zeitschrift für Anorganische und Allgemeine Chemie*, **623**, 719–723.

90 Barsotti, R.J. Jr, and Stellacci, F. (2006) *Journal of Materials Chemistry*, **16**, 962–965.

91 In, I., Jun, Y., Kim, Y.J. and Kim, S.Y. (2005) *Chemical Communications*, 800–801.

92 Fresco, Z.M. and Frechet, J.M.J. (2005) *Journal of the American Chemical Society*, **127**, 8302–8303.

III
High-Density Memories

Nanotechnology. Volume 3: Information Technology. Edited by Rainer Waser

ISBN: 978-3-527-31738-7

11
Flash-Type Memories

Thomas Mikolajick

11.1
Introduction

The trend towards mobile electronic devices drives an increasing demand for non-volatile memories [1]. In 2007, the market for NAND Flash memories approached the size of the dynamic random access memory (DRAM) market with regards to bit volume, and continues to grow. The hierarchy of today's non-volatile memories is illustrated schematically in Figure 11.1. From this, with the technologies available today, a trade-off between cost and flexibility must be made.

At the low end of flexibility there is the read-only memory (ROM), which can only be programmed during production, but delivers the lowest cost per bit. In a practical application the most important feature today is the *electrical rewritability*. The electrical erasable and electrical programmable ROM allows for this reprogrammability on a Byte level. To achieve this, each memory cell must be constructed from two transistors – the storage transistor and a select device – but this leads to a large cell size. The electrical programmable memory (EPROM), on the other hand, consists of only one memory transistor, but does not allow an electrical erase. The Flash-type memory combines the small cell size of the EPROM with the electrical erasability of the EEPROM simply by allowing the erase operation not on a Byte level but only on large blocks of 16 kB to 1 MB (see Figure 11.2). At the high end of flexibility there is a memory that allows random access-like operation as in DRAMs or static random access memories (SRAMs), and is non-volatile. Today, this non-volatile RAM can only be realized by the combination of DRAM or SRAM [2] with EEPROM or Flash, or by ferroelectric RAMs [3].

Today's standalone Flash memories can be divided into memories for code applications and for data applications. In the code application, the memory must allow a fast random access to enable real-time code execution. In the data application, the focus is on highest density and fast program and erase throughput. The implications of this difference on the array architecture and cell construction will be explained in Sections 11.3.2 and 11.3.3. Additionally a number of applications

Nanotechnology. Volume 3: Information Technology. Edited by Rainer Waser

ISBN: 978-3-527-31738-7

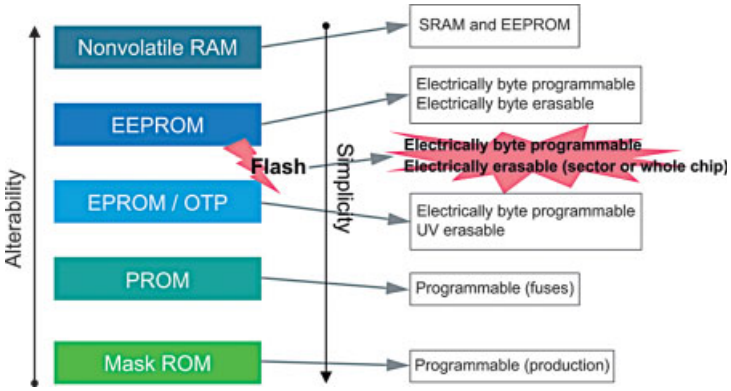


Figure 11.1 Hierarchy of today's non-volatile memory devices. With the available technologies a trade-off between simplicity (reflecting cost) and alterability exists. Flash memories that are programmable on a single byte or page level and erasable on large blocks have evolved as the best compromise for many mobile applications, such as cell phones.

such as "smart cards" call for Flash memory embedded into a high-performance logic circuit [4]. In the embedded Flash segment the density is typically much lower than in standalone memories. Therefore, the focus lies on easy integration into the standard complementary metal oxide–semiconductor (CMOS) flow and low design circuit overhead for the memory module. The requirements, however, are dependent on the actual application, which in turn leads to the development of a large number of different concepts for embedded Flash memories. In the standalone segment of the market, in contrast, one mainstream solution for code and one mainstream solution for data Flash memories have evolved.

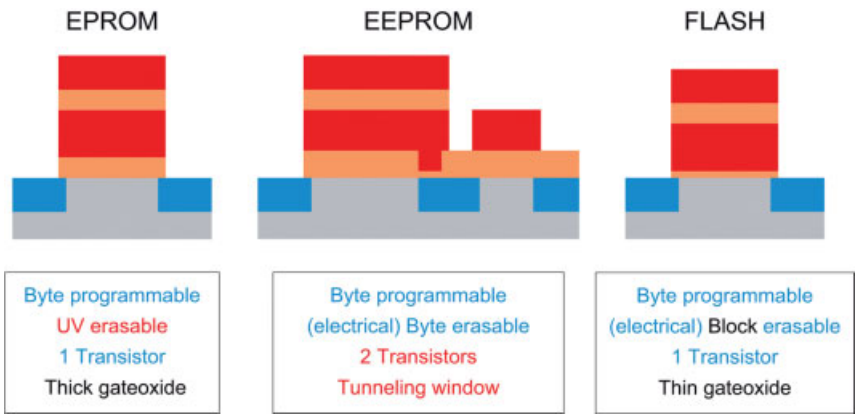


Figure 11.2 A comparison of floating-gate-based EPROM, EEROM and Flash memory cells. By sacrificing on the erase flexibility, a Flash memory cell can be realized with one transistor only.

At the heart of every non-volatile memory today there is either a floating-gate or a charge-trapping transistor, both of which were invented in 1967 [5, 6]. Due to the low oxide quality that was available during the late 1960s, a floating-gate transistor with good retention was difficult to achieve, and charge trapping was therefore very successful until the 1980s. However, with the improvement of oxides, the FAMOS cell [7], which is based on a floating gate that is programmed by avalanche injection and constructed similar to the cell shown at the left of Figure 11.2, was successfully introduced in 1971 for EPROM-type memories. Early EEPROM memories were based on charge-trapping devices using a trigate cell [8]. In 1980, the FLOTOX cell [9] (which is similar to the cell shown in the center of Figure 11.2) was demonstrated and became the mainstream for EEPROM memories. The first Flash memory was introduced at IEDM in 1985 [10], and in 1988 the ETOX cell [11] – which today is the mainstream for NOR-type memories – was proposed. Finally, NAND Flash [12] – the standard for data Flash memories – was first reported at the 1988 VLSI Technology Symposium. In 1987, charge trapping was revived by introducing a cell programmed by hot electrons and erased by hot holes [13]. This allowed the solution of the basic retention issue of charge-trapping devices, and in 1999 it was first demonstrated that this concept could be used to store two physically separated bits in one memory cell, thus reducing the effective cell size below the lithographic limit [14].

11.2
Basics of Flash Memories

In order to integrate a Flash memory into a product, two basic elements are required. First, a memory cell that can perform the program, erase and read operation with the required parameters is necessary. A generic charge storage memory cell, illustrating the program, erase and read operation, is shown in Figure 11.3. To build a large memory, these memory cells must be connected into memory arrays, with the final memory parameters being governed by the combination of memory cell and array architecture.

11.2.1
Programming and Erase Mechanisms

In programming and erase, the charge must be transferred to and from the charge storage layer, overcoming the large potential barrier of the bottom or tunneling dielectric. In principle, two different mechanisms are possible (see Figure 11.4). In the *hot carrier injection mode*, the energy of the carriers is heated up to a level which is sufficient to overcome the barrier. In the *tunneling mode* a large voltage is applied to the barrier in order to reduce its effective width. Variants of both effects are used in different type of Flash concepts. The ways in which several combinations of these effects are realized in Flash concepts are listed in Table 11.1, and the most important concepts will be explained in Section 11.3. Details of other concepts may be found in the references listed in Table 11.1. At this point it should be noted that in Table 11.1

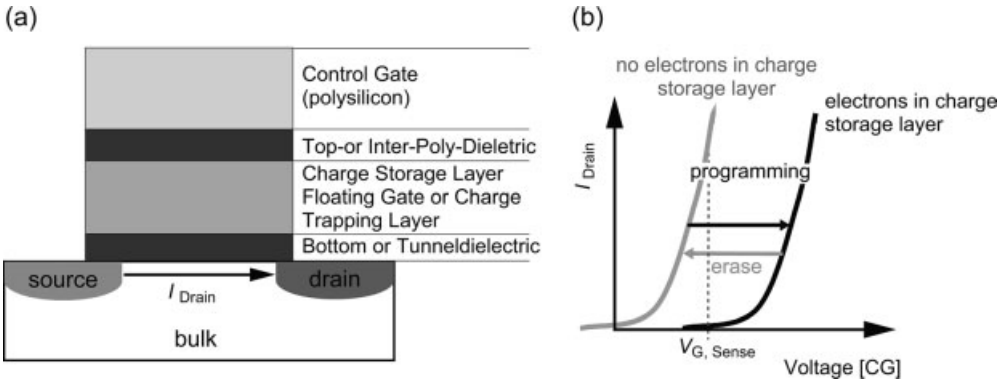


Figure 11.3 Generic charge storage memory cell (a) and basic cell operation (b). The charge storage layer, which can be either a floating gate (Section 11.3) or a charge-trapping layer (Section 11.4) is separated from both the control gate as well as the transistor channel by an insulator. By placing electrons or holes inside of the charge storage layer, the threshold voltage of the transistor can be controlled leading to a significant difference in drain current at a given gate voltage.

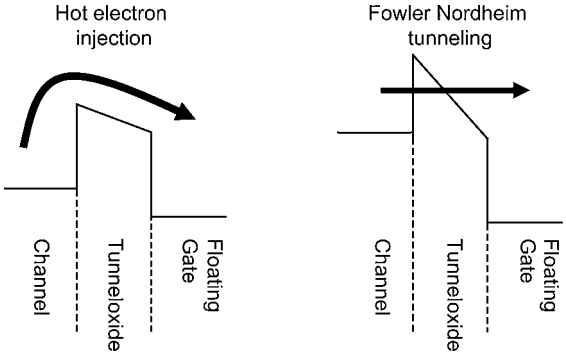


Figure 11.4 Two ways to overcome the silicon/silicon dioxide barrier. In the hot electron injection, carriers are accelerated until they have enough energy to surmount the barrier. In Fowler–Nordheim tunneling, a large electric field is applied to the barrier, leading to a reduction of the effective thickness of the barrier.

and the remainder of this chapter, the operation performed on a byte, word or page level is called "programming" and the operation performed on a block level is called "erase".

11.2.1.1 **Hot Carrier Injection**

In channel hot electron programming, the electrons are accelerated until they have enough energy to surmount the barrier between silicon and silicon dioxide. For electrons this barrier is about 3.1 eV [18], whilst for holes the silicon bandgap must be

Table 11.1 Combination of program and erase mechanisms used in different Flash cell concepts.

Erase	Programming		Hot Electrons	Hot Electrons with secondary impact ionization	Source Side Injection	Hot Holes (BBT)	Fowler Nordheim Tunneling		
							from channel	from Drain	from Source and Drain
Fowler	to Source	ETOX till 0.18 µm [11]	CHISEL [19]						
Nordheim	to Drain			HMOS [21]		AND [32] DINOR [16]	FLOTOX [9]	HiCR [17]	
Tunneling	to Channel	ETOX below 0.18 µm [15]		AG-AND [44]	PHINES [88]	NAND [12] UCP [34] SONOS [70]			
	to Poly	Triple Poly Split gate [31]		Field Enhancing Tunneling Injector [47, 48]	PHINES [89]				
Hot Holes (BBT)		NROM [14], TwinFlash [78], MirrorBit [76]		Twin MONOS [85, 86]					

added, resulting in a barrier of 4.2 eV. To achieve this energy, a high field must be generated in the channel by applying a sufficiently high drain voltage. Additionally, a gate voltage that attracts the generated carriers must be applied. This method has the advantage of microsecond programming speed for a single bit, as well as the fact that it is a three-terminal operation making the disturb optimization easy in a NOR-type architecture. On the other hand, the mechanism has the problem of being very ineffective, as typically approximately 10^5 to 10^6 channel electrons are needed to inject one electron into the storage layer. This leads to a current consumption which is in the range of 100 μA per cell. The consequence is a limited parallelism of cells during programming, and therefore a limited programming throughput.

In order to reduce both the programming current as well as the drain voltage required during programming, two alternative approaches are possible (see Figure 11.5). First, it is possible to apply a bulk voltage during programming, and in this case the field between drain and bulk is increased. As hot electrons at the drain side will lead to electron hole pair creation by impact ionization, the generated holes can be accelerated to the bulk and thereby create a second impact ionization. The so-generated tertiary electrons again may be accelerated towards the charge storage layer. By using this approach the drain voltage can be reduced below 3 V and the current can also be significantly reduced [19]. However, care must be taken to maintain the benefit when scaling down the channel length [20].

Another approach is to use a so-called "split gate transistor", where the channel region is divided into two serial regions with individual gates. The storage layer is present only below one of the two gates. The gate voltage at the gate close to the source is chosen at a value slightly above the threshold voltage, which limits the current flowing trough the channel. The voltage of the second gate is set to a high enough voltage to accelerate the carriers into the charge storage layer. By doing this, a high

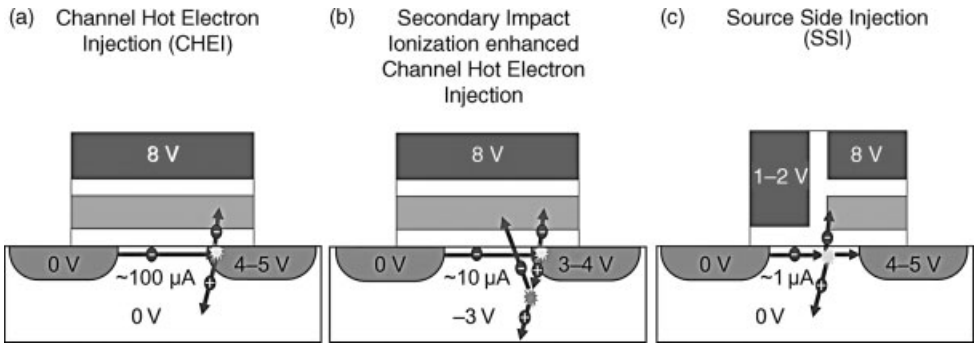


Figure 11.5 Channel hot carrier injection mechanisms. In the classical channel hot electron injection, the injection is mainly by primary channel electrons or secondary electrons (a). With applied back bias, the injection current is significantly increased by the carriers additionally generated during the secondary impact ionization event (b). In source side injection mode the channel current is limited by a second control gate and the field for hot carrier generation is decoupled from the field that attracts the carriers to the storage layer (c). This enhances the efficiency by about 2 orders of magnitude.

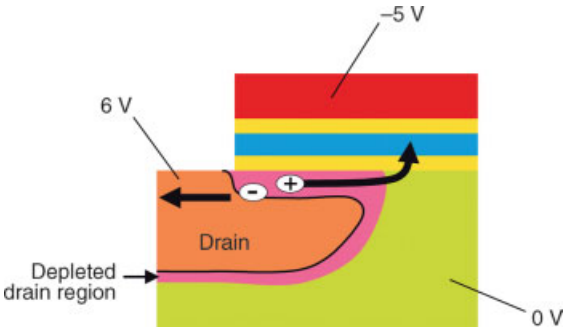


Figure 11.6 Generation of hot holes by band to band tunneling. The band bending in the highly doped region of the n^+ junction leads to generation of carriers by band to band tunneling. The carriers are heated by the lateral electrical field and attracted to the storage layer via the vertical electrical field.

field is created at the region between the two gates, such that the electrons will become hot in that area of the device. As the carriers are injected at a location close to the source, this method is referred to as "source side injection" (SSI) [21]. The channel current can be reduced to the single μA range. Due to the necessity of a second gate electrode, however, there is a cell size drawback and a circuit overhead associated with this source side injection.

Besides channel hot electron generation, carriers generated by band to band tunneling [27] may also be used for programming and erase. In this case, band-to-band tunneling in the drain junction is induced by applying a high drain potential while the gate is turned off. The so-generated carriers are accelerated towards the channel by the electrical field, collecting enough energy to surmount the potential barrier. This is illustrated in Figure 11.6 for the case of hot hole generation in a n-channel device. A modified version uses a highly doped buried layer to generate the band-to-band tunneling [28].

11.2.1.2 Fowler–NordheimTunneling

In Fowler–Nordheim (FN) tunneling, a high electric filed applied to the barrier creates a trapezoidal barrier which significantly reduced the effective barrier for the carriers (see Figure 11.4). The current can be calculated according to the well-known equation [29]:

$$I_G = A_{\mathrm{FN}} E_{\mathrm{ox}}^2 \exp\left(-\frac{B_{\mathrm{FN}}}{E_{\mathrm{ox}}}\right) \tag{11.1}$$

where E_{ox} is the electrical field in the oxide and A_{FN} and B_{FN} are material-specific constants. If a dielectric charge-trapping layer is used for charge storage, the material stack relevant for the tunneling will also depend on the applied field [33]. For low fields, the carriers must tunnel through part of the trapping layer in addition to the tunneling dielectric. For higher fields and very thin layers, direct tunneling is

dominant. Finally, for high fields the FN tunneling [as given by Equation 11.1] is the most important mechanism.

Until now, we have been considering charge transfer between the transistor channel and the charge storage layer. However, by using tunneling the charge can also be transferred to the gate electrode [30] or to a specially designed erase gate [31]. In this case the modified properties of the tunneling barrier created on a polysilicon electrode, as well as field enhancement by poly tips, must be taken into account in the FN tunneling current.

11.2.1.3 **Array Architecture**

When combining memory cells to a memory array, different versions are possible. The straightforward way is to connect the gate of every memory cell to a wordline, the drain of every memory cell to the bitline, and to connect all the sources of the memory cells to ground. By using this construction all n cells on one bitline are connected in parallel to each other. As this resembles the n-channel portion of a n-input NOR gate, this architecture is referred to as NOR, or more precisely common ground NOR. Other types of NOR architecture will be discussed later in this section. In contrast, it is also possible to connect the cells on the bitline in a serial connection, leading to a NAND-type array. In practical applications the number of cells in series must be limited to keep the read current on an acceptable level; therefore, typically 32 cells are placed between two select transistors and the so-constructed NAND strings are connected to the bitline in a similar manner as the individual cells in the standard NOR arrangement. When comparing NAND and NOR architectures, it is clear that the random access time is much faster in the NOR-type array, as every cell is directly accessed by a bitline. In the NAND architecture, in contrast, the serial connection of cells results in a high resistance through which the current must flow to the bitline. The result is a much lower read current – and therefore a much slower random access. On the other hand, the NAND arrangement has a distinct size advantage, as the contacts are shared between all cells of the string, whereas one contact for every two cells is necessary in common ground NOR.

The program and erase mechanism used in the cell also has an impact on the possible array architecture. For a cell programmed by channel hot electron injection, the common ground NOR arrangement is an ideal fit, as the necessary voltages can be precisely applied to the cell, thus minimizing any disturbance to other cells. For other injection mechanisms, or to minimize the cell size, the common ground NOR architecture may be modified. An overview of the possible array architectures is shown in Figure 11.7. In NOR, as well as the above-mentioned common ground NOR architecture, a separate source line may also be used for every bitline; such as array is commonly known as an AND [32] array. Although the most compact realization uses buried bitlines, some versions with metal bitlines are also used [34] if access time has to be minimized rather than cell size. Finally, in such an array each pair of neighboring bit line and source line can be combined to one line. Since here the ground is defined only by the operation of the array, this architecture is referred to as a "virtual ground NOR array" [35]. This has the advantage of a very small cell size (similar to NAND), and also enables a symmetrical operation of the cell, which is essential in multi bit operation (see Section 11.4.2).

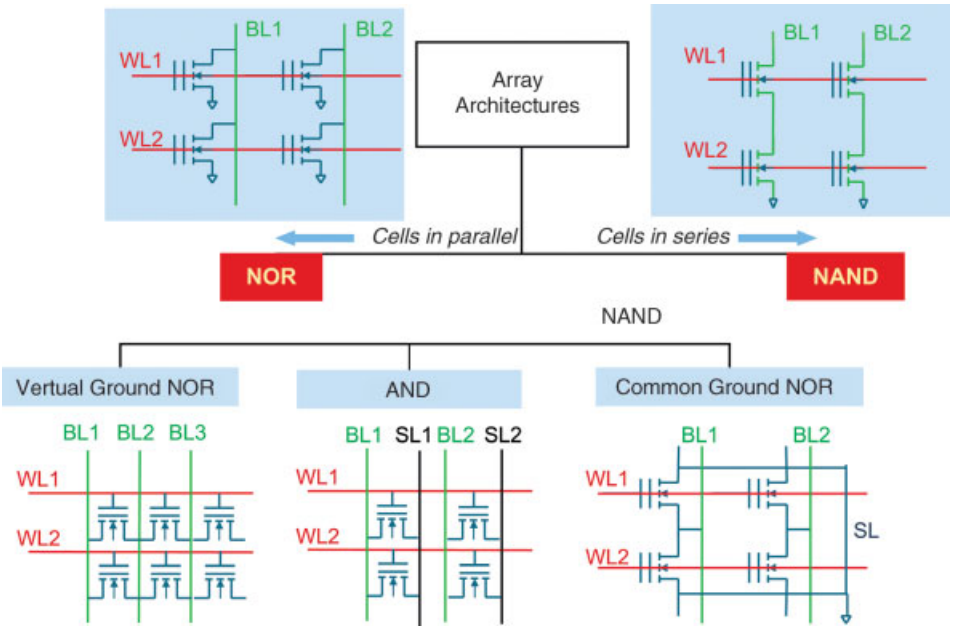


Figure 11.7 Architectures for Flash memory arrays. In principle, it is possible to connect the cells on one bitline in parallel leading to the NOR-type array and in series leading to the NAND-type array. For NOR-type arrays, many different variants have been proposed, whereas for NAND there is only one mainstream solution.

11.3 Floating-Gate Flash Concepts

11.3.1 The Floating-Gate Transistor

In essence, every flash cell is a metal oxide silicon (MOS) transistor with the charge storage layer placed in between the control gate and the channel. The drain current I_D of a MOS transistor can be expressed by the set of Equation 11.2 [36]

$$\begin{aligned} I_D &= \beta\left((V_G - V_T)V_D - \frac{V_D^2}{2}\right) && \text{for} \quad V_G - V_T > V_D > 0 \\ I_D &= \frac{\beta}{2}(V_G - V_T)^2 && \text{for} \quad 0 < V_G - V_T \leq V_D \\ I_D &= 0 && \text{for} \quad V_G - V_T < 0 \end{aligned} \tag{11.2}$$

where V_G is the gate voltage, V_D is the drain voltage, V_T is the threshold voltage, and β is the transconductance. β and V_T are given by Equation 11.3:

$$\beta = \frac{W}{L} C_{IS} \mu$$
$$V_T = \Phi_{MS} - 2\phi - \frac{Q_S}{C_{IS}} - \frac{Q_{IS}}{C_{IS}} \tag{11.3}$$

where W and L are the channel width and channel length of the device, C_{IS} is the total capacitance of the gate insulator, μ is the channel mobility, Φ_{MS} is the workfunction difference between the gate electrode and the channel, ϕ_B is the Fermi potential, Q_S is the charge in the depletion layer, and Q_{IS} is the total charge in the insulator normalized to the silicon/insulator interface. In principle, the stored charge in a non-volatile memory cell can be modeled by the insulator charge Q_{IS}. For a charge-trapping device (as introduced in Section 11.4) this holds true, without further modifications.

In the floating-gate device, however, the gate controlling the device is the floating gate (FG) where, from the outside world, only the control gate (CG) can be accessed. Therefore, the gate potential V_G in Equation 11.2 can only be controlled according to the capacitive coupling of the floating gate to the external terminals. The capacitive coupling of the floating gate to the external accessible terminals is illustrated schematically in Figure 11.8; from this figure, under the assumption that the bulk and source of the device are grounded, Equation 11.4 can readily be deduced giving the floating-gate voltage V_{FG} as a function of the gate and drain voltage.

$$V_{FG} = \alpha_G (V_{CG} + f V_D)$$
$$f = \frac{\alpha_D}{\alpha_G} = \frac{C_D}{C_C} \tag{11.4}$$

In Equation 11.4 α_G and α_D are the gate-coupling and the drain-coupling coefficients, which are a function of the respective capacitances according to Equation 11.5:

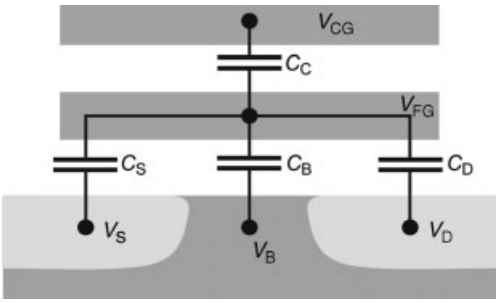


Figure 11.8 Schematic drawing of a floating-gate cell, illustrating the capacitive coupling of the floating gate to the external accessible terminals.

$$\begin{aligned} \text{gate coupling}: \quad \alpha_G &= \frac{C_C}{C_T} \\ \text{gate coupling}: \quad \alpha_D &= \frac{C_D}{C_T} \end{aligned} \tag{11.5}$$

$C_T = C_C + C_S + C_B + C_D$ is the total capacitance, where the individual capacitance are defined in Figure 11.8. By substituting Equation 11.4 into Equations 11.2 and 11.1 and rearranging, the threshold voltage with respect to the control gate is obtained

$$V_{T,CG} = \frac{V_{T,FG}}{\alpha_G} - \frac{\alpha_D}{\alpha_G} V_D - \frac{Q_{FG}}{C_C}. \tag{11.6}$$

The CG threshold voltage is a function of the floating-gate threshold voltage, the charge in the floating gate, and also a function of the applied drain voltage. This means that the threshold voltage is lowered when the drain voltage is increase or, in other words, the device can be turned on by the drain terminal. This effect must be considered when designing floating-gate memory arrays.

11.3.2 NOR Flash

Whilst, in Section 11.2.2 it was shown that a significant number of variants of different NOR-type concepts exist, this section will focus on the mainstream NOR concept. This uses a common ground NOR architecture (see Figure 11.7) in combination with the so-called EPROM tunnel oxide (ETOX) cell. The ETOX cell is, in principle, a stacked gate cell (see Figure 11.2, right and Figure 11.3a) that is programmed using channel hot electrons, and is erased using FN tunneling. In older generations the erasing was carried out towards the source terminal of the device, but as this calls for a large underdiffusion of the source junction under the gate in newer generations it was replaced by tunneling towards the channel.

The cell layout and cross-sections through a cell fabricated in 90-nm technology is illustrated in Figure 11.9. It can be seen that the wordline (WL) pitch is larger than the bitline (BL) pitch, as the channel length cannot be scaled to the minimum feature size using channel hot electron programming and the contact that must be placed in between two cells. The source line, in contrast, is fabricated self- aligned to the word line. The BL pitch on the other side is close to twice the minimum feature size, which is accomplished by self-aligning the bottom part of the floating gate to the shallow trench isolation (STI) and using an unlanded contact. Note that the top part of the floating gate is overlapping the STI, leading to a floating gate space below the minimum feature size. This, in combination with the portion of the control gate that is between the floating gate, provides more area to increase the gate coupling coefficient (see Figure 11.9).

The voltages applied to the terminals for programming, erasing and reading the cell are listed in Table 11.2. Programming is carried out using channel hot electron injection; optionally, the efficiency may be improved by applying a low bulk voltage.

Another important aspect that must be considered in the operation of every flash memory array are *disturbs*. These can cause an unwanted change of the cell content.

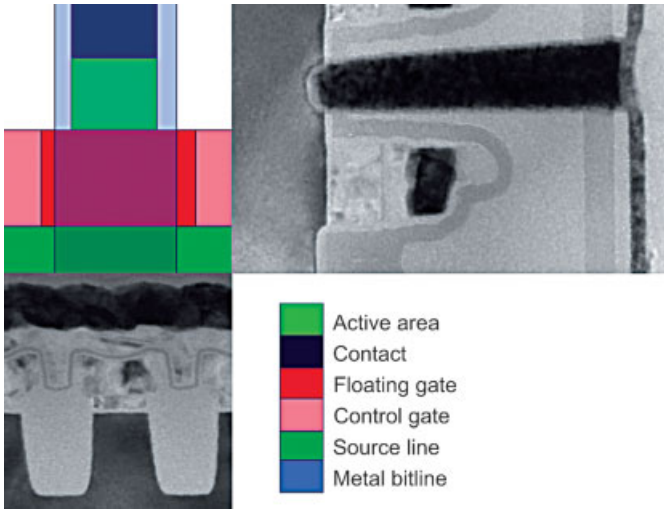


Figure 11.9 Layout and cross-sections of an ETOX-type Flash memory cell. The scanning electron microscopy cross-sections are taken from a 90-nm technology [37].

Table 11.2 Voltage conditions in read, program and erase for a typical ETOX-type cell.

	Gate	**Drain**	**Bulk**
Read	4 V	<1 V	0 V
Program	7 to 10 V	3 to 6 V	0 to −1.5 V (optional)
Erase	−6 to −8 V	float	6 to 8 V

A read disturb is caused by the fact that, for very low drain voltages, there is a probability of hot carriers being injected into the floating gate. Additionally, during read the applied gate voltage may cause FN tunneling into the floating gate. Both of these effects give rise to an unwanted programming of a cell during read. During programming, all cells connected to the same bitline are seeing a drain disturb, while all cells connected to the same wordline are seeing a gate disturb (see Figure 11.10). In

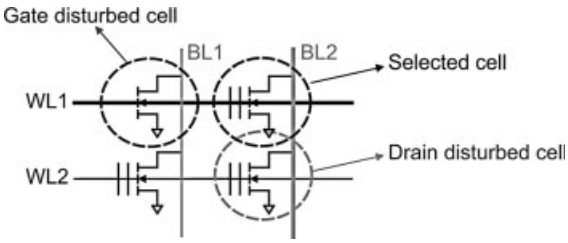


Figure 11.10 Gate and drain disturb in a common ground NOR-type Flash memory array.

erase, the disturb can be eliminated in flash-type memories by physically separating the erase sectors from each other.

11.3.3 NAND Flash

In NAND, only one concept exists, as illustrated in Figure 11.7. As with the NOR-type memory, the cell is a stacked gate but, due to the serial connection of cells, hot electron programming is not practical. The cells are therefore programmed and erased by FN tunneling between channel and floating gate. A schematic overview of a typical NAND string, including cell cross-sections taken from a 60-nm cell, is shown in Figure 11.11 [38]. Today, 32 cells are connected between two selects; this number has increased from eight cells via 16 cells, and may increase to 64 cells in the future [59]. The ground select connects the string to a sourceline, while the bitline select connects each string to a metal bitline.

In the wordline direction the cell resembles the ETOX cell described earlier. In the bitline direction, the cell is much denser due to a lack of contacts, as well as the channel length which can be scaled down to the minimum feature size due to the fact that the cell only has to isolate very low voltages. This becomes clear from Figure 11.12, where the main operations – read, write and erase – are explained in more detail. Erase is achieved simply by applying a sufficiently high voltage for FN tunneling to the well and applying 0 V to all the wordlines in the sector that must be erased, while leaving the wordlines of the non-erased sectors floating.

In the read operation a voltage higher than the highest V_T of a programmed cell must be applied to all the non-selected wordlines. In the erased state, the threshold voltage of the cells is chosen to be negative. Therefore, 0 V can be used on the wordlines that must be read. If a voltage of about 1 V is applied to the bitline, the selected cell will conduct if it is in the erased state and will be below threshold in the programmed state. For writing, a voltage high enough for FN tunneling (e.g., 15–20 V) is applied to the selected wordline. The channel of the selected cell is set to 0 V by applying 0 V to the selected bitline, turning the bitline selector on, and applying a high enough voltage to all the other wordlines; this will allow all other cells in the string to be turned on.

Care must be taken to avoid programming of the cells on the same wordlines. In early NAND Flash implementations a voltage of about 7 V was applied to the unselected bitlines and transferred to the channel of the cells on the selected wordline [39]. This, however, had two significant drawbacks. First, the junctions of the cells had to withstand the high voltage applied to the unselected bitlines, which restricts the scalability of the cell. Second, all the unselected bitlines had to be charged to a high voltage during programming. When the supply voltage was reduced from 5 to 3.3 V a new solution was implemented [40]. In that solution the bitline selects of the unselected wordlines are switched off by applying V_{DD} to the unselected bitlines. The channel of the inhibited cell will then raise its potential by capacitive coupling via the tunnel oxide capacitance C_{Tunnel} and the ONO capacitance C_{ONO} to the high voltage applied to the wordline. A high voltage applied to the passing wordlines will

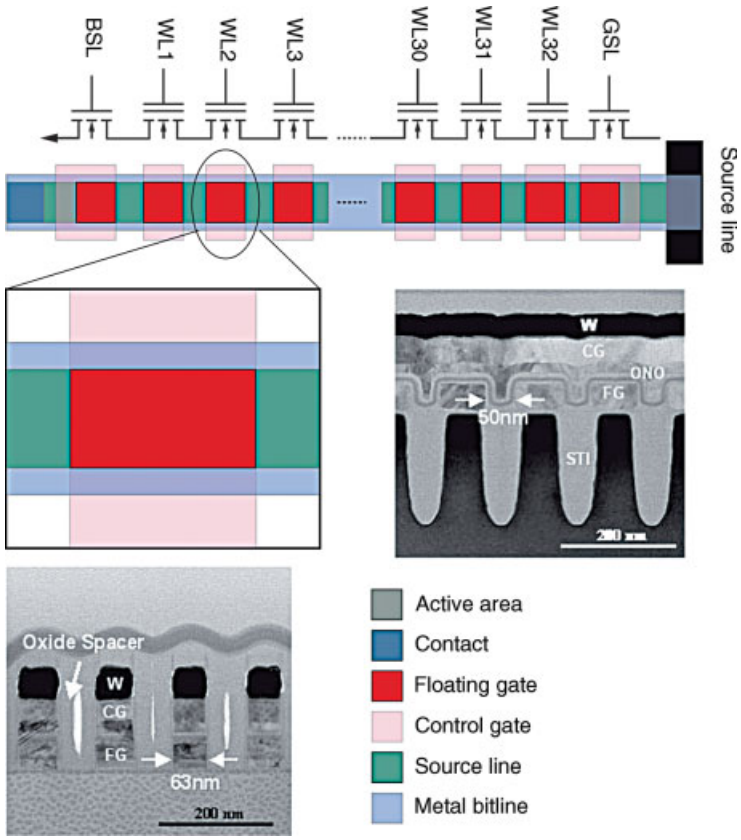


Figure 11.11 Layout and cross-sections of a string of NAND memory cells. In today's technologies, 32 cells are placed between a ground select and a bitline select. The ground select connects the string to a ground line, that in turn connects all strings to ground. The bitline select connects every string to a metal bitline via a bitline contact. The cross-sections are taken from a 60-nm technology [38].

further help to raise the channel potential in the disturbed cell. Now, the programming voltage and the voltage applied to the passing wordlines must be optimized to minimize the disturb effect. Examples for 120 and 90 nm Technology generations can be found in Ref. [41].

A large number of floating-gate concepts other than standard NOR or NAND have been proposed, and some of these are – or were – in production. Good overviews can be found in Refs. [42, 43]. Embedded flash memories have somewhat different requirements than standalone memories. Normally, much smaller memory densities are required than in standalone memories, and therefore the cell size is not as important but rather the size of the complete memory module including charge pumps, decoding, and so on must be minimized. This results in a high incentive to minimize the voltage requirements. Besides, every application focuses on different

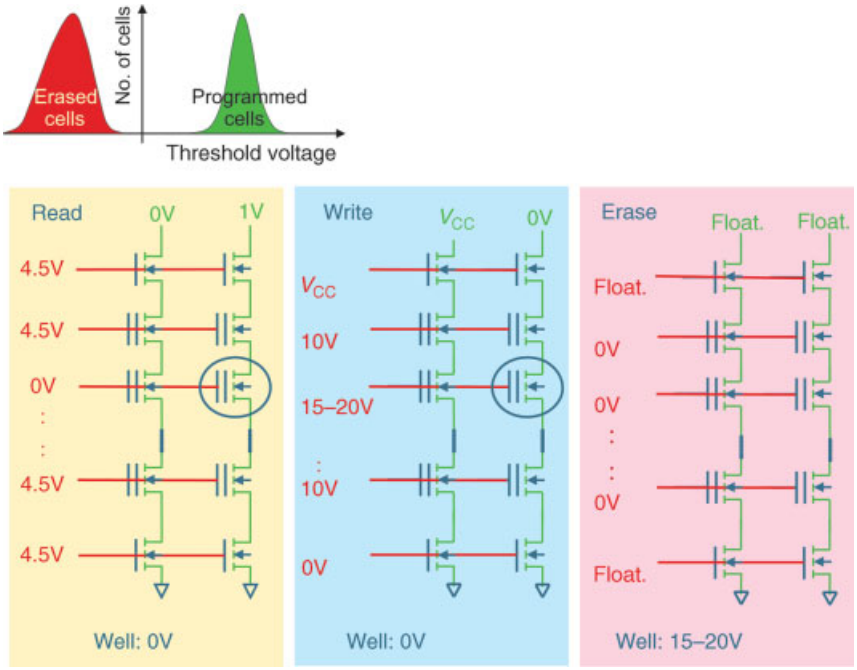


Figure 11.12 Read, write and erase operations of a NAND memory array.

requirements such as very fast random access, high endurance, high reliability, or low power, and therefore a large number of concepts coexist. In general, the ETOX is a good fit to many requirements, as long as the power consumption during programming can be tolerated. Among the large number of different concepts (some of these are referred to in Table 11.1) the field-enhancing tunneling injector cell [45, 47] is very popular and has been adopted by many foundries. Here, source side injection is used for the programming.

11.3.4 Reliability Aspects of Floating-Gate Flash

Every charge-based non-volatile memory inherently faces two reliability challenges. The first challenge is that the stored charge may be lost or charge may be gained during storage, leading to a loss of information; this phenomenon is referred to as *retention loss*. The second challenge is that the memory cell will degrade during repeated programming and erase cycles. Therefore, the endurance that a memory cell can achieve is a decisive quality criterion. Figure 11.13 illustrates how these basic properties manifest themselves in a 90 nm NAND Flash memory cell [46]. As program and erase cycling degrades the properties of the memory cell, the retention properties will also be affected by the precycling.

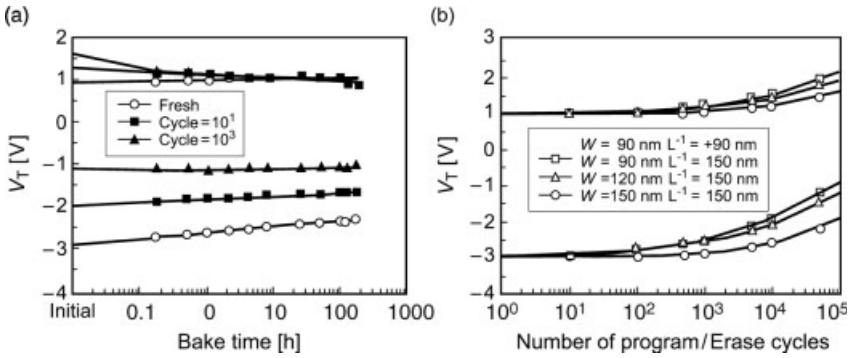


Figure 11.13 Retention (a) and endurance (b) characteristics of a 90-nm NAND Flash memory cell [49]. The dependence on precycling is shown for the retention characteristics, while the influence of device geometry is illustrated for the endurance curve.

In floating-gate memories, two specific effects must be mastered. The first is an effect that causes an abnormal erase behavior, leading occasionally to a much faster erase in some cells (see Figure 11.14a). As this effect is erratic in nature and affects single bits rather than the whole population, this phenomenon is called *erratic bit* [47]. The second phenomenon is also statistical in nature, in that during storage some cells may lose charge much faster than the main population. This effect occurs at lower temperatures, disappears at high storage temperatures, and becomes more pronounced after cycling (Figure 11.14b and c). This second effect is referred to as *anomalous SILC* or *moving bit effect*. Erratic bits are attributed to hole trapping in the bottom oxide. A small probability exists that clusters of three or more trapped holes exist. The overlap of the electrical field of the trapped holes leads to a strong increase in tunneling current [48], and thus to a much faster erase. The anomalous SILC or moving bit is explained by charge loss via neutral traps that are generated while cycling the cell. As a percolation path of such traps must exist for a charge loss to occur, the phenomenon is a strong function of oxide thickness, with thicker tunneling oxides showing much lower moving bit rates than their thinner counterparts [49].

11.3.5 Scaling of Floating-Gate Flash

The most severe issue in scaling down floating-gate devices is the non-scalability of the tunnel oxide and the inter poly dielectric. In order to achieve non-volatile retention, the tunneling dielectric must be at least 6 nm thick [50], although in practical memories thickness of 8–10 nm are used, based on the concrete reliability specification. This margin allows the covering of extrinsic effects such as moving bits. Scaling of the oxide-nitride-oxide (ONO) layer used as the interpoly dielectric is limited to an electrical effective thickness of about 13 nm due to retention and V_t stability constraints [51].

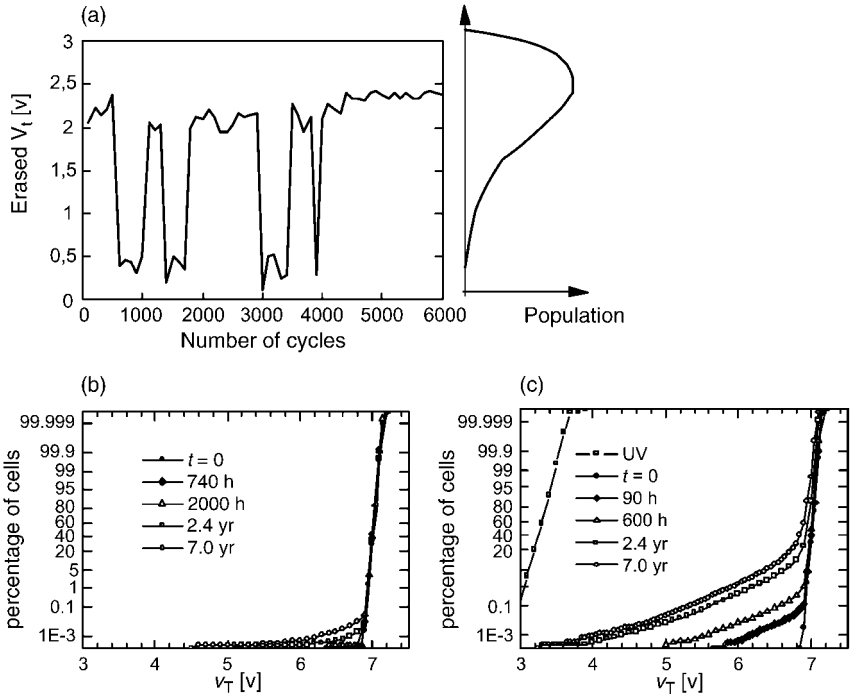


Figure 11.14 Specific reliability phenomena in floating gate memory cells [50]. Erratic bits (a) occasionally have a much faster erase than the majority of the population. Some bits show a pronounced low-temperature retention loss which is much higher after cycling (c: after 10 k cycles) than before cycling (b: after 10 cycles).

Further scaling of the tunneling oxide can only be obtained by radically re-engineering the tunnel barrier. Materials with a higher dielectric constant (e.g., HfO_2, ZrO_2) that are currently investigated for logic transistors can help. Crested barriers [52] could further improve the basic memory cell by increasing the ratio between the on and off current, leading to much faster write times as well as lower programming voltages. The principle of such an approach is shown schematically in Figure 11.15. The triangular shape of the barrier maintains the maximum barrier height if no voltage is applied (retention case), but drastically reduces the effective barrier in case of an applied voltage (programming or erase case). As a crested barrier is not achievable with those materials that have the required barrier heights, a staircase approximation using three layers with different band offsets, as well as different dielectric constants, is a reasonable approach. In the optimum structure the center layer would have a high band offset and a high dielectric constant, while the surrounding layer has lower band offset as well as a lower dielectric constant (Figure 11.15c). In most materials, however, a high band offset is correlated with a low dielectric constant and vice versa, making the optimum choice very difficult. A stack consisting of $Si_3N_4/Al_2O_3/Si_3N_4$ could be a reasonable and producible

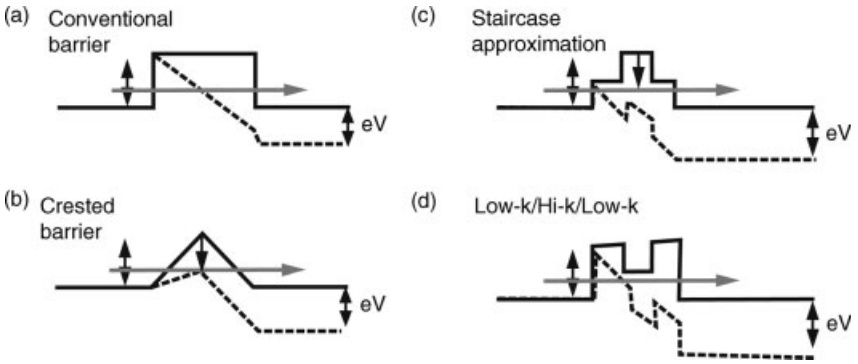


Figure 11.15 Different approaches to re-engineer the tunneling barrier of a floating-gate memory cell. (a) Conventional barrier; (b) ideal crested barrier; (c) staircase approximation of a crested barrier; (d) hi-k barrier sandwiched in between two low-k layers.

compromise [53]. An alternative – and perhaps better – manufacturing approach would be to sandwich a hi-k layer in between two low-k layers [54], and encouraging data have recently been demonstrated by using such an approach [55].

The non-scalability of the inter-poly dielectric thickness will eventually lead to a situation where the control gate will not fit into the space between two floating gates (Figure 11.16a). With the conventional ONO dielectrics this would lead to a situation where the control gate to floating gate coupling is significantly degraded [56]. To compensate for this, a high-k coupling dielectric is required, which has a k-value high enough to achieve the coupling only via the top of the floating gate (Figure 11.16b). Typical materials are very similar to those used for gate dielectrics, including HfO_2 and Hf/Al micro laminates [57].

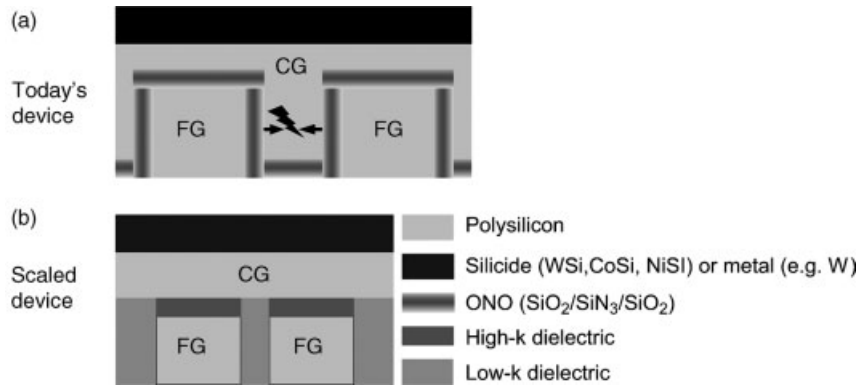


Figure 11.16 Possible scaling path for floating-gate memory cells. (a) Today's device architecture. (b) A scaled device using high-k inter-poly dielectrics as well as low-k dielectrics between floating gates. CG, control gate; FG, floating gate.

The reduced spacing between floating gates will also lead to higher capacitive coupling between floating gates, and result in severe crosstalk between cells [55]. This calls for a material with a lower dielectric constant between the floating gates, as shown in Figure 11.16b. This must also be implemented in the area between the word lines. Replacement of the silicon nitride spacer of the cell transistor with a silicon dioxide spacer (as shown in Ref. [58]) may help to significantly reduce the effect, but in the long term real low-k materials will be necessary. The recently demonstrated air gaps between the floating gates may represent an ultimate solution [59].

Although the scaling challenges presented so far are valid for all types of floating-gate flash memory devices, in the NOR-type architecture two more limitations must be considered [60]. First, a contact is required for every two cells, and this results in a significant area overhead. The overhead can be minimized by using a contact that is self-aligned to the control gate [61], or by using a virtual ground NOR array rather than a common ground NOR array [62]. The second limitation is that the channel length scaling is limited by the high voltages required during channel hot electron programming. However, a vertical device may be required to overcome this issue [63].

Instead of reducing the feature size, the cost reduction and density increase of a flash memory can also be achieved by increasing the number of bits stored on the same surface area, rather than reducing the size of a physical cell. As the charge storage is analogue in nature, more than two levels can be stored on one floating gate. To code n bits, 2^n levels are required. Figure 11.17 illustrates the corresponding V_T distributions for both NOR and NAND devices. For NOR devices, the multi-level approach was introduced to the market back in 1997 [64], and today in NAND devices

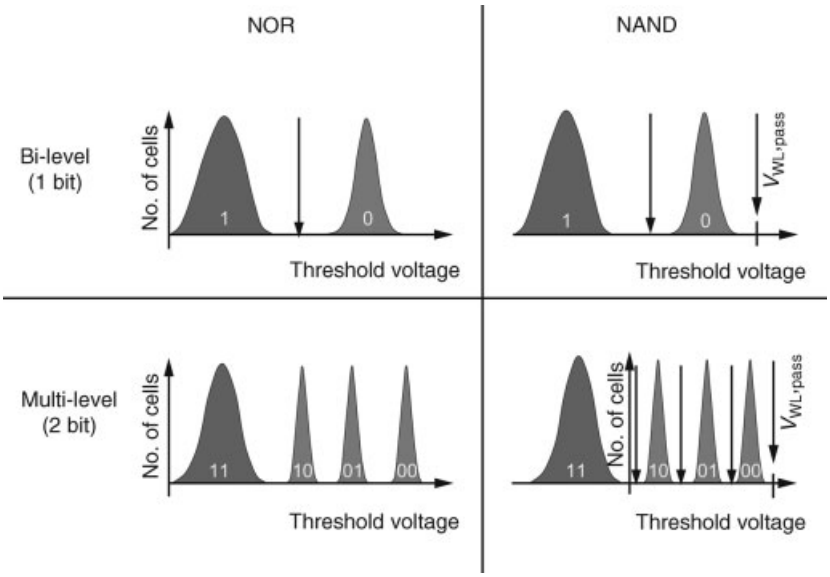


Figure 11.17 Multi level storage. The V_T distributions of a 1-bit and a 2-bit cell are compared for NOR as well as NAND Flash memories.

multi-level is also becoming increasingly popular as cost reduction is the most important aspect in data storage [65]. In the long term, the three-dimensional stacking of memory cells may also be an option; an example of this, namely the vertical integration of a NAND string, was demonstrated in Ref. [66].

11.4 Charge-Trapping Flash

Instead of using a floating gate to store the charge, an insulator with a high density of traps – a so-called *charge-trapping layer* – may also be used. Although this approach enjoyed some success during the 1970s and 1980s for EPROM and EEPROM memories [67, 68], in Flash memories the floating gate, in its ETOX and NAND versions, became dominant during the 1990s. The charge-trapping concept has some interesting advantages over floating-gate devices. As the charge is highly localized, no capacitive coupling effects (as described in Section 11.3.1) need to be considered. The cell can be described by Equations 11.2 and 11.3 with the storage charge included in Q_{IS}, which means that there is no drain turn on effect. The localization of charge also makes the cell less sensitive to any local defects responsible for erratic and moving bits in floating-gate cells. The coupling between cells is also much less pronounced. A few years ago the localization was utilized to store two physically separated bits in the same memory cell to create a multi-bit cell [69] (see Section 11.4.2), and this led to a revival of charge-trapping devices in the Flash world.

11.4.1 SONOS

Today, charge-trapping devices typically use a stack consisting of a polysilicon control gate electrode, a silicon dioxide topoxide, a silicon nitride storage layer, and a silicon dioxide bottom oxide placed on top of the active silicon. This stack (see Figure 11.18, left side) is referred to as a SONOS structure. Figure 11.18 also illustrates the programming and erase operation of such a structure using tunneling. During the programming and erasing operation, electrons or holes can be injected into or ejected out of the nitride storage layer. In programming, it is mainly electron injection from the silicon channel over the bottom oxide barrier into the nitride that takes place. Generally, for very long programming times and high V_T shifts, the hole injection from the control gate becomes significant and reduces any further programming. In the initial phase of the erase, the tunneling of electrons from traps via the nitride conduction band into the channel region is the dominant process, but this is increasingly being replaced by hole tunneling from the channel to the nitride as the erase progresses [70]. The erase process will result in less-negative charge in the trapping layer, leading to a reduction of the trapping layers potential. Therefore, the field over the bottom oxide is reduced and the field over the top oxide is increased, leading to an onset of electron tunneling from the gate to the trapping layer and a

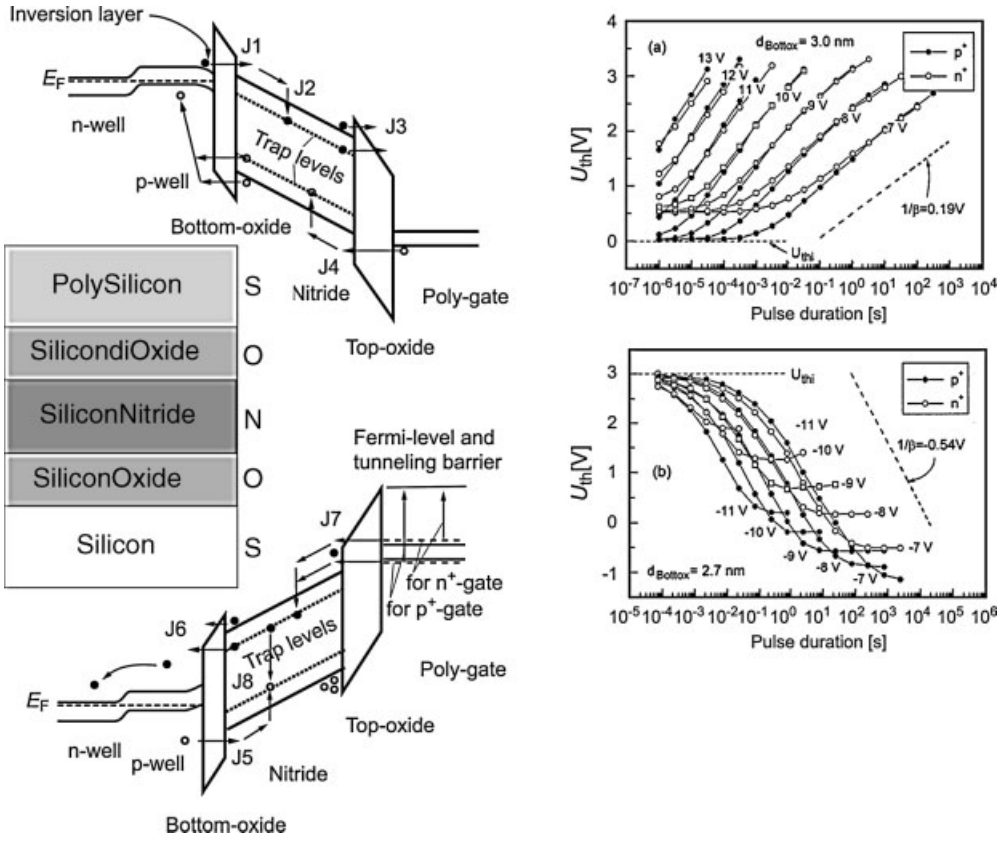


Figure 11.18 Programming (top row) and erase (bottom row) operation of a SONOS-type structure illustrated in the band diagram and a V_T over time characteristic [70]. During erase, the field over the bottom oxide will gradually decrease, while the field over the top oxide will increase; this will lead to a steady state and therefore to a saturation effect. The asymmetry of the two interfaces makes the saturation occur at higher V_T levels for higher gate voltages.

reduction of hole tunneling from the channel to the trapping layer. Finally, a steady state of the two processes – and therefore a saturation of the erase – will occur.

Due to the asymmetry of the two interfaces, the increase in erase voltage will lead to a strong increase in the injection from the gate, and hence to an even higher saturation level. This can be seen in the lower right-hand graph in Figure 11.18. In order to increase the erase speed, it is possible to increase the field over the bottom oxide, to reduce the field over the top oxide, or to increase the barrier for electrons at the control gate–top oxide interface. Although the simplest choice is to reduce the bottom oxide thickness, which will lead to a significantly increased field over the bottom oxide, it will also degrade the retention properties. For many years, therefore, the SONOS development was caught in the trade-off between bad retention and slow erase.

The increase in the barrier for electron injection from the gate electrode is very effective, as can be seen in Figure 11.18 (lower right), where n^+ doped gates and p^+ doped gates are compared. However, this alone is insufficient to achieve an erase performance that will be fast enough for a data Flash memory, yet still be secure after 10 years of retention. The field over the top oxide can be reduced by replacing the silicon dioxide with a high-k material such as Al_2O_3 [71]. Indeed, recently the combination of an Al_2O_3 topoxide and a high-workfunction TaN gate electrode was implemented in order to achieve NAND Flash-compatible performance on a 63 nm demonstrator using a bottom oxide as thick as 4 nm [72]. The structure used was referred to as a TANOS (Tantalum-nitride/Silicon-nitride/Silicon-dioxide/Silicon). The cross-section of the used stack, as well as the erase curves demonstrating a memory window of 5 V, are shown in Figure 11.19. This device represents a very promising candidate for NAND Flash memories with sub-40 nm ground rules.

11.4.2
Multi-Bit Charge Trapping

If the localized charge storage properties of a charge-trapping layer are combined with the localized injection by channel hot electrons, then the stored charge can be

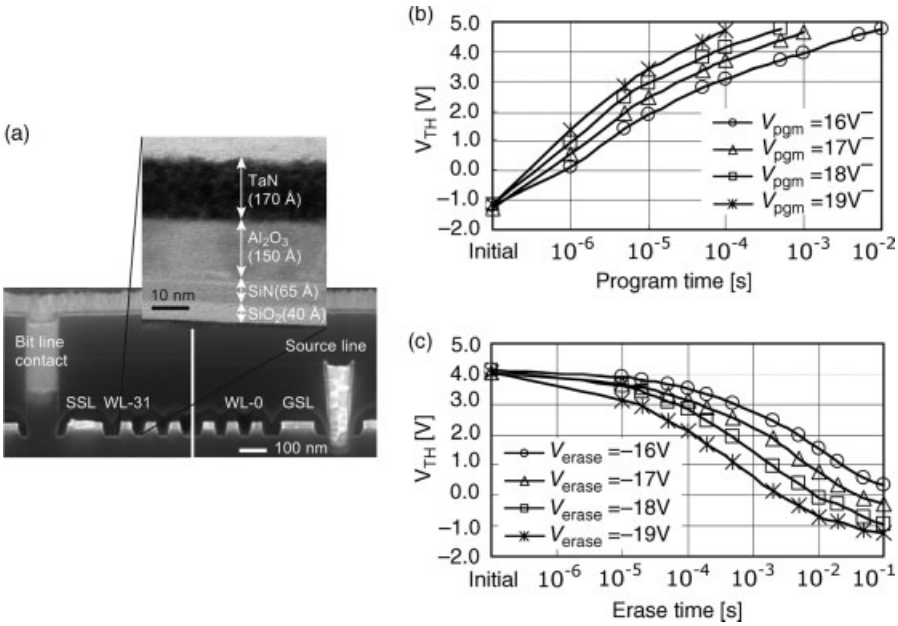


Figure 11.19 NAND Flash demonstrator using a TANOS cell on 63-nm ground rules [72]. The cross-section (a) shows the similarity to a floating-gate NAND. Program and erase performance is good enough to achieve 5 V memory window using a 4 nm-thick bottom oxide which ensures non-volatile retention.

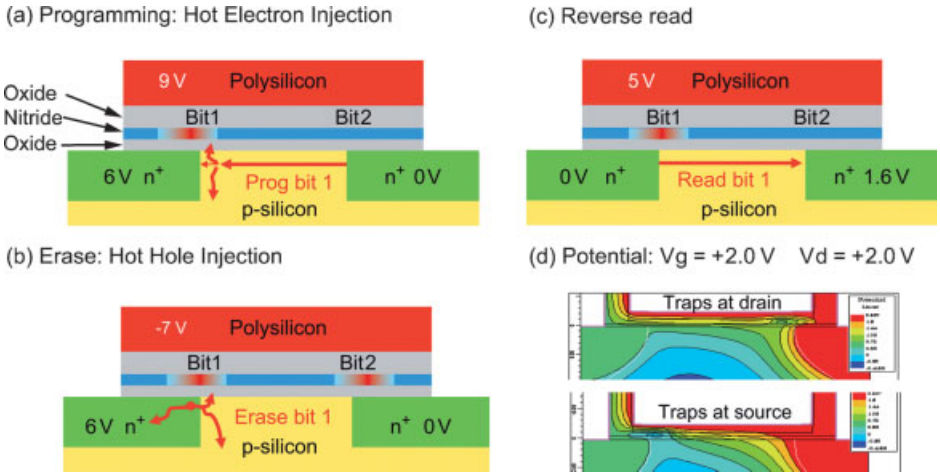


Figure 11.20 Multi-bit charge trapping memory cell. To inject charge self-aligned to the drain junction, programming is done by channel hot electron injection (a). Hot hole erase (b) enables the use of a thick bottom oxide. In reverse read (c), the charge of the second bit in the same cell is screened by applying a sufficiently high drain potential, while the role of drain and source are interchanged as compared to the programming conditions. The potential diagrams (d) show how the potential of the region below the charge is controlled by the drain potential rather than the stored charge if the drain potential is high. In (d), in red = high potential; blue = low potential.

placed in very narrow region which is self-aligned to the drain junction of the device. By interchanging the source and drain of such a device, two physically separated bits can be stored (see Figure 11.20a and Ref. [69]). In order to read the bits separately from one another, the source and drain must be interchanged compared to the programming conditions, and a drain voltage large enough to punch through the region below the charge above the drain region used in read must be applied (Figure 11.20c and d). As the charge is localized in a region close to the drain junction during programming, hot holes generated by band-to-band tunneling can be used for erase to compensate the stored charge (Figure 11.20b; see also Figure 11.7 and Ref. [13]). This allows for a sufficiently thick bottom oxide layer so as to avoid vertical charge loss and circumvent the erase saturation issue described in Section 11.4.1.

This device is implemented under different trade names such as NROM [13], MirrorBIT [73], NBit [74] and Twin Flash [75], both in code and data Flash products.

In order to operate the multi-bit charge-trapping cell described above, it is essential to have an architecture where no difference between bitlines and sourceline exists. The virtual ground NOR array (see Figure 11.7) therefore is the natural choice. This can be constructed by the structuring of a ONO layer, implanting the bitlines through the openings, growing an oxide for isolation, and finally forming the wordlines perpendicular to the bitlines. This method was used in the first-generation systems (see Figure 11.21a), but it has the drawback of a high thermal budget that must be applied to the bitline implants.

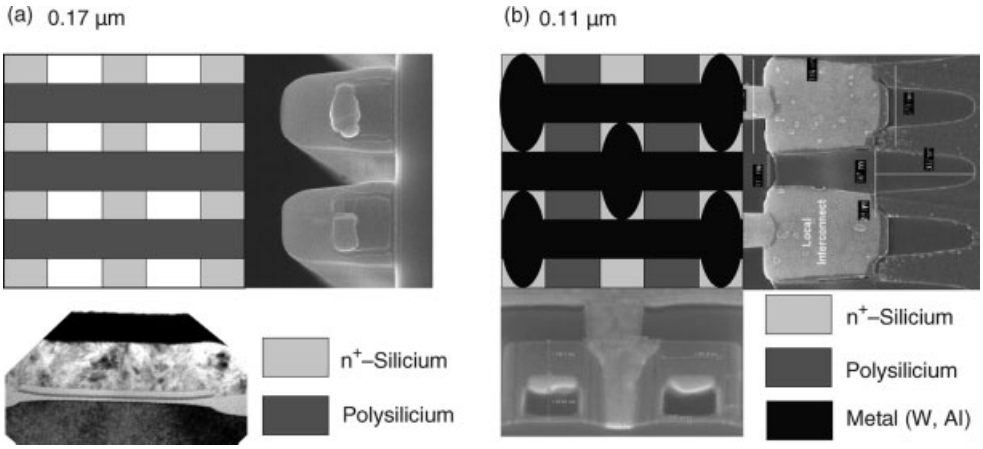


Figure 11.21 Multi-bit charge-trapping memory cells. In the 0.17 µm generation a buried bitline with crossing wordlines was used [79]. A more advanced version uses a device that resembles a standard MOS transistor [80, 81]. To end up with a virtual ground NOR architecture, a local interconnect must connect two neighboring devices over an isolation region. This local interconnect is then connected to the metal bitline.

Although the multi-bit charge-trapping type of memory cell has all the advantages described in Section 11.4.1, plus the inherent two bit per cell operation, there remain two challenges that must be considered. First, the unique mechanism of localized charge storage makes an understanding of the reliability-governing factors more complex. On an empirical basis, all effects that are necessary to create a reliable product are well understood and under control [76]. The physical basis for the observed results, however, remains the subject of debate among the scientific community. In principle, two effects may occur: First, the injection of hot holes during erase may damage the bottom oxide, leading to traps that can cause a vertical loss of the stored electrons [77]. Second, due to the fact that in programming and erase two localized mechanisms are used that will not be totally aligned to each other, a dipole will be created. This dipole may lead to lateral charge movement and therefore a change in V_T. In practice, however, both mechanisms may be involved, leading to a well-behaved and predictably reliable unit [78].

A number of modifications of the above-described multi-bit charge-trapping cell have been reported. For example, by adding an assist gate programming can be carried out using source side injection, which significantly reduces the cell current during programming. Examples of multi-bit charge-trapping cells using source side injection may be found in Refs. [82–84]. Another interesting variant is created by programming the cell with hot holes rather than hot electrons. In that case, the erase may be performed by FN tunneling either to the channel or to the gate [85, 86]. This concept, which is referred to as PHINES (programming by hot-hole injection nitride electron storage), has one main drawback in that programming of the second bit on

the same junction must be avoided. However, the same basic cell can also be used in a NAND-type architecture [87].

11.4.3
Scaling of Charge-Trapping Flash

When further scaling down the planar type of SONOS cell, the cell properties will suffer from low gate control, low read current, and a small number of electrons. In principle, the options of tailoring the tunneling barrier described for floating-gate device scaling can also be applied to the bottom oxide of a charge-trapping cell. Charge trapping, however, also enables another scaling path, by utilizing a FinFET device [88]. In principle, this could also be achieved with a floating-gate device, but two problems are encountered: First, the stack of tunnel oxide, floating gate and interpoly dielectric is too thick to fit the space between two neighboring cells; and second, the high coupling of the floating gate to the channel in a FinFET device will cause a deterioration in the gate coupling ratio [see Equation 11.2]. In a charge-trapping device the implementation of a FinFET device is straightforward; the general concept, as well as the excellent programming and erase curves that may be achieved with devices as short as 20 nm [89], are illustrated in Figure 11.22.

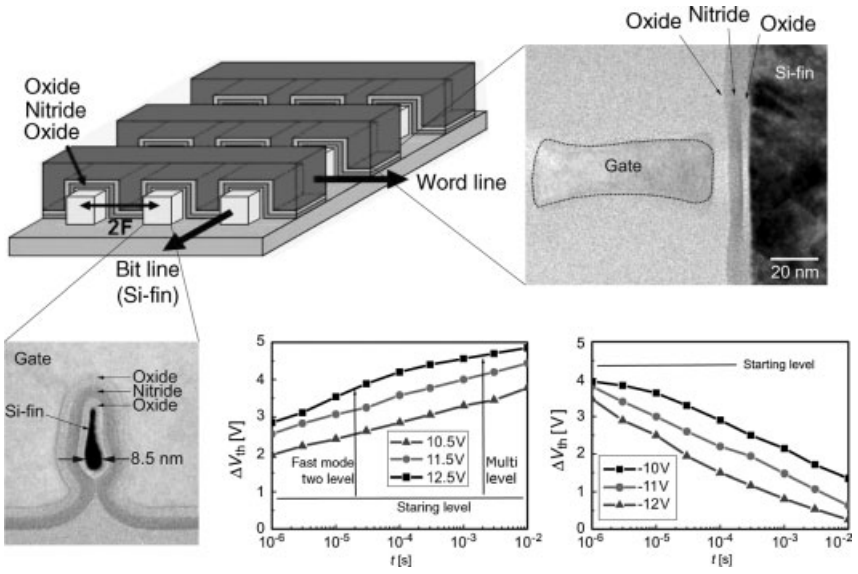


Figure 11.22 FinFET-based charge-trapping NAND [89]. The SEM images show cross-sections of a fabricated device with a channel length of 20 nm and a fin width below 10 nm. The programming and erase characteristics demonstrate a memory window of more than 4 V with fast program and erase.

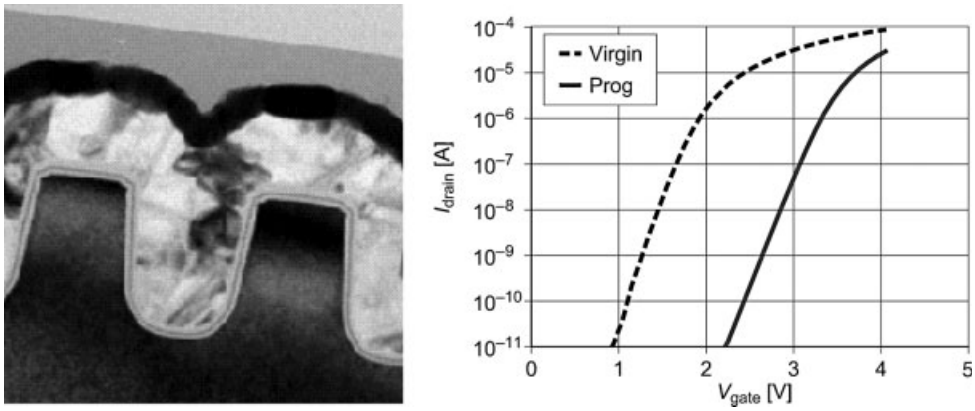


Figure 11.23 U-shaped multi-bit charge-trapping cell [93]. The U-shaped channel allows the surface area usage to be reduced, without any reduction of the effective channel length. The current–voltage (I–V) curves demonstrate the excellent separation of the two bits. Virgin = unprogrammed; Prog = programmed.

In order to further increase the memory density, a 3-D NAND-type memory would be very beneficial. Although the straightforward approach here would be to use thin-film transistors, charge trapping is again the favored solution, as it is much easier to integrate in a stacked manner compared to a floating-gate device. The first results of a thin-film transistor-based charge-trapping memory cell have recently been published by Walker and colleagues [90].

For the multi-bit charge-trapping memory cell it is essential to scale down the channel length, and indeed some excellent cell properties have been demonstrated down to 60 nm generation with the type of cell shown in Figure 11.21b [75, 81]. For further scaling down, a cell which uses structured nitride areas to store the charge was proposed. This has the advantage of controlling the cross-talk between bits, as well as being able to us thinner gate dielectrics between the ONO layers [91, 92]. A more radical approach that utilizes the third dimension by adopting a U-shaped device [93] is shown in Figure 11.23. Another approach to increase the storage density and reduce the area usage per bit is to combine the multi-bit concept with a multi-level approach. As a result, with four levels on each side of the cell (a total of eight levels), 4 bits can be stored [94], whereas by comparison a floating-gate cell requires 16 levels to store 4 bits (see Section 11.3.5).

11.5
Nanocrystal Flash Memories

In addition to floating gates and charge-trapping layers, nanocrystals are also currently under investigation for future use as flash devices [95]. A typical

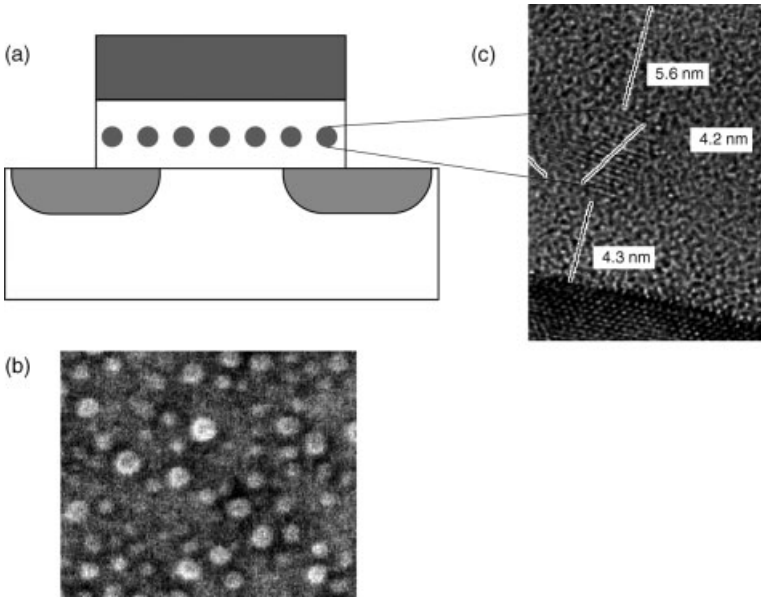


Figure 11.24 Silicon nanocrystal Flash cell. (a) Conducting nanocrystals are embedded into the gate dielectrics of a MOS device. The SEM images show (b) a top view and (c) a cross-section of a typical silicon nanocrystal layer [99].

nanocrystal device is illustrated schematically in Figure 11.24, whereby small conductive crystals with a diameter of about 2–5 nm are embedded into a silicon dioxide layer. Most of these investigations have been conducted on silicon nanocrystals, although germanium [96] or metal [97, 98] nanocrystals have also been studied. As with charge-trapping devices, nanocrystal devices are much more robust with respect to local defects in the bottom oxide layer. They also show the unwanted erase saturation that is observed in SONOS. However, in the case of silicon nanocrystals there are two identical interfaces/barriers between the control electrode and topoxide, as well as between the nanocrystal and bottom oxide. Hence, a higher voltage will not lead to an even higher erase saturation level, and the erase can be accelerated to the appropriate level by using a higher erase voltage [100].

Nanocrystal devices can be programmed and erased either by tunneling [101] or by hot electron/hot hole injection [102]. In the latter case, as with charge-trapping memories, a multi-bit device can be realized, whereas in the former version nanocrystals may be candidates for future NAND-type memories. Although the erase saturation is somewhat relaxed, the trade-off between fast program and erase and sufficient non-volatile retention is a key issue for nanocrystal devices which are programmed and erased using tunneling. One way to improve the retention is to use a self-aligned double stack of nanocrystals [103]; in this way, by utilizing the coulomb blockade effect and quantum confinement, the retention can be improved if

a small nanocrystal embedded into a thin oxide layer is placed below the larger nanocrystal which actually carries the stored charge.

In most cases, either a low-pressure chemical vapor deposition (LPCVD) from SiH_4 [104] or ion-implantation with subsequent thermal treatment [105] are used to fabricate nanocrystals. Although other techniques have shown promise [106], LPCVD and ion implantation are the easiest procedures for integration into a standard CMOS process. In both cases, the nanocrystal formation is a statistical process leading to controllability issues in scaled down devices [107]. Methods for controlled fabrication of nanocrystal size and distance by using templates or self-organization would, therefore, significantly improve the outcome [108]. When further scaling down the nanocrystal device, this path may in time lead to a single-electron memory [109].

11.6 Summary and Outlook

The trend towards mobile electronic devices has created – and continues to create – a rapidly increasing demand for non-volatile memories. Today, Flash memories represent the best solution for most of these applications, where coded Flash applications are typically covered by NOR Flash devices and data Flash applications by NAND Flash devices. Currently, the floating-gate transistor is seen as the "workhorse" of those cell devices used in many of today's technologies. Indeed, floating-gate technology shows a scaling potential for further generations if innovations such as high-k coupling dielectrics or low-k isolation oxides can be mastered. By contrast, charge-trapping devices are possibly due to make a return, with multi-bit charge-trapping having recently emerged in a number of applications. In fact, a modified version of the classical SONOS device, programmed and erased by tunneling, may replace the floating-gate transistor in future generations of NAND Flash. Nanocrystals represent another option to replace the floating gate, although at present the challenges that they face seem much more severe than for the charge-trapping case. However, in the long term this development may lead to a single-electron device.

Unfortunately, flash-type memories based on charge storage in either floating gates or charge-trapping layers still suffer from important drawbacks, including limited endurance, slow write/erase, and no direct overwrite. Hence, for many years research groups have sought new storage mechanisms that could supply a non-volatile memory without such shortcomings. To achieve this goal, new materials with innovative switching effects must be integrated into the CMOS flow [110, 111]. Although these technologies are beyond the scope of this chapter, it is important to note that although they may have distinct advantages over Flash memories, and indeed some have now reached the production stage (see Chapters 13–16), the scaling of Flash memories has to date been much more successful. Such scaling possibilities provide Flash with a major competitive advantage in terms of system cost.

References

1 Niebel, A. (2004) *Proceedings of the 20th Nonvolatile Semiconductor Memory Workshop*, p. 14.

2 Harari, E., Schmitz, L., Troutman, B. and Wang, S. (1978) *ISSCC Digest of Technical Papers*, 108.

3 Mikolajick, T. *et al.* (2001) *Microelectronics Reliability*, **7**, 947.

4 Yoshikawa, K. (1999) *VLSI Symposium on Technology, Systems and Applications*, p. 183.

5 Kahng, D. and Sze, S.M. (1967) *BELL System Technical Journal*, **46**, 1288.

6 Wegener, H.A.R. *et al.* (1967) *IEDM Digest of Technical Papers*, 70.

7 Frohman-Bentchkowsky, D. (1971) *ISSCC Digest of Technical Papers*, 80.

8 Cricchi, J.R., Blaha, F.C. and Fitzpatrick, M.D. (1974) *IEDM Digest of Technical Papers*, 204.

9 Johnson, W. *et al.* (1980) *ISSCC Digest of Technical Papers*, 152.

10 Masuoka, F. *et al.* (1984) *IEDM Digest of Technical Papers*, 464.

11 Kynett, V.N. *et al.* (1988) *ISSCC Digest of Technical Papers*, 132.

12 Shirota, R. *et al.* (1988) *Symposium on VLSI Technology*, 33.

13 Chan, T.Y., Young, K.K. and Hu, C. (1987) *IEEE Electron Device Letters*, **8**, 93.

14 Eitan, B. *et al.* (1999) *Proceedings SSDM*, 522.

15 Keeney, S. (2001) *IEDM Digest of Technical Papers*, 41.

16 Onoda, H. *et al.* (1992) *IEDM Digest of Technical Papers*, 599.

17 Hisamune, Y.S. *et al.* (1993) *IEDM Digest of Technical Papers*, 19.

18 Sze, S.M. (1981) *Physics of Semiconductor Devices*, John Wiley & Sons, New York. p. 397.

19 Bude, J.D. *et al.* (1997) *IEDM Digest of Technical Papers*, 279.

20 Mahapatra, S., Shukuri, S. and, Bude, J.D. (2002) *IEEE Transactions on Electron Devices*, **7**, 1296.

21 Van Houdt, J. *et al.* (1992) *IEEE Transactions on Electron Devices*, **39**, 1150.

22 Ingrosso, G. *et al.* (2002) *European Solid-State Device Research Conference*, 187.

23 Tam, K., Ko, P.K. and Hu, C. (1984) *IEEE Transactions on Electron Devices*, **31**, 1116.

24 Mikolajick, T. *et al.* (2004) *Proceedings of the Nonvolatile Semiconductor Memory Workshop*, p. 98.

25 Hagenbeck, R. *et al.* (2004) *Journal of Computational Electronics*, **3**, 239.

26 Meinerzhagen, B. (1988) *IEDM Digest of Technical Papers*, 504.

27 Wolf, S. (1995) *Silicon Processing for the VLSI Era. Volume 3: The Submicron MOSFET*, Latice Press, Sunset Beach, p. 198.

28 Sim, J.S. *et al.* (2005) *Symposium on VLSI Technology*, 122.

29 Lenzliner, M. and Snow, E.H. (1969) *Journal of Applied Physics*, **40**, 278.

30 Kianian, S. *et al.* (1994) *Symposium on VLSI Technology*, 71.

31 Mehrotra, *et al.* (1992) *Symposium on VLSI Circuits*, 24.

32 Kume, H. *et al.* (1992) *IEDM Digest of Technical Papers*, 991.

33 Libsch, F.R. and White, M.H. (1990) *Solid State Electronics*, **33**, 105.

34 Peters, C. *et al.* (2004) *Proceedings of the 20th Nonvolatile Semiconductor Memory Workshop*, p. 55.

35 Ohi, M. *et al.* (1993) *Symposium on VLSI Technology*, 57.

36 Wolf, S. (1995) *Silicon Processing for the VLSI Era. Volume 3: The Submicron MOSFET*, Latice Press, Sunset Beach, p. 134.

37 Song, Y. *et al.* (2003) *Symposium on VLSI Technology*, 91.

38 Park, J.H. (2004) *IEDM Digest of Technical Papers*, 873.

39 Yoshihisa, I. *et al.* (1995) *IEEE Journal of Solid State Circuits*, **30**, 1157.

40 Suh, K.-D. *et al.* (1995) *IEEE Journal of Solid State Circuits*, **30**, 1149.

41 Kim, D.-C. *et al.* (2002) *IEDM Digest of Technical Papers*, 919.

42 Paven, P. *et al.* (1997) *Proceedings of the IEEE*, **85**, 1248.

43 Eitan, B. and Roy, A. (1999) *Flash Memories* (eds P. Cappelletti, C. Golla, P. Olivo, and E. Zanoni), Kluwer, Boston, p. 91.

44 Kianian, S. *et al.* (1994) *Symposium on VLSI Technology*, 71.

45 Kotov, A. *et al.* (2002) *Proceedings of the NVMTS*, 110.

46 Lee, J.-D. *et al.* (2003) *IEEE International Reliability Physics Symposium*, 497.

47 Cappelletti, P., Bez, R., Modelli, A. and Visconti, A. (2004) *IEDM Digest of Technical Papers*, 489.

48 Cappelletti, P., Bez, R., Cantarelli, D. and Fratin, L. (1994) *IEDM Digest of Technical Papers*, 291.

49 Degraeve, R. *et al.* (2004) *IEEE Transactions on Electron Devices*, **51**, 1392.

50 Lai, S. (1998) *Proceedings of the Seventh Biennial IEEE International Nonvolatile Memory Technology Conference*, p. 6.

51 Mori, S. *et al.* (1991) *IEEE Transactions on Electron Devices*, **38**, 386.

52 Likharev, K. (1998) *Applied Physics Letters*, **73**, 2137.

53 Casperson, J. (2002) *Journal of Applied Physics*, **92**, 261.

54 Specht, M., Staedle, M. and Hofmann, F. (2002) *European Solid-State Device Research Conference*, 599.

55 Liu, R. *et al.* (2005) *IEDM Digest of Technical Papers*, 22.3.1.

56 Kim, K. (2006) *Proceedings of the 21st Nonvolatile Semiconductor Memory Workshop*, p. 9.

57 Lee, W.-H. (1997) *VLSI Technology Digest of Technical Papers*, 117.

58 Lee, J.-D. (2002) *IEEE Electron Device Letters*, **23**, 264.

59 Kang, D. *et al.* (2006) *Proceedings of the 21st Nonvolatile Semiconductor Memory Workshop*, p. 36.

60 Atwood, G. (2004) *IEEE Transactions on Device and Material Reliability*, **4**, 3001.

61 Watanabe, H. *et al.* (1998) *IEDM Digest of Technical Papers*, 975.

62 Koval, R. (2005) *Symposium on VLSI Technology Digest of Technical Papers*, 2004.

63 Pein, H. and Plummer, J.D. (1993) *IEEE Electron Device Letters*, **14**, 415.

64 Atwood, G. *et al.* (1997) *INTEL Technology Journal*, **1**, 1.

65 Byeon, D.S. (2005) *ISSCC Digest of Technical Papers*, 46.

66 Endoh, T. *et al.* (2003) *IEEE Transactions on Electron Devices*, **50**, 945.

67 Libsch, F.R. and White, M.H. (1998) *Nonvolatile Semiconductor Memory Technology* (ed. W.D. Brown and J.E. Brewer), IEEE Press, New York. p. 309.

68 Jones, F. (1983) *Wescon Conference Record*, **27**, 28.1.1.

69 Eitan, B. *et al.* (2000) *IEEE Electron Device Letters*, **21**, 543.

70 Bachhofer, H. *et al.* (2001) *Journal of Applied Physics*, **89**, 2791.

71 Specht, M. *et al.* (2003) *European Solid-State Device Research Conference*, 155.

72 Shin, Y. *et al.* (2005) *IEDM Digest of Technical Papers*, 13.6.1.

73 van Buskirk, M. (2006) *Proceedings of the 21st Nonvolatile Semiconductor Memory Workshop*, p. 8.

74 Zous, N.K. *et al.* (2004) *IEEE Electron Device Letters*, **25**, 649.

75 Stein, E. *et al.* (2005) *Proceedings of the NVMTS*, 5.

76 Janai, M. (2003) *Proceedings of the IRPS*, 502.

77 Tsai, W.J. *et al.* (2002) *Proceedings of the IRPS*, 34.

78 Janai, M. *et al.* (2004) *IEEE Transactions on Device Materials Reliability*, **4**, 404.

79 Maayan, E. *et al.* (2002) *ISSCC Digest of Technical Papers*, 100.

80 Willer, J. *et al.* (2004) *Symposium on VLSI Technology*, 76.

81 Nagel, N. *et al.* (2005) *Symposium on VLSI Technology*, 120.

82 Hayashi, Y. *et al.* (2000) *Symposium on VLSI Technology*, 122.

83 Ogura, T. *et al.* (2003) *Symposium on VLSI Technology,* 207.

84 Tomiye, H. *et al.* (2002) *Symposium on VLSI Technology,* 206.

85 Yeh, C.C. *et al.* (2002) *IEDM Digest of Technical Papers,* 931.

86 Yeh, C.C. *et al.* (2005) *IEEE Transactions on Electron Devices,* **52**, 541.

87 Yeh, C.C. *et al.* (2006) *Proceedings of the 21st Nonvolatile Semiconductor Memory Workshop,* p. 76.

88 Hisamoto, D. *et al.* (2000) *IEEE Transactions on Electron Devices,* **47**, 2320.

89 Specht, M. *et al.* (2004) *IEDM Digest of Technical Papers,* 1083.

90 Walker, A.J. *et al.* (2003) *Symposium on VLSI Technology,* 29.

91 Lee, Y.K. (2004) *Proceedings of the 20th Nonvolatile Semiconductor Memory Workshop,* p. 96.

92 Choi, B.Y. (2006) *Proceedings of the 21st Nonvolatile Semiconductor Memory Workshop,* p. 72.

93 Willer, J. *et al.* (2003) *Proceedings of the 19th Nonvolatile Semiconductor Memory Workshop,* p. 42.

94 Eitan, B. (2005) *IEDM Digest of Technical Papers,* 22.1.1.

95 Tiwari, S. (1996) *Applied Physics Letters,* **68**, 1377.

96 Bostedt, C. *et al.* (2004) *Applied Physics Letters,* **84**, 4056.

97 Tseng, J.Y. *et al.* (2004) *Applied Physics Letters,* **85**, 2595.

98 Samanta, S.K. *et al.* (2005) *Applied Physics Letters,* **87**, 113110.

99 Muralidhar, R. *et al.* (2003) *IEDM Digest of Technical Papers,* 26.1.1.

100 Sadd, M. *et al.* (2004) *Proceedings of the 20th Nonvolatile Semiconductor Memory Workshop,* p. 75.

101 Compagnioni, C.M. *et al.* (2005) *IEEE Transactions on Electron Devices,* **52**, 569.

102 Perniola, L. *et al.* (2005) *IEEE Transactions on Nanotechnology,* **4**, 360.

103 Ohba, R. *et al.* (2002) *IEEE Transactions on Electron Devices,* **59**, 1392.

104 Gerardi, C. *et al.* (2004) *IEEE International Conference on Integrated Circuit design and Technology,* p. 37.

105 Borani, J.v. *et al.* (2002) *Solid State Electronics,* **46**, 1729.

106 Zacharias, M. *et al.* (2002) *Applied Physics Letters,* **80**, 661.

107 Perniola, L. *et al.* (2003) *Solid State Electronics,* **47**, 1637.

108 Guarini, K.W. *et al.* (2003) *IEDM Digest of Technical Papers,* 22.2.1.

109 Kim, I. *et al.* (1999) *IEEE Electron Device Letters,* **20**, 630.

110 Mikolajick, T. and Pinnow, C.U. (2002) *Proceedings of the NVMTS,* 3.

111 Pinnow, C.-U. and Mikolajick, T. (2004) *Journal of the Electrochemical Society,* **151**, K12.

12
Dynamic Random Access Memory

Fumio Horiguchi

12.1
DRAM Basic Operation

Dynamic random access memories (DRAMs) use the charge stored in a capacitor to represent binary digital data values. They are called "dynamic" because the stored charge leaks away after several seconds, even with power continuously applied. Therefore, the cells must be read and refreshed at periodic intervals. Despite this complex operating principle, their advantages of small cell size and high density have made DRAMs the most widely used semiconductor memories in commercial applications. In 1970, the three-transistor cell used for the 1 kbit DRAM was first reported [1], and the one-transistor (1T-1C) cell became standard use in 4 kbit DRAMs [2]. During the following years, the density of DRAMs increased exponentially, with rapid improvement to the cell design, its supporting circuit technologies, and fine patterning techniques.

The equivalent circuit of the 1T-1C DRAM cell is shown in Figure 12.1. The array transistor acts as a switch and is addressed by the word line (WL), which controls the gate. The storage capacitor, C_S, represents the charge storage element containing the information and is connected to the bit line, BL, via the array transistor. When the array transistor switch is closed, the voltage level $+V_{DD}/2$ or $-V_{DD}/2$ is applied to C_S via the bit line. The corresponding charge on C_S represents the binary information, "1" or "0". After this write pulse, the capacitor is disconnected by opening the array transistor switch.

The memory state is read by turning on the array transistor and sensing the charge on the capacitor via the bit line, which is precharged to $V_{DD}/2$ (where V_{DD} is the power supply voltage). The cell charge is redistributed between the cell capacitance, C_S, and the bit line capacitance, C_B, leading to a voltage change in the bit line. This voltage change is detected by the sense amplifier in the bit line and amplified to drive the input/output lines. Because a read pulse destroys the charge state of the capacitor, it must be followed by a rewrite pulse to maintain the stored information. The plate, PL, is kept at $V_{DD}/2$ to reduce the electric voltage stress on the capacitor dielectric, which is charged to $+V_{DD}/2$ or $-V_{DD}/2$ instead of being discharged to 0 V and charged to the full power supply voltage, V_{DD}.

Nanotechnology. Volume 3: Information Technology. Edited by Rainer Waser

ISBN: 978-3-527-31738-7

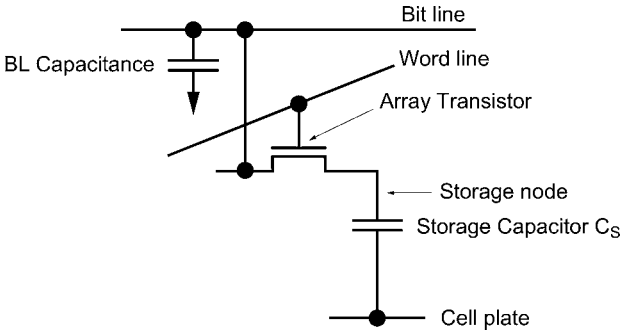


Figure 12.1 DRAM memory cell equivalent circuit. See text for details.

DRAM has required almost constant storage capacitance (more than 40 fF, i.e., 4×10^{-14} F) among the generations, despite the scaling to smaller cell sizes, and this is the reason for the requirement of three-dimensional (3-D) structures such as trench or stacked capacitor. A trench capacitor uses the inner surface of a Si hole to store charge, while a stacked capacitor uses a poly Si capacitor above the array transistor and bit line (see Figure 12.2).

12.2 Advanced DRAM Technology Requirements

Historically, the cost of DRAM has been forced to decrease to retain its share in the huge and competitive market for high-density memory. This has resulted in a

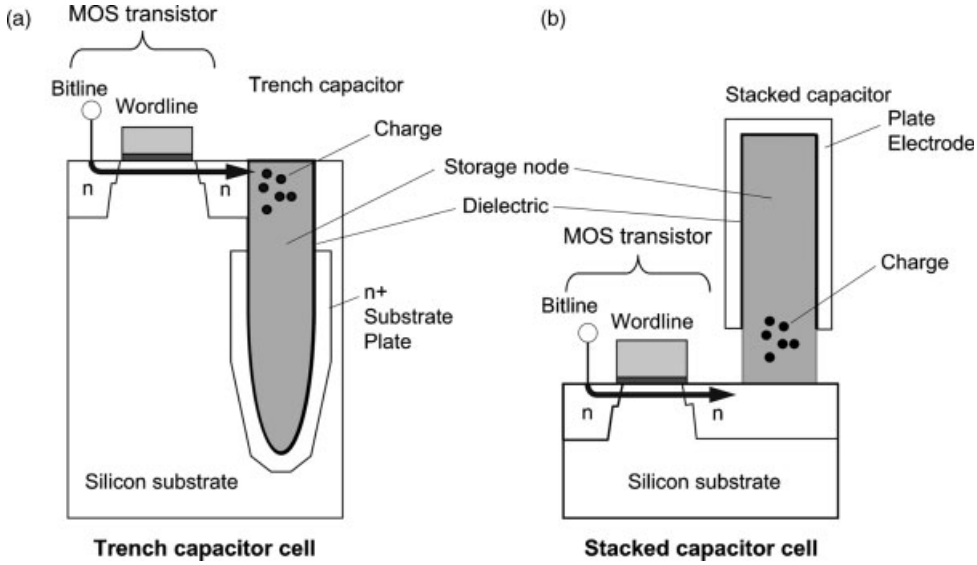


Figure 12.2 Conventional DRAM cells.

Table 12.1 Technology requirements for a DRAM memory cell.

Cell area	Smaller cell area $<8\,F^2$
Storage capacitor	>40 fF, low leakage $<1\times10^{-16}$ A
Array transistor	High drivability; low leakage current ($<1\times10^{-16}$ A)
Bit line contact	Self-aligned to the word line
Storage node contact	Self-aligned to the word line (and trench capacitor or bit line)

decrease in DRAM cell size because the chip cost is directly related to the cell area. Thus, every part of the cell is required to be as small as possible, and the cell size must be less than or equal to $8\,F^2$ (where F is the feature size). A summary of the technological requirements for a DRAM memory cell is provided in Table 12.1. The most critical part in the cell shrinkage is the capacitor; thus, a 3-D structure such as a trench or stacked capacitor has been adopted to retain sufficient capacitor area within a limited space.

As the DRAM cell size shrinks to sub-100 nm, it becomes critically important to realize a sufficient on-off-current ratio in the array transistor. In general, the scaling approach implies that the transistor sizes L and W, the gate oxide thickness T_{ox}, the supply voltage, and the threshold voltage V_{th} should be reduced by a factor of $1/k$ (k: scale factor), and that channel doping should be increased by a factor of k in order to sustain or improve the transistor performance. In a DRAM cell, the charge must be stored in storage capacitors; therefore, an extremely low off-current in the array transistor is required for data retention. Thus, V_{th} should be made as small as possible to decrease the channel leakage current, and the supply voltage should be minimized so that sufficient charge can be written into the capacitor. Gate oxide thickness must also be reduced to maintain a sufficient breakdown voltage for the gate dielectrics. All this means that the array transistor cannot be scaled down in a conventional manner. On the other hand, a sufficient on-current in the array transistor is required for fast writing characteristics. This suggests a short L and large W, but scaling difficulties prevent the reduction of L and the cell size limits W. Thus, maintaining sufficient on-current in the array transistor is difficult and, moreover, increasing channel doping degrades the channel mobility, which further decreases the on-current in the array transistor.

In view of these considerations, a different structural approach for array transistors in DRAMs is necessary.

12.3 Capacitor Technologies

The first generation of DRAMs used a planar storage capacitor for memories of up to 1 Mbit. However, from the 4 Mbit generation onwards, trench or stacked capacitors were used for maintaining the same storage capacitance within a limited cell area. A comparison between the stacked capacitor cell and the trench capacitor cell is shown

Table 12.2 Stacked versus trench cells.

	Stacked	Trench
Complexity of capacitor formation	×	×
Decrement of memory cell parasitics	○	△
Shrinkability of memory cell	△	○
Compatibility with logic process integration	△	○
Compatibility with logic device characteristics	○	◎
Compatibility with logic layout design rules	△	◎
Additional mask steps: @130 nm node		6–9

in Table 12.2. The differences arise mainly from the transistor formation process compared with the capacitor formation process. The array transistor in a stacked capacitor is formed before the capacitor; thus, the transistor source/drain junctions can easily be extended after the stacked capacitor is fabricated by a thermal process, which results in a degradation in transistor performance. Figures 12.2 to 12.4 show

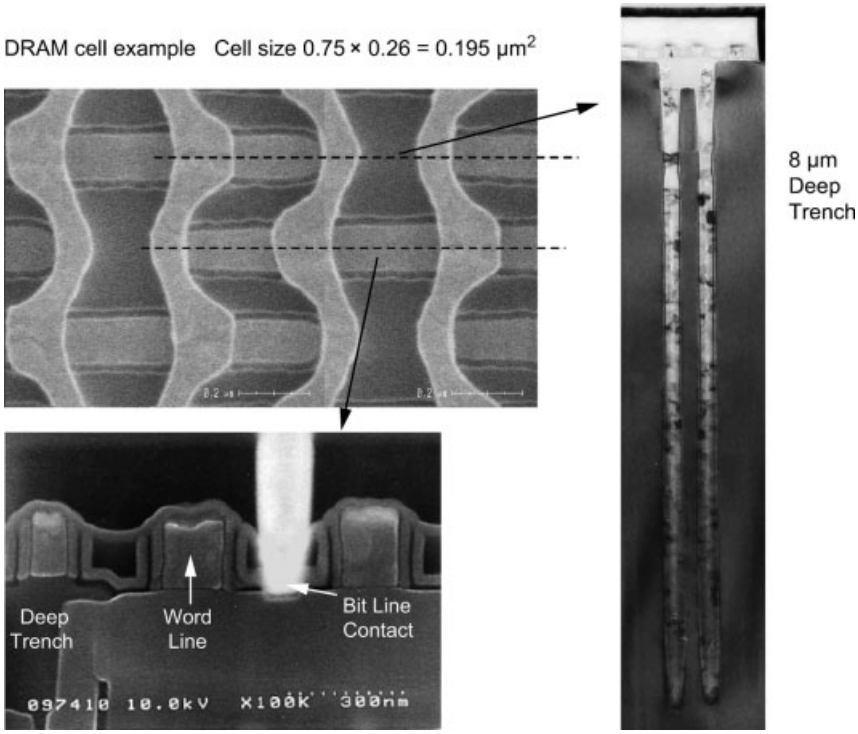


Figure 12.3 Trench capacitor cell. This has a 8 μm-deep trench capacitor and a sidewall storage-node contact for enough storage capacitance and small contact within a small area. (After Yamada, VLSI Tech. Short Course 2003 and Ref. [3]).

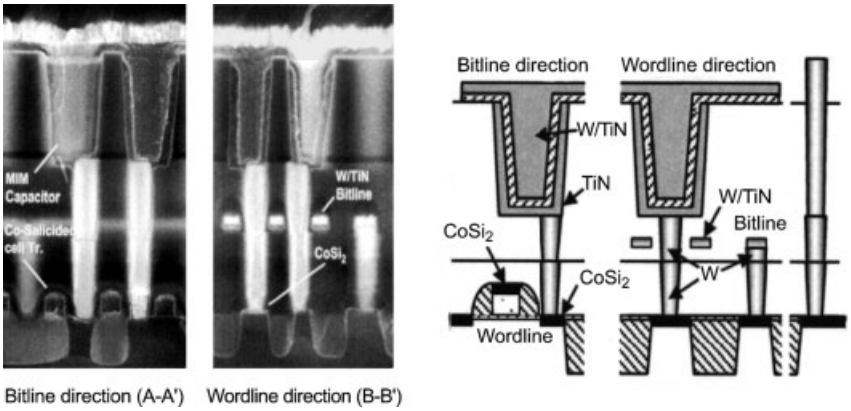


Figure 12.4 Stacked capacitor cell. This has a high-aspect-ratio storage node over the bit line and word line. The capacitor is composed of two electrodes and nitride-oxide or high-k dielectrics [4].

examples of trench and stacked capacitor cells, where each cell has high-aspect-ratio storage nodes under (trench) and over (stacked) the array transistor [3, 4].

Figure 12.5 shows the relationship between C_S and the trench diameter. C_S becomes smaller than 40 fF for a design rule of less than 90 nm because of the smaller capacitor area. To overcome the capacitor area reduction, hemispherical grains (HSG) [5, 6] can be used to enhance the surface area of the storage node in a deep trench cell (see Figure 12.6). HSG technology is an approach more typically used in the stacked capacitor cell. Another technique to enhance the capacitance is to use "high-k" dielectric materials such as Al_2O_3 or Ta_2O_5, where k denotes the dielectric permittivity. An example of a trench capacitor with a high-k Al_2O_3 dielectric is shown in Figure 12.7 [7]. The capacitance is increased by more than 30% compared with the standard nitride-oxide (NO) dielectric, which must be scaled to 1–2 nm thickness in a typical 65-nm process. As capacitor dielectrics are scaled, the leakage current increases (1.2 nm SiO_2 consists of only five atomic layers). Consequently, high-k dielectrics have been

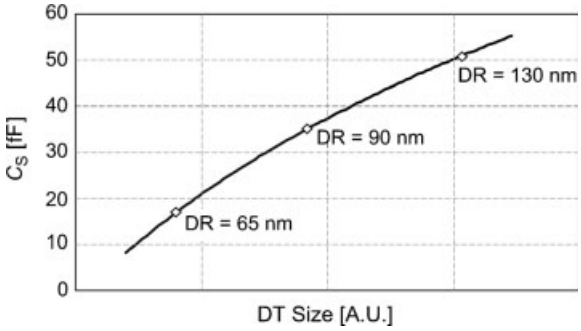


Figure 12.5 The relationship between capacitance (C_S) and trench diameter (DT).

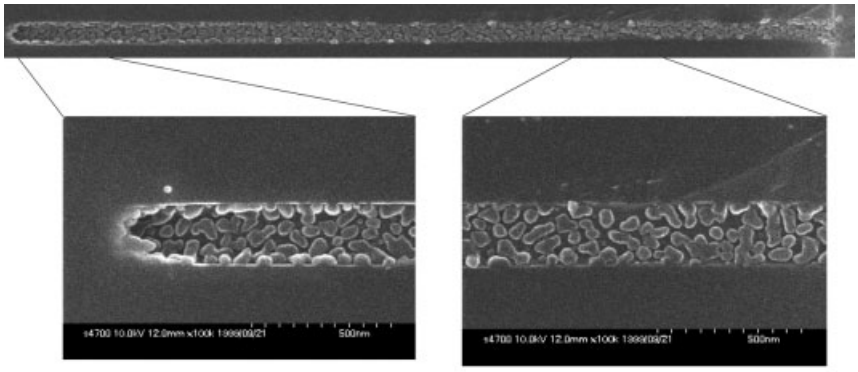


Figure 12.6 Trench capacitor with hemispherical grains (HSG).

proposed as alternatives for capacitor dielectrics to address the leakage problem. The most common high-k capacitor dielectrics used for gate dielectrics are Al_2O_3, Ta_2O_5 and hafnium-based dielectrics (e.g., HfO_2, HfSixOy). However, high-k gate dielectrics are not compatible with the polysilicon gate electrodes commonly used in today's integrated circuit technology. Their combination leads to threshold voltage uncontrollability and on-current reduction. Thus, metal gate electrodes with appropriate work functions must be used. Despite this, the implementation of high-k

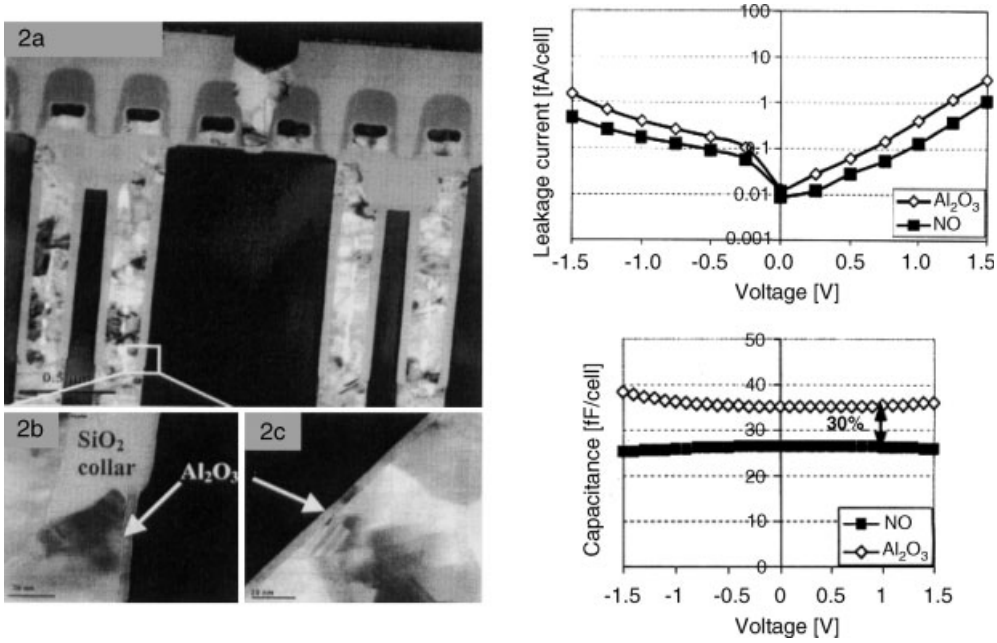


Figure 12.7 Trench capacitor with a high-k dielectric (Al_2O_3).

dielectrics and metal gate electrodes into complementary metal oxide–semiconductor (CMOS) technology is difficult, and involves many technical issues such as deposition methods, dielectrics reliability, charge trapping and interface quality. For capacitor dielectrics, it is much easier to implement high-k dielectrics because the threshold voltage shift induced by the charge trapping and the interface quality do not affect the capacitor characteristics compared with their effect on gate dielectrics. Thus, Al_2O_3 or Ta_2O_5 have been used as capacitor high-k dielectrics for a few DRAM products. For future DRAMs, high-k dielectrics will most likely be used not only for capacitor dielectrics but also for peripheral transistor gate dielectrics to overcome scaling problems.

12.4
Array Transistor Technologies

The recess-channel-array transistor (RCAT) [8] is used to reduce the electrical field near the drain to achieve a long data retention time in a stacked capacitor cell. Figure 12.8 shows the RCAT structure, which increases the effective gate length of the array transistor and mitigates the short-channel effect without increasing area.

Usually, channel doping enhances the electric field near the drain and degrades data retention characteristics because of the increased drain leakage current. To overcome this effect, the RCAT is used to reduce the electric field by separating the channel from the drain using an engraved channel region. This is effective in reducing the electric field and short-channel effects; however, the longer channel results in a small on-current in the array transistor. Thus, the RCAT is not suitable for high-speed writing.

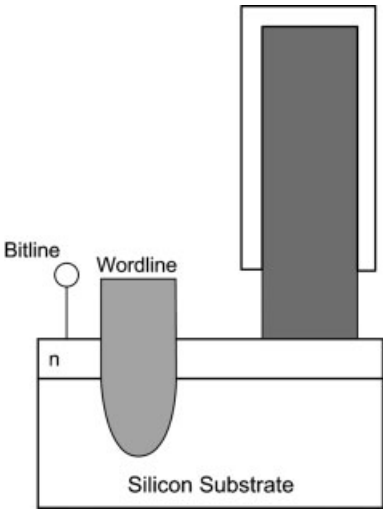


Figure 12.8 The recess-channel-array transistor (RCAT).

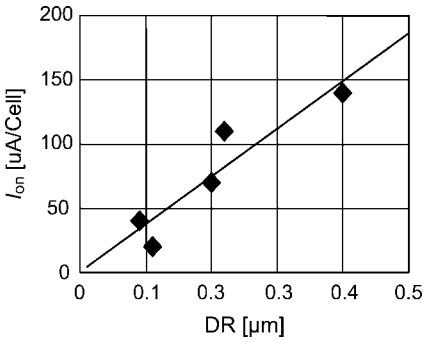


Figure 12.9 On-current (I_{on}) trend of array transistors.

The RCAT can be used down to around the 50 nm node, but below this another 3-D approach will be needed to satisfy the requirement of current drivability and to reduce the short-channel effect.

Figure 12.9 shows the on-current (I_{on}) trend of array transistors. The I_{on} decreases with transistor size in accordance with the design rule scaling; this in turn increases the signal delay in data sensing on the bit line (see Figure 12.10), in the case of reading a "1". When I_{on} is small, the signal appears on the bit line with a delay and approaches the "1" target level slowly.

To overcome these constraints in the array field-effect transistor (FET), a trench isolated transistor using sidewall gates (TIS) or a fin-array-FET can be adopted to improve the transistor performance, as in the case of silicon-on-insulator (SOI) transistors [9–12]. Figure 12.11 shows a "bird's-eye view" of a TIS-array FET where the TIS gate structure, which consists of a top gate and a sidewall gate enables a high on-current and a low off-current simultaneously because of the double-gate structure and high gate controllability. Figure 12.12 shows the T_{fin} (the width of the fin) dependence of the minimum gate length (L_g). A thinner T_{fin} would be expected to result in a marked reduction of off-current, which means that the TIS gate structure is very suitable for array transistors.

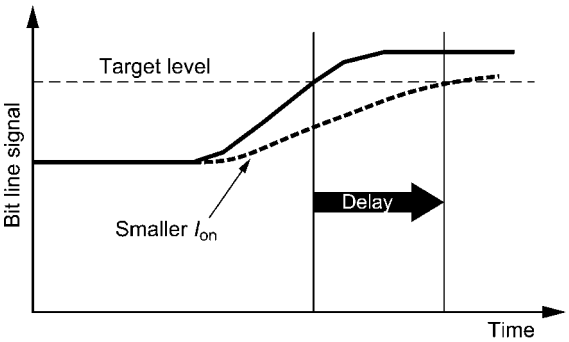


Figure 12.10 Signal delay by the smaller I_{on} current.

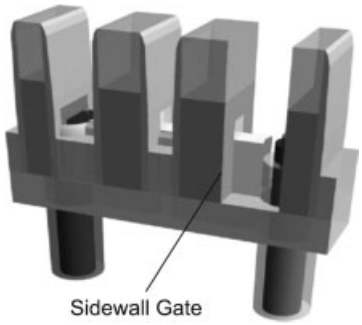


Figure 12.11 TIS/Fin array field-effect transistor (FET) DRAM.

In the TIS structure, the fin substrate is fully depleted and the double side gates contribute to the potential of each side channel. The subthreshold swing of the TIS transistor is smaller than that of the conventional planar transistor because of the strong effect of the sidewall gates. Thus, a small gate voltage difference can rapidly change the drain current from a small off-current to a large on-current. Moreover, the constant threshold voltage characteristics without a back-gate bias effect contribute to the large on-off-current ratio.

A more advanced array transistor is the *vertical transistor*, in which the source, gate and drain are arranged vertically. There are two types of vertical transistors. One uses the inner sidewall of the trench hole, while the other uses the outer sidewall of a silicon pillar for the channel. The former is suitable for trench capacitor cells [13], while the latter is known as a surrounding gate transistor (SGT) [14, 15]. The gate electrode of the SGT surrounds a pillar of silicon, and the gate length of the SGT is

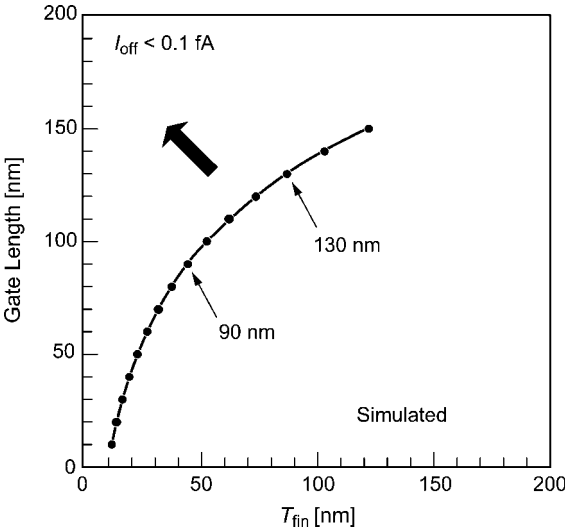


Figure 12.12 The dependence of fin width (T_{fin}) on minimum gate length (L_g).

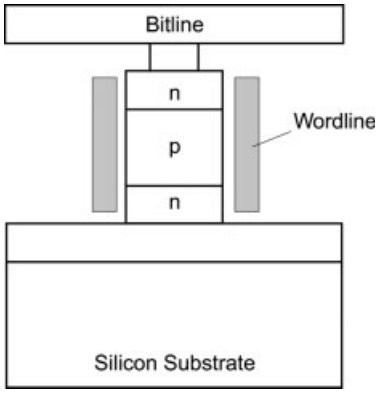


Figure 12.13 A surrounding gate transistor. The gate electrode of SGT surrounds a silicon pillar, with the gate length being adjusted by the pillar height.

adjusted by the pillar height, as shown in Figure 12.13. Therefore, the SGT has the merits of short-channel-effect immunity and superior current drivability resulting from the excellent gate controllability.

Planar array transistors cannot easily be scaled down (as noted above), and the TIS has good on-off-current ratio characteristics. The vertical transistor is different from the planar type in that the channel length is defined by the depth of the hole or the height of the pillar. Thus, the gate length is free from the minimum design rule and the cell area limitations, and can be selected to be sufficiently large so as to avoid the short-channel effect. Similar to the trench-type capacitor, the vertical transistor and the capacitor are formed in the same hole, and this contributes to the small cell size of less than $6\,F^2$. In the SGT cell, it is more difficult to form the capacitor and array transistor, although it has ideal array transistor characteristics. The SGT substrate is fully depleted and the surrounding gate contributes to the potential of the pillar surface channel. The subthreshold swing of the SGT cell is smaller than that of the conventional planar transistor and the TIS because of the stronger effect of the surrounding gate. Also, surrounding gate structures contribute to the large width of the transistor by using the entire perimeter of the pillar. Thus, a large on-off-current ratio can be attained without a back-gate bias effect, and the SGT can be used for $4\,F^2$ small-cell-size DRAMs.

DRAM scaling will continue to enable the integration of many advanced technologies in view of the huge size of the DRAM market. Thus, these advanced technologies will be used in future-generation DRAMs.

For the array transistor, the TIS/fin type structure is expected to be adopted using p^+ poly for obtaining a suitable threshold voltage with low channel doping. For the capacitor, a high-k dielectric (e.g., barium strontium titanate, BST) may be used in future DRAMs. For the peripheral transistor, mobility enhancement technologies such as the use of SiGe or a linear strain technique and high-k gate dielectrics will be adopted to achieve large driveability for a high-speed operation. A $4\,F^2$ cell layout is

predicted to be introduced to obtain higher-density DRAMs, and during this generation, new structures such as the SGT will be adopted.

12.5
Capacitorless DRAM (Floating Body Cell)

The difficulties of DRAM integration are mainly attributable to the necessity for constant capacitance, even when the cell size is reduced. For this reason, the integration of capacitors is very complicated for trench or stacked capacitors. The floating body cell (FBC) is a new concept of a DRAM without a capacitor. Because the cell is composed of one transistor, the FBC has a simple and compact structure.

Figure 12.14 shows the principle of the FBC, which involves the storage of the signal charge in the body of the cell transistor. To write "1", V_{WL} is biased to 1.5 V and V_{BL} to 2 V, so that the body potential (V_{body}) is increased by the holes that accumulate by impact ionization. To write "0", V_{WL} is biased to 1.5 V and V_{BL} to -1.5 V, so that V_{body} is decreased by ejecting holes from the body. The body potential difference (ΔV_{body}) is stored by setting V_{WL} to -1.5 V and V_{BL} to 0 V. In order to read the stored data, V_{BL} is biased to 0.2 V and V_{WL} is swept up to a certain level, while the bit line current (I_{read}) is measured. The I_{read}–V_{WL} characteristics are shown in Figure 12.15. The threshold voltage difference between a "0" cell and a "1" cell (ΔV_T), which is an index of the data reading margin, is about 0.32 V. In order to increase ΔV_T or ΔI_{read}, C_S (the body capacitance for data storage) plays an important role, because the ΔV_{body} of the hold state is reduced by WL-body and BL-body capacitance coupling. A back gate is used to enable charge accumulation in the body [16–18], and also to increase C_S, which stabilizes the body potential. The structure of the FBC has a modified double-gate configuration. A transmission electron microscopy cross-section of a fully depleted FBC, with thin SOI and BOX layers, is shown in Figure 12.16.

The body capacitance is small compared to the standard DRAM capacitor, typically by two orders of magnitude. However, the leakage current of the FBC storage node is small because of the small p–n junction area, which is located only at the channel-side

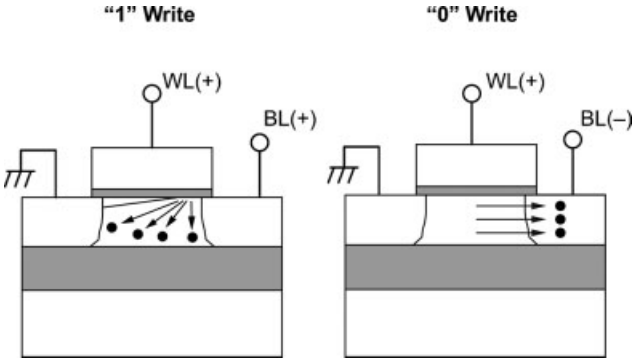


Figure 12.14 The write operation of the floating body cell (FBC). See text for details.

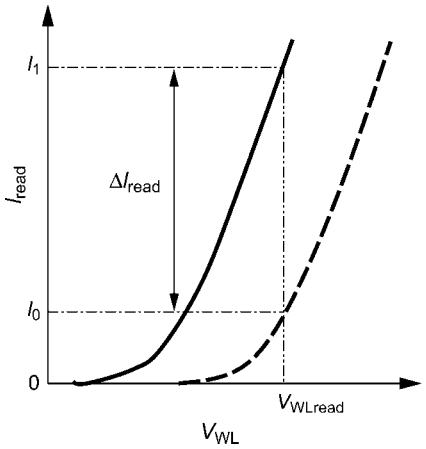


Figure 12.15 The read operation of the floating body cell (FBC). See text for details.

edges of the source and drain. Thus, the retention time of the FBC is reduced slightly compared to the standard DRAM. As a consequence, and because of the short retention time, the FBC is suitable for high-performance embedded DRAM applications rather than low-power applications.

The SOI structure has been widely used for high-performance applications, particularly game processors, and is expected to be used in the embedded DRAM for on-processor caches. An FBC using a SOI substrate can easily be used for these applications with the same compatibility as the SOI substrate. As the FBC is composed of one transistor and has no capacitor, it is scalable down to the 32 nm node. Details of this structure are provided in Ref. [17]. An image of an FBC with 128 Mb DRAM, along with the chip features, is shown in Figure 12.17. The FBC, which has dimensions of 7.6×8.5 mm, contains all of the necessary circuits (including internal voltage generators) and operates using a single 3.3 V power supply [18].

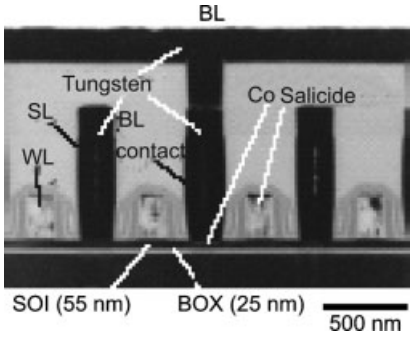
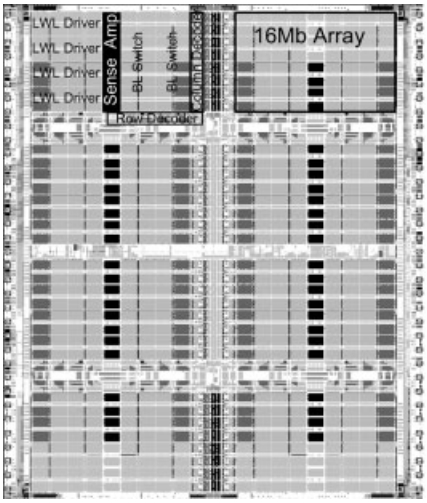


Figure 12.16 Transmission electron microscopy cross-section of a fully depleted floating body cell (FBC), showing the thin SOI and BOX layers.



Bit Organization	8 M word × 16 bit
Cell Structure	$t_{ox}/t_{Si}/t_{BOX}$ = 6 nm/55 nm/25 nm, L_g = 150 nm
Cell Size	0.33 μm × 0.515 μm = 0.17 μm^2
Peripheral	body tied SOI, t_{ox} = 6 nm, L_{gn} = 450 nm, L_{gp} = 400 nm
Random Access	18.5 ns (Normal Mode) 25.7 ns (VSRAM Mode)
Refresh Cycle	4 K
Redundancy	8 Red. LWL/1 Mb & 16 Red. LBL/2 Mb
Special Modes	Fast Page Mode, VSRAM Mode
Chip Size	7.6 mm × 8.5 mm = 64.6 mm^2

Figure 12.17 Floating body cell (FBC) 128 Mb DRAM and its features.

12.6 Summary

Today, while the demand for DRAM remains greater than for any other type of memory, the capacity of DRAM is continually increasing such that variations are now becoming available for both low-power and high-speed applications. Because of the scaling limitations, the TIS/fin array-transistor is expected to be used in future-generation DRAMs, with more advanced DRAMs – such as vertical transistors such as the SGT – most likely being used for DRAMs with a smaller cell size. In addition, the capacitorless DRAM – the FBC – shows great promise as a candidate for next-generation embedded DRAMs offering both high density and high speed. Clearly, the use of 3-D structures should help to overcome the scaling problems likely to be encountered in future-generation memories.

References

1 Regitz, W.M. *et al.* (1970) *IEEE Journal of Solid-State Circuits*, **SC-5**, 181–186.
2 Boonstra, L. *et al.* (1973) *IEEE Journal of Solid-State Circuits*, **SC-8**, 305–310.
3 Yanagiya, N. *et al.* (2002) *IEDM Technical Digest*, 58–61.
4 Arai, S. *et al.* (2001) *IEDM Technical Digest*, 403–406.
5 Saida, S. *et al.* (2000) *Proceedings of ISSM*, 177.
6 Amon, J. *et al.* (2004) A highly manufacturable deep trench based DRAM cell layout with a planar array device in a 70 nm technology. *IEDM Technical Digest*, 73–76.
7 Seidl, H. *et al.* (2002) *IEDM Technical Digest*, 839–842.
8 Kim, J.Y. *et al.* (2003) The breakthrough in data retention time of DRAM using Recess-Channel-Array Transistor (RCAT)

for 88 nm feature size and beyond. *VLSI Technical Digest*, 11–12.

9 Hieda, K. *et al.* (1987) New effects of trench isolated transistor using side-wall gates. *IEDM Technical Digest*, 736–737.

10 Hisamoto, D. *et al.* (2000) FinFET – a self-aligned double-gate MOSFET scalable to 20 nm. *IEEE Transactions on Electron Devices*, **47**, 2320–2325.

11 Katsumata, R. *et al.* (2003) Fin-Array-FET on bulk silicon for sub-100 nm trench capacitor DRAM. *VLSI Technical Digest*, 61–64.

12 Weis, R. *et al.* (2001) A highly cost efficient $8F^2$ DRAM cell with a double gate vertical transistor device for 100 nm and beyond. *IEDM Technical Digest*, 415–418.

13 Lee, D.-H. *et al.* (2007) Improved cell performance for sub-50 nm DRAM with manufacturable bulk FinFET structure. *VLSI Technical Digest*, 164–165.

14 Sunouchi, K. *et al.* (1989) A surrounding gate transistor (SGT) cell for 64/256 Mbit DRAMs. *IEDM Technical Digest*, 23–26.

15 Goebel, B. *et al.* (2002) Fully depleted surrounding gate transistor (SGT) for 70 nm DRAM and beyond. *IEDM Technical Digest*, 275–278.

16 Shino, T. *et al.* (2004) Fully-depleted FBC (floating body cell) with enlarged signal window and excellent logic process compatibility. *IEDM Technical Digest*, 281–284.

17 Shino, T. *et al.* (2006) Floating body RAM technology and its scalability to 32 nm node and beyond. *IEDM Technical Digest*, 569–572.

18 Ohsawa, T. *et al.* (2005) An 18.5 ns 128 Mb SOI DRAM with a floating body cell. *ISSCC Technical Digest*, 458–459.

19 Ranica, R. *et al.* (2004) A capacitor-less DRAM cell on 75 nm gate length, 16 nm thin fully depleted SOI device for high density embedded memories. *IEDM Technical Digest*, 275–280.

13
Ferroelectric Random Access Memory

Soon Oh Park, Byoung Jae Bae, Dong Chul Yoo, and U-In Chung

13.1
An Introduction to FRAM

An ideal non-volatile memory should possess the required characteristics such as high density (high scalability and compact cell size), high reliability (excellent retention and endurance), low cost, and high performance (random access, high read/write speed, and low power consumption) [1, 2]. Although Si-based Flash memory with high density and low cost is the leading non-volatile memory, it cannot basically meet the needs of high endurance and performance characteristics. Therefore, new concepts for non-volatile memory such as FRAM (Ferroelectric RAM), PRAM (Phase change RAM), MRAM (Magnetoresistive RAM), and RRAM (Resistive RAM) have been demonstrated as strong candidates for an ideal non-volatile memory.

Among these emerging memories, PRAM uses phase-change material as a storage element, and shows high scalability and compact cell size owing to its simple cell structure [3]. However, it has disadvantages such as long crystallization time, high power consumption for phase-change switching, and low endurance performance. MRAM uses magnetic material for data storage and shows excellent high speed and good reliability performance, but it requires a large cell area to make a unit cell [4]. Recently emerged RRAM uses resistive switching material as a storage element, but its technology is not yet matured [5]. Finally, FRAM uses ferroelectric materials as a storage element and has been a strong candidate to a universal memory since the late 1980s [6–8]. Because of its similar structure and operation scheme to DRAM (Dynamic RAM) and additional non-volatility, FRAM has been developed as a universal memory for one-chip solution. The operation scheme, reliability of ferroelectric capacitor, and the technology of high-density FRAM for a universal memory will be introduced in the following sections.

Nanotechnology. Volume 3: Information Technology I. Edited by Rainer Waser

ISBN: 978-3-527-31738-7

13.1.1
1T1C and 2T2C-Type FRAM

The operation and architecture of capacitor-type FRAM is almost identical to that of dynamic random access memory (DRAM). It should be noted that every cell has its own separate plate line in capacitor-type FRAM, whereas DRAM uses common plate-line in the level of a half V_{dd}. Necessarily, a ferroelectric capacitor replaces the linear capacitor in a metal-insulator-metal (MIM) storage element.

A schematic view of the two-transistor–two-capacitor (2T2C) -type FRAM and the one-transistor–one-capacitor (1T1C) -type FRAM [9] are shown in Figure 13.1. In the 2T-2C type, the switching and non-switching charges of two adjacent ferroelectric capacitors are used as data "1" or "0" charges, which have a large sensing window and uniform cell operation. However, the cell area is too large to be used for high-density ferroelectric devices because two capacitors can store only a single bit. On the other hand, the 1T1C type provides an advantage of small cell area because single capacitor stores single bits. However, the sensing margin of the 1T1C type is reduced as a half of that of 2T2C type by setting a reference level in the middle of data "1" and "0". The sensing margin might be further reduced due to the variation of reference ferroelectric capacitors. Nevertheless, the 1T1C type is used as the cell structure due to its small cell size in most high-density ferroelectric devices.

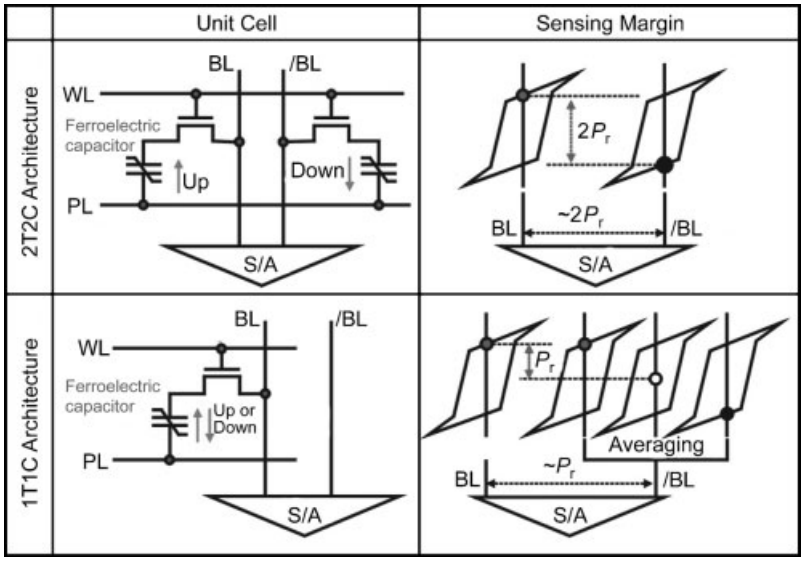


Figure 13.1 Comparison of 2T2C and 1T1C FRAM architectures in respect of cell size and sensing margin.

13.1.2
Cell Operation and Sensing Scheme of Capacitor-Type FRAM

Figure 13.2 illustrates the writing operation of 1T1C-type FRAM. Figure 13.2a is the schematic of 1T1C FRAM which is composed of the word line (WL), bit line (BL), and plate line (PL). Figure 13.2b shows the charges preserved in the hysteresis curve, while Figures 13.2c and d show the timing diagrams of writing data "1" and "0". To write "1" into the memory cell, the BL is raised to V_{pp} and PL is kept as ground (GND). The polarization directions are from PL to BL and the $-P_r$ value is preserved. To write "0", the BL is kept as GND and the PL is kept high as V_{pp}. Thus, the opposite direction of the polarization is generated and $+P_r$ value is preserved.

Figure 13.3 illustrates the reading operation of 1T1C-type FRAM [10]. A read access begins by precharging the BL to GND, after which the PL is raised to V_{pp}. This establishes serial two capacitors consisting of C_s and C_{BL} between the PL and the GND, where C_s is the capacitance of ferroelectric storage element and C_{BL} is the parasitic capacitance of BL. Therefore, the V_{pp} is divided into V_f and V_{BL} between C_s and C_{BL} according to their relative capacitance. Depending on the data stored, the voltage developed on the ferroelectric capacitor and BL can be approximated as follows:

$$V_f = C_{BL} \times V_{PP}/(C_s + C_{BL}) \tag{13.1}$$

$$V_{BL}(\text{Data"1"}) = dQ_{sw}/(C_{sw} + C_{BL}) \tag{13.2}$$

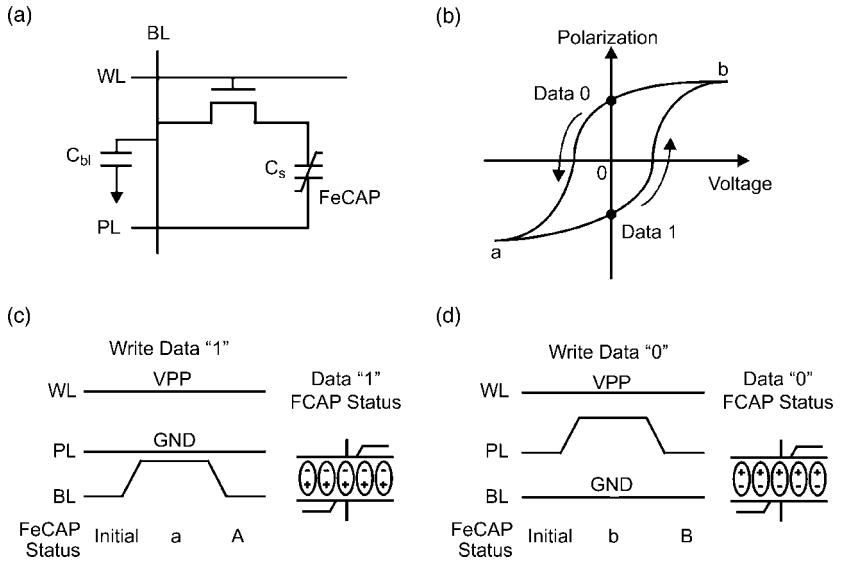


Figure 13.2 The writing operation of 1T1C-type FRAM. (a) Schematic of 1T1C FRAM which is composed of word line (WL), bit line (BL), and plate line (PL). (b) Charges preserved in hysteresis curve. (c,d) Timing diagrams of (c) writing data "1" and (d) writing data "0".

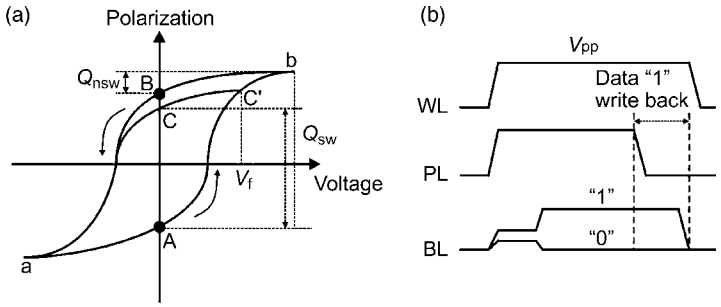


Figure 13.3 Schematic illustration of read operation procedures in 1T1C FRAM.

$$V_{BL}\,(\text{Data“0”}) = dQ_{nsw}/(C_{nsw} + C_{BL}) \tag{13.3}$$

In general, the voltage developed in the BL is too small to sense charge differences. Therefore, a sensing amplifier should be used in order to drive the BL to full V_{pp} if the data is "1", or to 0 V if the data is "0". The structure of the sensing amplifier of capacitor-type FRAM includes the cross-coupled latch sense amplifier of DRAM. It can be classified to a folded bit line and an open bit line according to the cell array, as shown in Table 13.1 and Figure 13.4 [11]. The open bit-line scheme is applicable to 1T1C structure, and the folded bit-line scheme can be applied to both 1T1C and 2T2C structures.

Table 13.1 Comparison of FRAM sense amplifier types.

	Sense amplifier	Realization	Noise immunity	Sensibility
Folded bit-line	1 ea/2 BL	Easy layout	Same noise environment	Good
Open bit-line	1 ea/1 BL	Difficult layout	Different noise environment	Not good

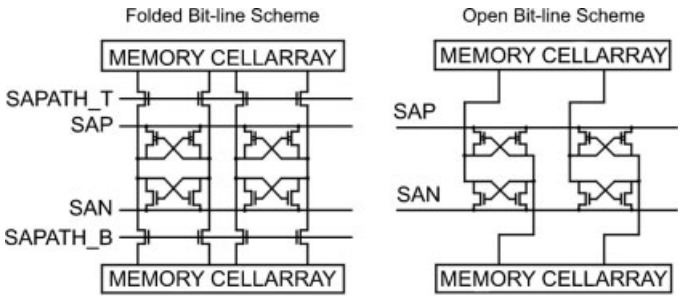


Figure 13.4 Schematic diagrams of FRAM sense amplification.

13.2
Ferroelectric Capacitors

Similar to the DRAM device, the capacitor technology concerning ferroelectrics serves as a guideline to the development of FRAM devices. In this regard, the material characteristics and reliability of typical ferroelectric capacitors will be considered in this section.

13.2.1
Ferroelectric Oxides

Representative ferroelectric materials for complementary metal oxide–semiconductor (CMOS) integration can be divided to two groups, including perovskite-structured (PZT and $BiFeO_3$) and Bi-layer structured (SBT, BLT, and BTO) materials. The characteristics of these are summarized in Table 13.2 [12, 13]. The crystal structure of PZT can be either tetragonal or rhombohedral according to the Zr/Ti composition ratio below Curie temperature. In SBT, two $SrTaO_3$ perovskite blocks and one $(Bi_2O_2)^{2+}$ layer constitute one unit cell, as shown Figure 13.5b. In BLT, La atoms are partially substituted to Bi atom in $Bi_4Ti_3O_{12}$ (BTO) crystal which is composed of three TiO_6 octahedra and one $(Bi_2O_2)^{2+}$ layer leading the wanted crystal structure. These differences of unit cell structure largely determine the characteristics of corresponding ferroelectrics. Therefore, the typical properties of PZT, SBT and BLT can be compared from this point of view.

First, the remanent polarization value (P_r) determines the sensing margin between data "0" and "1" (the larger $2P_r$, the better sensing-margin), while the coercive voltage (V_c) or coercive field (E_c) decides the operating voltage in an FRAM device (the smaller V_c, the better operation voltage). PZT shows large P_r and E_c values because of strong interactions between neighboring perovskite unit cells. In contrast, SBT and BLT request the anisotropic growth along the a-b axis to attain direct interactions between the neighboring perovskite unit cells toward electrical field direction. The smaller P_r and E_c values of SBT, compared to PZT, tend to increase by Nb doping.

Table 13.2 The features of typical ferroelectrics used for FRAM.

Ferroelectrics	$Pb(Zr,Ti)O_3$ (PZT)	$SrBi_2Ta_2O_9$ (SBT)	$(Bi,La)_4Ti_3O_{12}$ (BLT)
P_r [$\mu C\ cm^{-2}$]	10–40	5–10	10–15
E_c [$Kv\ cm^{-1}$]	50–70	30–50	30–50
Endurance	Poor on Pt electrode Good on oxide electrode	Good on Pt electrode	Good on Pt electrode
Crystallization temperature [°C]	450–650	650–800	650–750
Curie temperature [°C]	~400	~400	~400

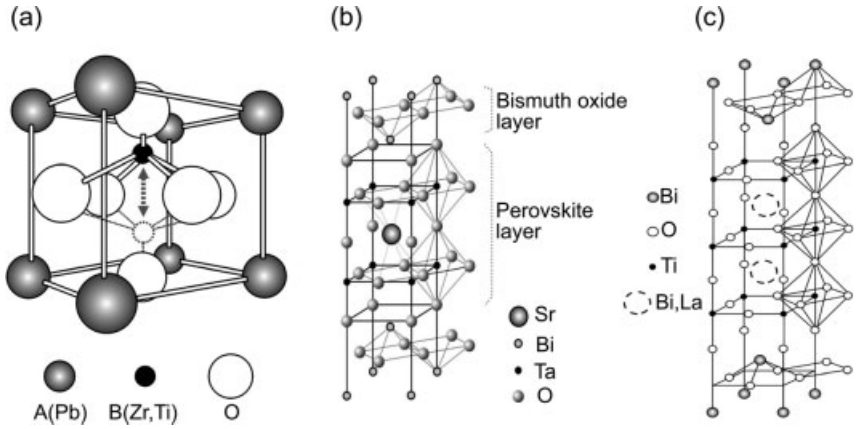


Figure 13.5 The crystal structures of ferroelectric (a) PZT, (b) SBT, and (c) BLT.

Ferroelectric films must attain a crystallized perovskite structure in order to show the polarization behavior. Therefore, the plentiful oxygen atmosphere and high substrate temperature in order to crystallize ferroelectric oxide lead to the demand for a noble metal electrode and oxidation barrier to achieve CMOS integration. In a stacked FRAM cell with COB (capacitor over bit-line) structure, the capacitor is located directly on the top of the MOSFET drain, which requires the low-temperature process to realize high-density CMOS integration. In this respect, the lower crystallization temperature of PZT than that of SBT and BLT is advantageous for the fabrication of future high-density FRAM.

To date, PZT has long been the leading material considered for ferroelectric memories, and has been superior to SBT and BLT. The higher P_r value and lower process temperature of PZT can act as strong merits for the fabrication of high-density COB cells in CMOS integration. On the other hand, lead-free SBT and BLT can be considered for environmentally friendly FRAMs. Recently renewed multiferroic $BiFeO_3$ exhibits both ferroelectric and magnetic properties, but it is unclear whether this is useful for memory application, or not [14]. In particular, the proper ferroelectrics for CMOS integration should be chosen by serious consideration for the degradation induced by hydrogen, plasma, stress, and heat in the succeeding integration processes [15].

13.2.2 Fatigue

"Fatigue" is a term describing the fact that the remanent polarization becomes small when a ferroelectric film experiences numerous polarization reversals. When PZT on Pt electrodes suffers from reading/writing cycles over 1E5 cycles, the P_r value shows a conspicuous reduction, which limits the repeated use of a memory. A few reports about non-fatigue phenomena of Pt/PZT/Pt capacitors should be regarded with great care because this type of behavior can be observed when the applied voltage is less

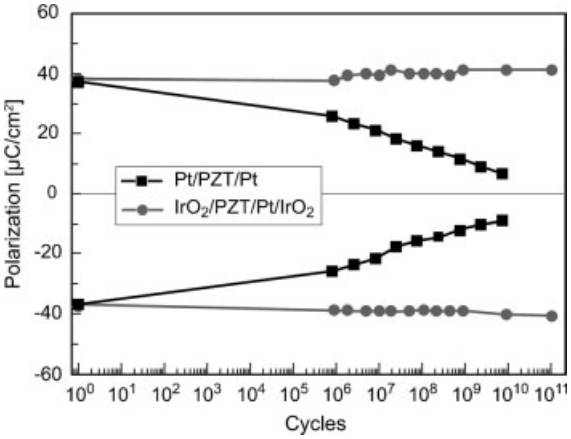


Figure 13.6 Fatigue properties of PZT and an IrO_2 electrode.

than $V_{(90\%)}$. Fatigue behavior is strongly related to the generation of oxygen vacancy by the repeated cycles, which induces dipole-pinning electrons for charge-neutrality; this is why oxygen vacancies are the only mobile ionic species in the lattice even at the room temperature on the basis of defect chemistry model. However, the fatigue problems of PZT can be almost solved at present by the use of conducting oxide electrodes (e.g., IrO_2, RuO_2, $SrRuO_3$, $CaRuO_3$, $LaNiO_3$, and LSCO), ensuring no degradation of the P_r value even up to 1E12 cycles. Figure 13.6 shows a comparison of fatigue properties in PZT capacitor with Pt and IrO_2 electrodes [16]. This improvement of fatigue property can be explained by the fact that oxygen in the IrO_2 electrodes reduces oxygen vacancies, which prevents fatigue degradation reducing the dipole-pinning effect. As Ir is stably converted into IrO_2 under oxygen at ambient temperature, the fatigue problem can be remarkably enhanced in the case of using an IrO_2 oxide electrode.

In contrast to the PZT film, an SBT film does not show the fatigue phenomenon up to 1E13 switching cycles, even if Pt electrodes are used. It was speculated that the $(Bi_2O_2)^{2+}$ interlayer can compensate the produced oxygen vacancy. However, similar Bi-layer structured BTO shows fatigue problems on a Pt electrode, which suggests that the simple charge-compensation role of the $(Bi_2O_2)^{2+}$ layers is not sufficient to make the fatigue-free films. This reduction in polarization could be much alleviated by using La-doped BTO (so-called BLT). Accordingly, the limited switching cycles of dipoles are no longer a serious problem for any ferroelectric materials, as shown in Figure 13.7 [13, 14].

13.2.3 Retention

Polarization retention is the ability of poled ferroelectric capacitors to preserve the poled state over time (generally 10 years into the future at 85 °C). The retention property represents an important reliability issue for non-volatile ferroelectrics

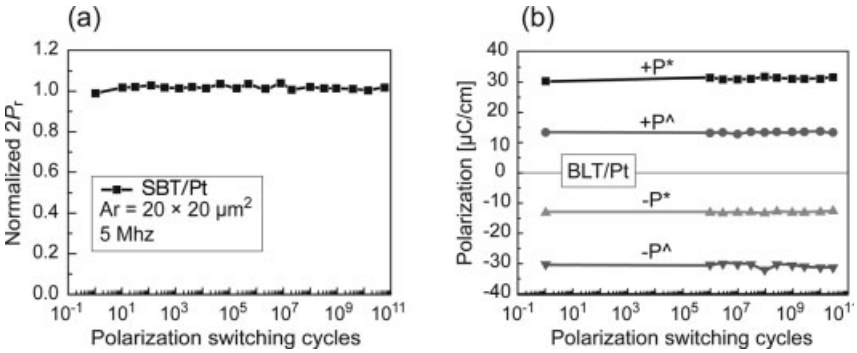


Figure 13.7 Fatigue properties of (a) SBT and (b) BLT film on Pt electrode. The symbols are indicated in Figure 13.9.

memories. Most commonly, the retention of ferroelectrics can be classified into the same-state and opposite-state retention.

The same-state retention, which is closely related to aging, represents the loss of polarizability when one first writes the datum of "0" or "1" in a capacitor with electrical pulses, and reads the datum again after long period without changing the initial status. Therefore, the same-state retention failure can occur when the relaxation component in the opposite-polarity state increases at the expense of the relaxation component in the stored polarity state. The stored polarity status can be stabilized by the use of ferroelectrics with "high" Curie temperature, after which the same-state retention loss can be improved from the viewpoint of thermodynamics. For instance, $BaTiO_3$ is not applicable for non-volatile FRAM because of its low Curie temperature (~140 °C), although this can be raised to 500 °C by imposing biaxial compressive strain.

The opposite-state retention, which is closely related to imprint, represents the loss of polarizability when one first writes the datum of "0" or "1" in a capacitor with electrical pulses and reads the "changed" datum again. In words, the same-state retention is a longstanding problem of the read-only memory (ROM), while the opposite-state retention is that of the random-access memory (RAM), because information must be modifiable (as shown in Figure 13.8) [17].

The opposite-state retention failure occurs when a capacitor, which has aged considerably in one state, is switched to the opposite state. In this case, the capacitor behaves as if it would prefer to remain in the original state. The charge defects are activated by thermal energy and redistributed by the polarization field. Therefore, the resulting internal field causes a lower energy barrier and invokes polarization back-switching during the delay time, as shown in Figure 13.9 [18]. Accordingly, the opposite-state retention can be solved by minimizing the space charges, which results from defects inside the ferroelectrics, domain wall motion, or defects near the electrode-ferroelectric interfaces.

The thickness scaling of ferroelectric films is indispensable when pursuing a low switching voltage, making this suitable for integrated electronics applications. However, to date, thinner ferroelectric films have shown serious degradation of

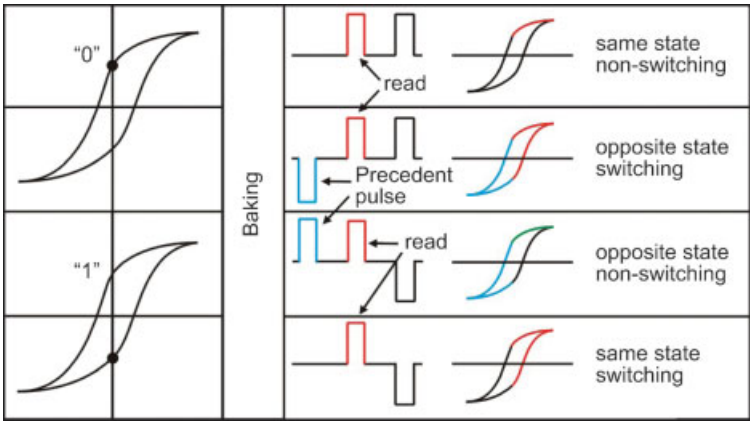


Figure 13.8 Retention pulse sequence.

opposite-state retention compared to same-state in cumulative studies. Consequently, during the past few years much attention has been focused on the failure mechanism of the opposite-state retention. In order to solve the opposite-state retention failure problem, it is necessary to utilize frequently used technologies such as seeding, metalorganic chemical vapor deposition (MOCVD), and perovskite oxide electrode in the case of PZT ferroelectrics. Thus, these technologies will be reviewed briefly in the following sections.

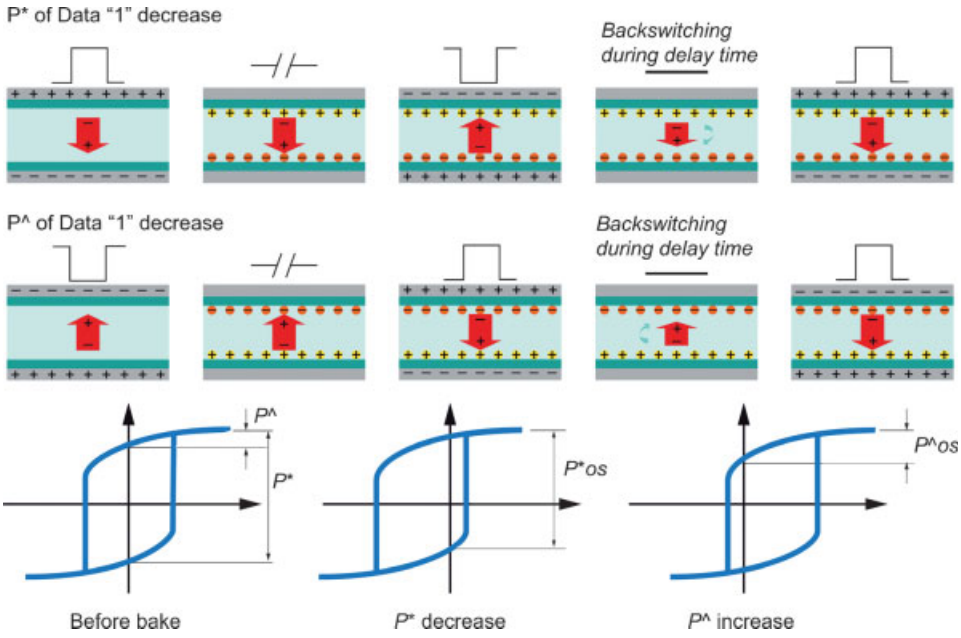


Figure 13.9 Retention failure mechanism.

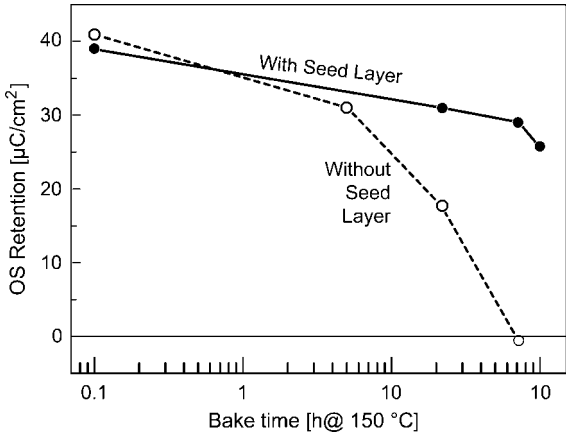


Figure 13.10 Improved opposite-state retention by the use of $PbTiO_3$ seed layer.

13.2.3.1 Crystallinity of PZT Film

Frequently, the perovskite-structured PZT phase can be generated by utilizing nucleation and the grain growth process from the pyrochlore phase. Because the nucleation process is strongly dependent on the substrates, an appropriate seed layer can supply nucleation sites in order to decrease activation energy for crystallization of the perovskite phase. A $PbTiO_3$ seed layer is very effective for supplying a high density of nuclei in the initial stage of deposition, because the crystallization temperature (350∼680 °C) is lower than that of PZT (>650 °C). The succeeding PZT film shows a much enhanced crystallinity and preferred orientation, and thereby exhibits improved retention result (see Figure 13.10 [19, 20]. For this purpose, an optimum thickness of $PbTiO_3$ is essential because a thinner $PbTiO_3$ layer cannot play a sufficient role for the seed layer, while a thicker one may cause adverse effects on the electrical properties of the overall film. A $PbTiO_3$ seed layer is helpful in the initial stage of film growth, but still constitutes a portion of the ferroelectric films. Therefore, the use of a perovskite oxide electrode as a seed layer may provide a better means of preparing reliable ultrathin ferroelectric films, because the seed layer belongs to the electrode and not to the ferroelectrics.

13.2.3.2 The MOCVD Deposition Process

Most current FRAM cells below 64 Mb density are based on a planar capacitor stack. In CMOS integration, it is commonplace to use either chemical solution deposition (CSD) or a sputtering technique for the deposition of planar films, and chemical vapor deposition (CVD) for a conformal deposition, based largely on an economics viewpoint. Somewhat ironically, however, such common sense has caused a "dribbling" (slow movement, low amplitude) of technological developments in this field. For example, the current deposition method used for "planar" PZT films is mostly based on an "MOCVD process" in order to pursue the excellent opposite-state retention properties [21]. The comparative retention of a PZT film deposited by CSD

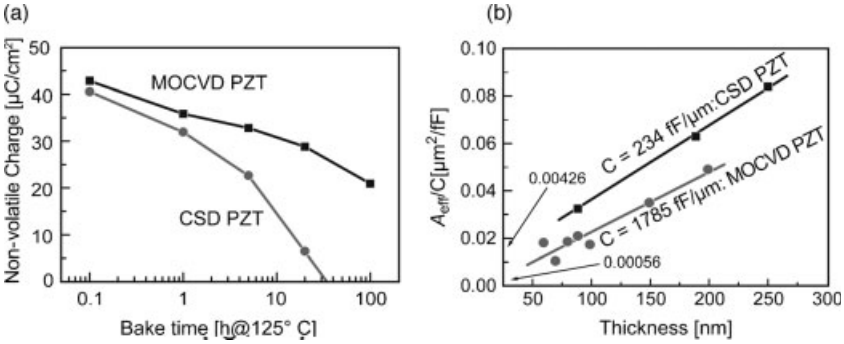


Figure 13.11 Improved opposite-state retention by using the metalorganic chemical vapor deposition (MOCVD) deposition process.

and MOCVD method is shown in Figure 13.11, where the MOCVD PZT film shows superior retention properties to those of CSD films due to the low defect density in ferroelectrics and/or interfaces. It may be speculated that an as-crystallized PZT film on an Ir electrode can be obtained by using the MOCVD process, such that the non-switching layer at the interface between the electrode and ferroelectrics is thinner, without the formation of Pt_3Pb alloys.

13.2.3.3 **Perovskite Oxide Electrode**

Most ferroelectric materials have a perovskite crystal structure, as outlined in the previous section. Therefore, if a conducting oxide electrode having a perovskite structure is use, then ferroelectric properties such as reliability can be greatly improved due to the reduction of any non-ferroelectric dead layer at the interface between the ferroelectrics and electrodes. Such remarkable improvement of retention properties by using an $SrRuO_3$ electrode with a perovskite structure is illustrated in Figure 13.12 [22].

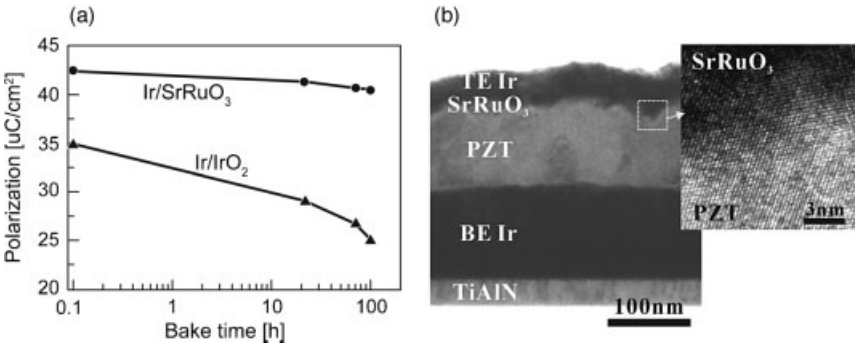


Figure 13.12 (a) Retention properties of a PZT capacitor with $Ir/SrRuO_3$ and Ir/IrO_2 electrodes. (b) Transmission electron microscopy image of the interface between $SrRuO_3$ and PZT films.

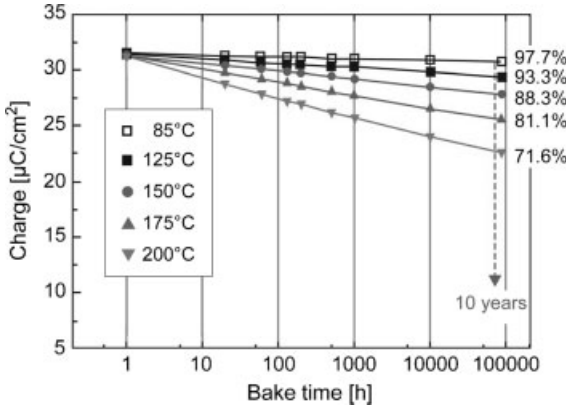


Figure 13.13 The ultrahighly reliable properties of the FRAM device.

During recent years, although perovskite oxide electrodes such as $SrRuO_3$, $LaNiO_3$ and $CaRuO_3$ have undergone intense investigation, the problems of high leakage currents – which inevitably are induced by high defect densities in the oxide electrode – remain to be overcome.

Recently, the successful development of an ultrahighly reliable FRAM device has been reported, and the retention properties of this fully integrated device, at different temperatures, are illustrated graphically in Figure 13.13 [23]. Based on these findings, the FRAM device could be expected to maintain >80% of any initial charge, even after 10 years at 175 °C.

13.3
Cell Structures

A vertical scanning electron microscopy image of the FRAM cell structure is shown in Figure 13.14 [24]. The cell is composed of a cell transistor, capacitor, buried contact, bit line, word line, and plate line. The cell structure can be divided into the CUB (capacitor under bit line) and COB (capacitor over bit line) structures, the merits and demerits of which are considered in the following section.

13.3.1
CUB Structure

In the CUB cell structure, the ferroelectric capacitor is formed beside the cell transistor, as shown in Figure 13.15 [25]. This requires a large cell area compared to the COB cell structure, in which the ferroelectric capacitor is formed over the cell transistor. The CUB scheme has no thermal budget limitations on the ferroelectric film deposition, and the subsequent anneal process for crystallization of the ferroelectric film, because the ferroelectric capacitor formation processes (including stack deposition and dry etching) are completed before the metallization process is

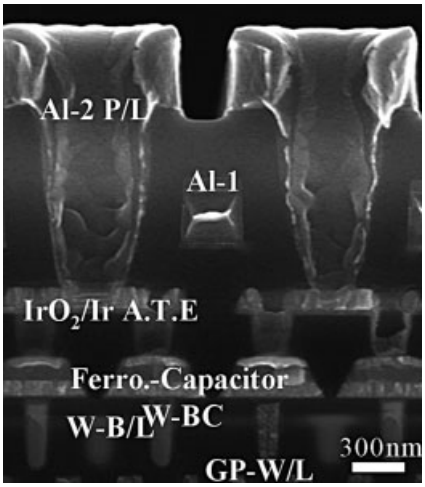


Figure 13.14 A vertical scanning electron microscopy image of the FRAM cell.

carried out. Due to technical difficulties in realizing a ferroelectric film with low thermal budget processes, and of identifying a suitable oxidation barrier metal which is stable above 600 °C, the early FRAMs were developed with a CUB cell structure, thereby sacrificing cell size efficiency.

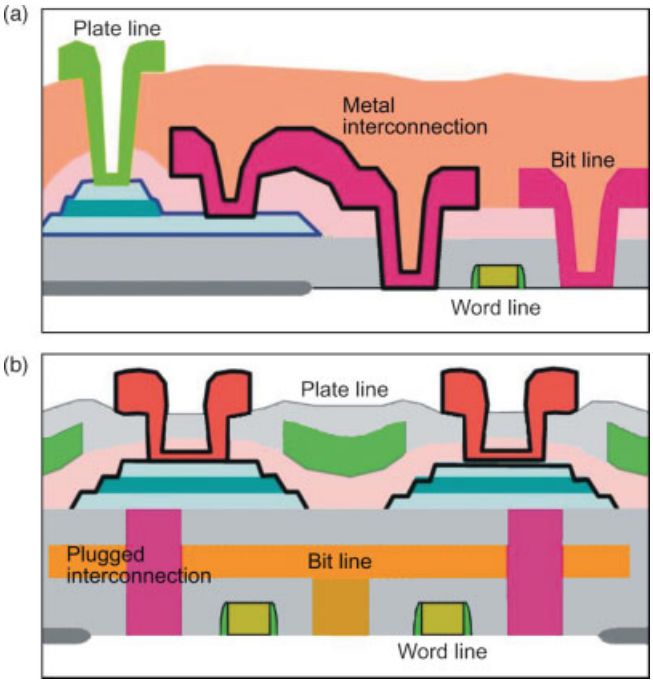


Figure 13.15 Schematic diagram of (a) capacitor under bit line (CUB) and (b) capacitor over bit line (COB) cell structures.

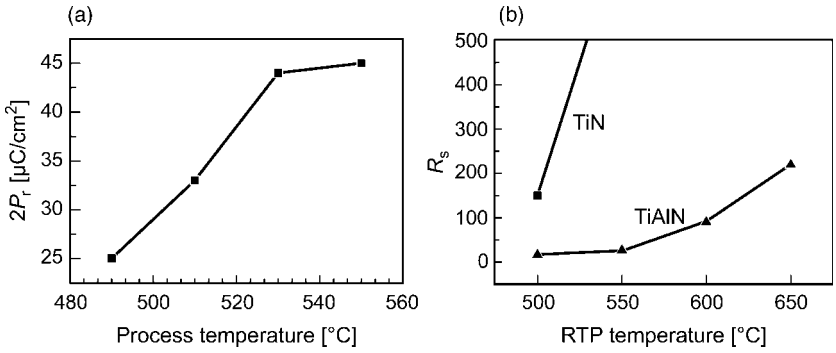


Figure 13.16 (a) $2P_r$ variation versus MOCVD process temperature and (b) comparison with oxidation resistance of TiAlN and TiN.

13.3.2
COB Structure

In the COB cell structure, the ferroelectric capacitor is formed over the bit line. Thus, the realization of a COB cell structure requires both a new buried contact (BC) plug and new metal technologies for the oxidation barrier. A stable contact between the BC plug and the bottom electrode must be provided when the ferroelectric capacitor has been processed at a high temperature of 600 °C or above [26]. As shown in Figure 13.16a, a high-temperature process is essential to obtain a sufficient polarization value in an MOCVD PZT process [27]. In order to prevent oxidation of the BC plug, various oxidation-barrier metals have been widely investigated; among these, a TiAlN film proved successful in preventing oxidation of the BC plug. The oxidation resistance properties of TiAlN and TiN thin films, as a function of temperature, are illustrated graphically in Figure 13.16b.

As mentioned above, as the COB structure is more beneficial with regards to high-density integration than are CUB structures, an increasing proportion of FRAM devices are today adopting the COB structure.

13.4
High-Density FRAM

In this section, the current status of planar capacitor technology, together with the technical issues involved in the development of 3-D capacitors for high-density FRAM device application, will be discussed.

13.4.1
Area Scaling

In order to achieve a high-density FRAM, the cell size must be scaled down as much as possible. Unfortunately, however, there exists a scaling limit because the

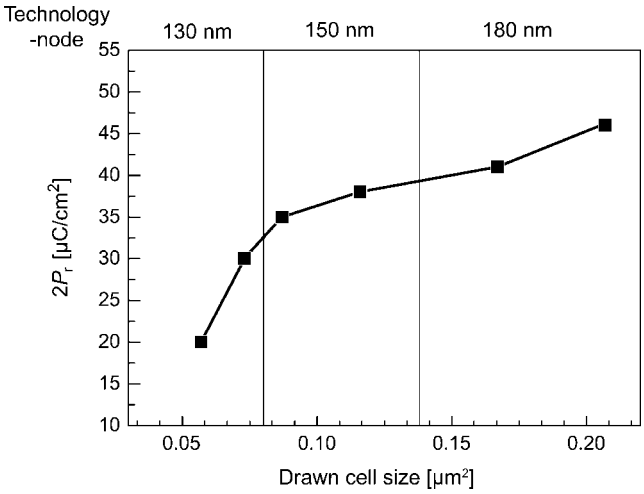


Figure 13.17 Polarization decay as a scale-down of the drawn cell size.

polarization value ($2P_r$) decreases in proportion to the cell size, such that etching damage on the capacitor becomes increasingly critical. The data in Figure 13.17 show that the polarization decays as the drawn cell size decreases. With a planar capacitor structure, although the polarization degradation is negligible down to the 150-nm technology node, the polarization value decreases rapidly below that level. This effect is mainly caused by the difference between the drawn area and the effective area, and indicates that the etched slope is no longer steep enough to provide both a designed top-electrode area and sufficient spacing between adjacent bottom electrodes at the 130-nm technology node. Therefore, in order to increase the effective capacitor area below the critical cell size, both thickness scaling and the high-etched slope of the capacitor stack should be guaranteed.

In order to maximize effective capacitor area, the most important technology is to achieve the high-etched slope of the capacitors, but this is difficult because both top and bottom electrodes are usually noble metals, and the noble metal etch process has remained an unanswered question since the initial stages of FRAM development. Even until quite recently, the limitation of the capacitor etched slope was about 60∼65°, mainly owing to the loss of hard-mask from the sputtering condition of the noble metal etch. This lower capacitor slope can lead to a decrease in capacitor area of the top electrodes, or to a short circuit between the cap-to cap at the bottom electrode. However, based on some experimental findings (see Figure 13.18), new technology has been successfully developed in order to obtain a high-etched slope of about 80∼85° [28]. This new etching scheme was tested at high temperature with chlorine and fluorine chemistry, and a dual hard mask (oxide and metal). As a result, the noble metal was successfully etched with a high slope by improving the reactivity between the noble metal and etch gases, and by increasing the process temperature and reinforcing the robustness of the hard-mask.

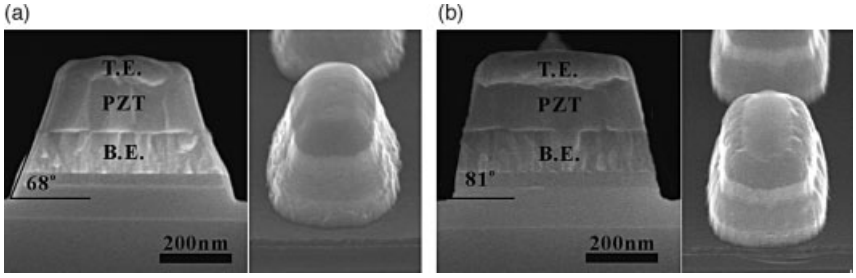


Figure 13.18 Scanning electron microscopy image showing the improvement of capacitor etch slope. (a) Normal cap etch condition; (b) enhanced cap etch condition.

13.4.2
Voltage Scaling

With the advent of the "mobile" era, low-voltage operation has become increasingly important in the reduction of power consumption. In the case of FRAM devices, the operation voltage is directly related to the thickness of the ferroelectric film; hence, the latter dimension should be minimized for low voltage application. As shown in Figure 13.19a, a PZT capacitor prepared by the CSD process shows a drastic degradation of ferroelectric properties below 100nm thickness. This is clearly a critical problem which must be solved in the case of high-density FRAM devices. As described above, both ferroelectric properties and reliability are greatly improved when the PZT films are prepared with MOCVD process; thus, even an 80 nm-thick PZT film prepared in this way demonstrates highly reliable ferroelectric properties [29] (Figure 13.19b).

It is difficult to prevent ferroelectric degradation at the interface between electrodes and ferroelectric material, even when the MOCVD process is employed. However, if perovskite oxide electrodes are used, the dead layer effect at the interface may be remarkably reduced. Recently, it has been reported that a high reliability can be achieved even with a 50 nm-thick PZT capacitor [30]. The charge-to-voltage (Q–V)

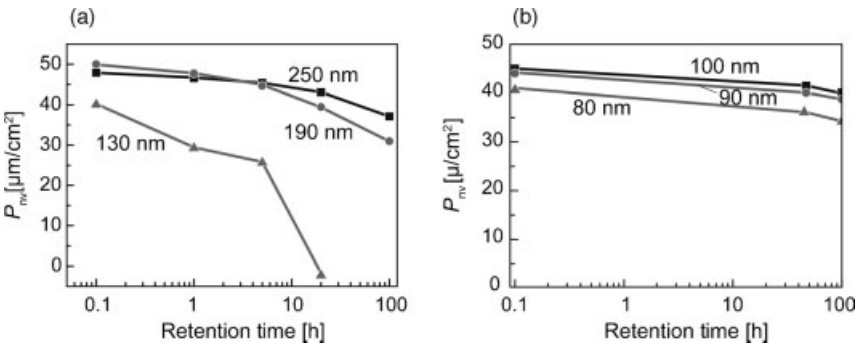


Figure 13.19 Retention properties as PZT thickness is scaled down. (a) CSD-processed PZT; (b) MOCVD-processed PZT.

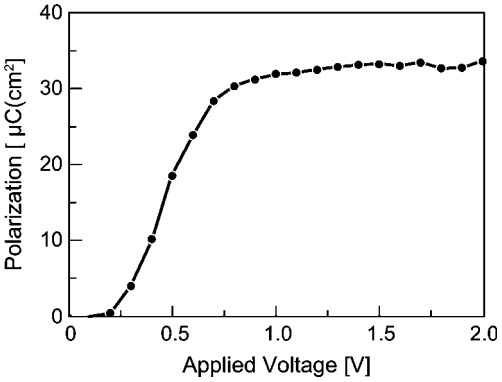


Figure 13.20 Charge–voltage (*Q*–*V*) diagram of 50 nm-thick PZT capacitor.

diagram of such as capacitor is illustrated graphically in Figure 13.20, with the capacitor being fully polarized well below an operation voltage of 1 V.

With thinner PZT films, however, several problems persist, including roughness and high leakage current. In order to overcome these difficulties, a chemical mechanical polishing (CMP) process has been introduced for PZT films. As the increase in leakage current for thin PZT film depends mainly on the surface roughness, a CMP process for PZT films can greatly reduce the leakage current [31]. The atomic force microscopy (AFM) findings and ferroelectric properties for PZT films, with or without the CMP process, are shown in Figure 13.21.

13.4.3 3-D Capacitor Structure

13.4.3.1 Limitation of Planar Capacitor

Today, many technical challenges remain to be solved for high-density planar capacitor structured FRAMs, including the limitation of capacitor stack thickness, the noble metal etch process, and thin PZT degradation. In addition, an optimum cell size is clearly required for a sufficient sensing window in FRAM devices (Figure 13.22) [32].

It can be seen from Figure 13.22 that it is difficult to achieve the 200 mV sensing margin which is required in 1T1C cell structure with sub-130 nm design rules. From this point of view, even if a thin capacitor stack and a high-etch slope were to be realized in the planar capacitor structure, it would appear difficult to embody a high-density FRAM device in excess of 256 Mb. Therefore, in order to overcome this limitation, FRAM development should ideally be pursued with a 3-D capacitor structure similar to the present-day DRAM.

13.4.3.2 Demonstration of a 3-D Capacitor

As mentioned above, the requirement for a 3-D capacitor structure is inevitable for high-density FRAM development, and the structure – together with the necessary technologies to develop a 3-D FRAM cell – are shown schematically in Figure 13.23 [33].

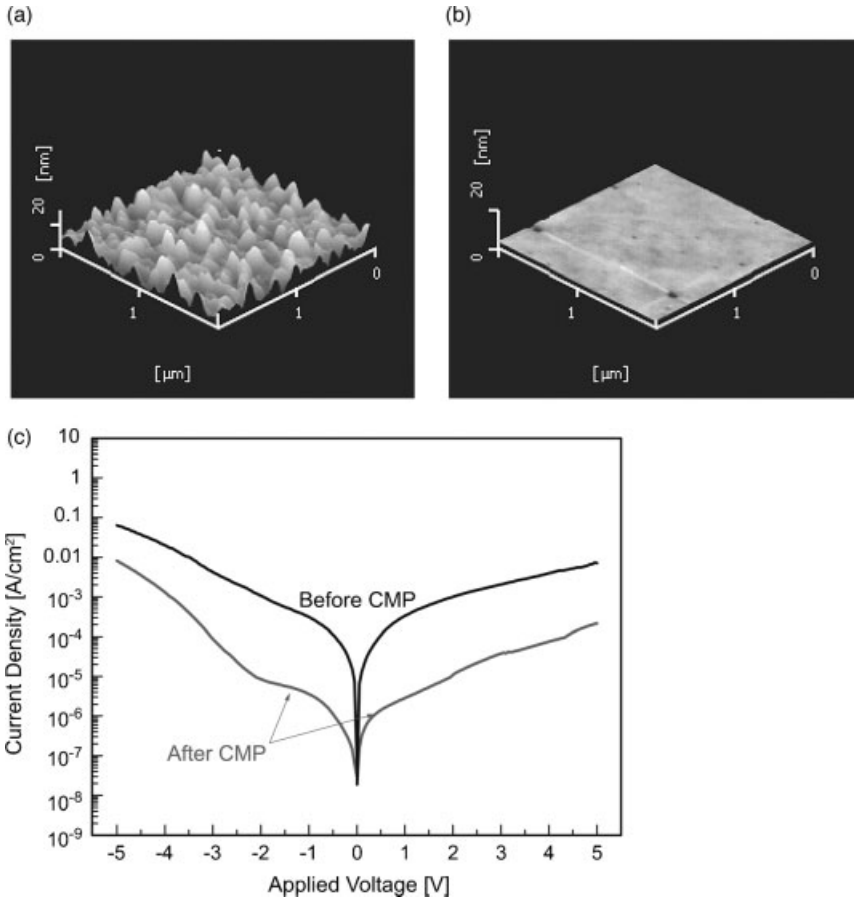


Figure 13.21 (a,b) Atomic force microscopy images and (c) leakage current characteristics of PZT films before and after CMP processing.

A prototype 3-D capacitor has recently been demonstrated, and a TEM image representing a 3-D PZT capacitor is shown in Figure 13.24. Although some pyrochlores remained in the trench capacitor, the columnar grains were well established at the side-wall of trench, with optimized deposition condition.

Figure 13.25 illustrates, graphically, the ferroelectric properties with different-sized trench structures. The polarization–voltage characteristics of a planar capacitor and trench capacitors are shown in Figure 13.25a. Under 2.1 V external bias, and an electric field of $350\,\mathrm{kV\,cm^{-1}}$, these capacitors produced no current leakage and showed quite good hysteretic behavior compared to their planar counterpart. The remnant polarization ($2P_r$) plotted against the external maximum voltage is shown in Figure 13.25b; these data showed that $2P_r$ is very similar to that for the planar capacitor in the case of a 0.32 μm trench-diameter 3-D capacitor. However, a 0.25 μm trench-diameter capacitor showed a $2P_r$ value of $19\,\mu\mathrm{C\,cm^{-2}}$ under an external

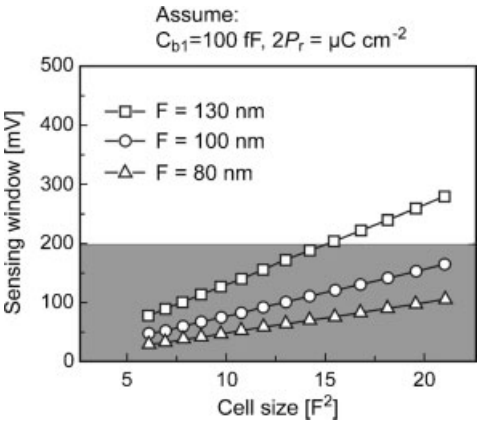


Figure 13.22 Cell size limitations of planar capacitors.

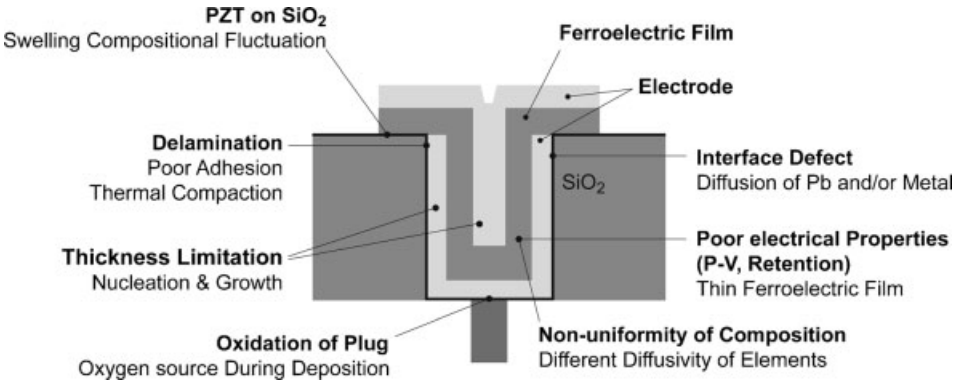


Figure 13.23 Schematic representation of 3-D capacitor structure, and the technical issues encountered.

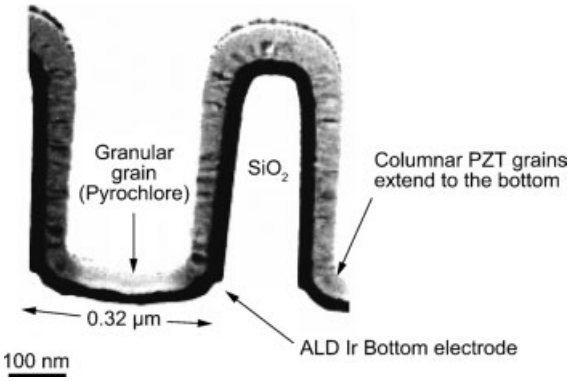


Figure 13.24 Transmission electron microscopy image of 3-D FRAM cell structure during the development of SAIT (Samsung Advanced Institute of Technology).

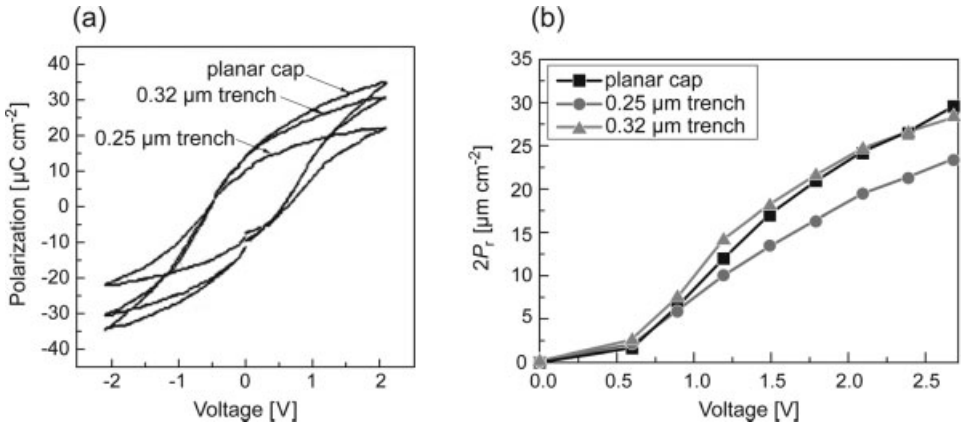


Figure 13.25 Ferroelectric properties of 3-D FRAM cell structure during the development of SAIT. (a) Polarization–voltage (*P*–*V*) loops and (b) charge–voltage (*Q*–*V*) results with different trench sizes.

maximum voltage of 2.1 V, which was 80% of the $2P_r$ values in either planar or 0.32 μm trench-diameter cases. This difference may be derived from an incomplete extension of the columnar grains on the 0.25 μm trench side-wall. Based on these findings, it is quite possible that the side-wall PZT film has the same ferroelectric properties as the planar PZT film.

In order to realize Giga-bit FRAMs with a 3-D capacitor, it has been necessary to develop the atomic layer deposition (ALD) process for the PZT and electrode material. As shown in Figure 13.23, the thickness of the ferroelectric material should be less than 50 nm because the bottom/top electrode and ferroelectric films may be formed inside a trench of 200 nm diameter. This means that the ferroelectric properties of sub-50 nm-thick PZT capacitors should be obtained for 3-D capacitor research. In addition, a step coverage of the PZT film becomes important as the aspect ratio of the capacitor increases. Because the PZT film should have a uniform composition at the bottom and side-wall, the ALD method is regarded as the best choice among other deposition methods, such as PVD and CVD. Although ALD for PZT has been investigated by many research groups, process optimization is still required. Moreover, both noble electrode metals and ferroelectric materials may be prepared using ALD. Recently, although iridium was successfully deposited using ALD, additional improvements of properties should also be investigated. In contrast, a CMP technology for noble metal electrodes may need to be introduced in order to separate each capacitor within this structure. Although noble metal CMP has not yet been achieved, it is currently undergoing extensive investigation.

Unfortunately, as of today several technical difficulties, including the reliability of the 3-D capacitor, have not been fully solved. Nonetheless, the activities of many research groups have provided much promise for 3-D FRAM development. It follows that, if some of the above-mentioned problems are solved in the near future, then the Giga-bit FRAM era will be well and truly opened.

13.5
Summary and Conclusions

FRAM technology, which has been undergoing continuous development since the early 1990s, has been used to target a universal memory in the semiconductor industry. Although reliability – notably endurance and retention – was initially a major challenge, recent findings have shown that this is no longer a key issue for FRAM devices. Rather, it is scalability which has become an important issue, following the development of 64 Mb FRAM through material and cell structural innovations. At this density, FRAM may be applied to low-density embedded memory (e.g., a smartcard), based on the demands of non-volatility, rapid access, high read/write endurance, low-power operation, and high security level. In order to produce high-density FRAM devices for use in major applications, a conventional planar-type capacitor technology is insufficient for further cell size scaling. Rather, breakthrough technologies such as the 3-D capacitor must be developed in order for the FRAM device to serve as an ideal, non-volatile memory in the future.

References

1 Kim, K.N. and Lee, S.Y. (2004) *Integrated Ferroelectrics*, **64**, 3–14.

2 Kim, K.N. (1999) *Integrated ferroelectrics*, **25**, 149–167.

3 Jeong, G.T., Hwang, Y.N., Lee, S.H., Lee, S.Y., Ryoo, K.C., Park, J.H., Song, Y.J., Ahn, S.J., Jeong, C.W., Kim, Y.T., Horii, H., Ha, Y.H., Koh, G.H., Jeong, H.S. and Kim, K.N. (2005) *IEEE International Conference on Integrated Circuit and Technology*, pp. 19–22.

4 Kim, H.J., Oh, S.C., Bae, J.S., Nam, K.T., Lee, J.E., Park, S.O., Kim, H.S., Lee, N.I., Chung, U.I., Moon, J.T. and Kang, H.K. (2005) *IEEE Transactions on Magnetics*, **41**, 2661–2663.

5 Baek, I.G., Lee, M.S., Seo, S., Lee, M.J., Seo, D.H., Suh, D.S., Park, J.C., Park, S.O., Kim, H.S., Yoo, I.K., Chung, U.I. and Moon, J.T. (2004) *IEDM Technical Digest*, pp. 587–590.

6 Ishiwara, H., Okuyama, M. and Arimoto, Y. (eds) (2004) *Ferroelectric Random Access Memories – Fundamentals and Applications*, Spinger-Verlag.

7 Rameash, R. (1997) *Thin Film Ferroelectric Materials and Devices*, Kluwer Academic Publishers.

8 Scott, J.F. and Paz De Araujo, C.A., (1989) *Science*, **246**, 1400.

9 Kim, K.N. (2001) *International Symposium on VLSI Technology*, pp. 81–84.

10 Ishiwara, H., Okuyama, M. and Arimoto, Y. (eds) (2004) *Ferroelectric Random Access Memories – Fundamentals and Applications*, Spinger-Verlag, pp. 149–163.

11 Choi, M.K. and Jeon, B.G. *et al.* (2002) *IEEE Journal of Solid-State Circuits*, **37**, 1472–1478.

12 Araujo, C.A., Cuchiaro, J.D., McMillan, L.D., Scott, M.C. and Scott, J.F. (1995) *Nature*, **374**, 627–629.

13 Park, B.H., Kang, B.S., Bu, S.D., Noh, T.W., Lee, J. and Jo, W. (1999) *Nature*, **401**, 682–684.

14 Wang, J., Neaton, J.B., Zheng, H., Nagarajan, V., Ogale, S.B., Liu, B., Viehland, D., Vaithyanathan, V., Schlom, D.G., Waghmare, U.V., Spaldin, N.A., Rabe, K.M., Wuttig, M. and Ramesh, R. (2003) *Science*, **299**, 1719–1722.

15 Joo, H.J., Song, Y.J., Kim, H.H., Kang, S.K., Park, J.H., Kang, Y.M., Kang, E.Y., Lee, S.Y., Jeong, H.S. and Kim, K.N. (2004) *International Symposium on VLSI Technology,* pp. 148–149.

16 Nakamura, T., Nakao, Y., Kamisawa, A. and Takasu, H. (1994) *Applied Physics Letters,* **65**, 1522–1524.

17 Kang, B.S., Yoon, J.G., Kim, D.J., Noh, T.W., Song, T.K., Lee, Y.K., Lee, J.K. and Park, Y.S. (2003) *Applied Physics Letters*, **82**, 2124–2126.

18 Shin, S., Hofmann, M., Lee, Y.K., Cho, C.R., Lee, J.K., Park, Y., Lee, K.M. and Song, Y.J. (2003) *Materials Research Society Symposium Proceedings*, **748**, U4.1.1–U4.1.10.

19 Bae, B.J., Lim, J.E., Yoo, D.C., Nam, S.D., Heo, J.E., Im, D.H., Cho, B.O., Park, S.O., Kim, H.S., Chung, U.I. and Moon, J.T. (2005) *Integrated Ferroelectrics,* **75**, 235–241.

20 Shimizu, M., Sugiyama, M., Fujisawa, H. and Shiosaki, T. (1994) *Japanese Journal of Applied Physics*, **33**, 5167.

21 Lee, M.S., Park, K.S., Nam, S.D., Lee, K.M., Seo, J.S., Joo, S.H., Lee, S.W., Lee, Y.T., An, H.G., Kim, H.J., Cho, S.L., Son, Y.H., Kim, Y.D., Jung, Y.J., Heo, J.E., Park, S.O., Chung, U.I. and Moon, J.T. (2002) *Japanese Journal of Applied Physics*, **41**, 6709–6713.

22 Heo, J.E., Bae, B.J., Yoo, D.C., Nam, S.D., Lim, J.E., Im, D.H., Joo, S.H., Jung, Y.J., Choi, S.H., Park, S.O., Kim, H.S., Chung, U.I. and Moon, J.T. (2006) *Japanese Journal of Applied Physics*, **45**, 3198–3201.

23 Lee, S.Y. and Kim, K.N. (2005) *International Symposium on Integrated Ferroelectrics 2005, Shanghai.*

24 Kang, S.K., Song, Y.J., Joo, H.J., Kim, H.H., Park, J.H., Kang, Y.M., Kang, E.Y., Lee, S.Y. and Kim, K.N. (2004) *Integrated Ferroelectrics*, **66**, 29–34.

25 Lee, S.Y., Kim, H.H., Jung, D.J., Song, Y.J., Jang, N.W., Choi, M.K., Jeon, B.K., Lee, Y.T., Lee, K.M., Joo, S.H., Park, S.O. and Kim, K.N. (2001) *International Symposium on VLSI Technology,* pp. 111–112.

26 Choi, D.Y., Park, J.H., Rhie, H.S., Joo, H.J., Kang, S.K., Kang, Y.M., Kim, J.H., Koo, B.J., Lee, S.Y., Jeong, H.S., and Kim, K.N. (2005) *International Symposium on Integrated Ferroelectrics 2005,* Shanghai.

27 Lee, J.K., Lee, M.S., Hong, S., Lee, W., Lee, Y.K., Shin, S. and Park, Y. (2002) *Japanese Journal of Applied Physics*, **41**, 6690–6694.

28 Ko, H.Y., Byun, K.R., Jung, Y.J., Im, D.H., Yoo, D.C., Joo, S.H., Ham, J.H., Park, S.O., Kim, H.S., Chi, K.K., Kang, C.J., Cho, H.K., Jung, U.I. and Moon, J.T. (2005) *AVS 52nd International Symposium & Exhibition,* Boston, USA.

29 Bae, B.J., Lee, K.M., Lim, J.E., Nam, S.D., Park, K.S., Yoo, D.C., Lee, C.M., Lee, M.S., Park, S.O., Kim, H.S., Chung, U.I. and Moon, J.T. (2004) *Integrated Ferroelectrics,* **68**, 123–128.

30 Yoo, D.C., Bae, B.J., Lim, J.-E., Im, D.H., Park, S.O., Kim, H.S., Chung, U.I., Moon, J.T. and Ryu, B.I. (2005) *Symposium on VLSI Technology Digest*, pp. 100–101.

31 Choi, S.H., Bae, B.J., Son, Y.H., *et al.* (2005) *Integrated Ferroelectrics*, **75**, 215–223.

32 Kang, Y.M., Kim, J.H., Joo, H.J., Kang, S.K., Rhie, H.S., Park, J.H., Choi, D.Y., Oh, S.G., Koo, B.J., Lee, S.Y., Jeong, H.S. and Kim, K.N. (2005) *Symposium on VLSI Technology Digest,* pp. 102–103.

33 Koo, J.M., Seo, B.S., Kim, S.P., Shin, S.M., Lee, J.H., Baik, H.S., Lee, J.H., Yang, M., Bae, B.J., Lim, J.E., Yoo, D.C., Park, S.O., Kim, H.S., Han, H., Baik, S., Choi, J.Y., Park, Y.J. and Park, Y. (2005) *Symposium on IEDM Technology Digest,* pp. 340–343.

14
Magnetoresistive Random Access Memory

Michael C. Gaidis

14.1
Magnetoresistive Random Access Memory (MRAM)

Through the merging of magnetics (spin) and electronics, the burgeoning field of "spintronics" has created MRAM memory with characteristics of non-volatility, high density, high endurance, radiation hardness, high-speed operation, and inexpensive complementary metal oxide–semiconductor (CMOS) integration. While MRAM is unique in combining all of the above qualities, it is not necessarily the best memory technology for any single characteristic. For example, SRAM is faster, flash is more dense, and DRAM is less expensive. Stand-alone memories are generally valued for one particular characteristic: speed, density, or economy. MRAM therefore faces difficult odds in competing against the aforementioned memories in a stand-alone application. However, embedded memory for application-specific integrated circuits or microprocessor caching often demands flexibility over narrow performance optimization. This is where MRAM excels: it can be called the "handyman of memories" for its ability to flexibly perform a variety of tasks at a relatively low cost [1]. Whilst one may hire a specialist to rewire the entire electrical circuitry of a house, or install entirely new plumbing, a handyman with a flexible toolbox is a much more reasonable option for repairing a single electrical outlet or a leaky sink. Moreover, the handyman may be able to repair a defective electrical circuit discovered while in the process of repairing leaky plumbing!

A semiconductor fabrication facility that has MRAM in its toolbox is more likely to tailor circuit designs to a customer's individual needs for optimal performance at reasonable cost. The ways in which the characteristics of MRAM compare to those of other embedded memory technologies at the relatively conservative 180 nm node are listed in Table 14.1. In the remainder of this chapter, the state of the art in MRAM technology will be reviewed: how it works; how its memory circuits are designed; how it is fabricated; the potential pitfalls; and an outlook for future use of MRAM as devices are scaled smaller.

Nanotechnology. Volume 3: Information Technology. Edited by Rainer Waser

ISBN: 978-3-527-31738-7

Table 14.1 Embedded memory comparison at the 180 nm node.

Parameter		eSRAM	eDRAM	eFlash	eMRAM
Size	Cell area (μm^2)	3.7	0.6	0.5	1.2
Size	Array efficiency	65%	40%	30%	40%
Cost	Additional process	0	20% (4 msk)	25% (8 msk)	20% (3 msk)
Speed	Read access	3.3 ns	13 ns	13 ns	15 ns
Speed	Write cycle	3.4 ns	20 ns	5000 ns	15 ns
Power	Data retention	400 µA	5000 µA	0	0
Power	Active read	15 pC b^{-1}	5.4 pC b^{-1}	28 pC b^{-1}	6.3 pC b^{-1}
Power	Active write	15 pC b^{-1}	5.4 pC b^{-1}	31 000 pC b^{-1}	44 pC b^{-1}
Endurance	Write	Unlimited	Unlimited	1e5 cycles	Unlimited
Rad Hard	—	Average	Poor	Average	Excellent

The shaded cells indicate where MRAM has a distinct advantage. Relative comparisons should hold through scaling to the 65 nm node [2].

14.2 Basic MRAM

MRAM (magnetoresistive RAM) differs from earlier incarnations of magnetic memory (magnetic RAM) in that MRAM tightly couples electronic readout with magnetic storage in a compact device structure. During the early second half of the twentieth century, the most widely used RAM was a type of magnetic RAM called *ferrite core memory*. These memories utilized tiny ferrite rings threaded by multiple wires used to generate fields to write or to sense the switching of the magnetic polarity in the rings [3]. Highly valued for its speed, reliability, and radiation hardness, approximately 400 kB of this core memory was used in early IBM model AP-101B computers on the space shuttle. However, with the advent of compact, reliable, and inexpensive semiconductor memory, the 1 mm^2 cell size of the core memory could no longer compete, and in 1990 the space shuttle converted to battery-backed semiconductor memory with around 1 MB capacity [4].

In order for magnetic memory to compete again in the RAM arena, miniaturization on the scale of semiconductor integrated circuitry had to be implemented. This was stimulated by the discovery in 1988 of giant magnetoresistance (GMR) structures which provided an elegant means of coupling a magnetic storage (spin) state with an electronic readout, thereby creating the field of spintronics [5]. Spintronics relies on the phenomenon wherein electrons in certain ferromagnetic materials will align their spins with the magnetization in the ferromagnet. In essence, this is a result of a greater electron density of states at the Fermi level for electrons with spin aligned parallel to the magnetization in the ferromagnet. The passing of a current along two ferromagnetic films in close proximity allows the transport of the electrons to be influenced by adjusting the relative orientation of the two films' magnetization. As shown in Figure 14.1, although for parallel orientation, electrons are less likely to suffer resistive spin-flip scattering events, for antiparallel orientation they will exhibit a stronger preference for scattering and thus an increase in resistance will be

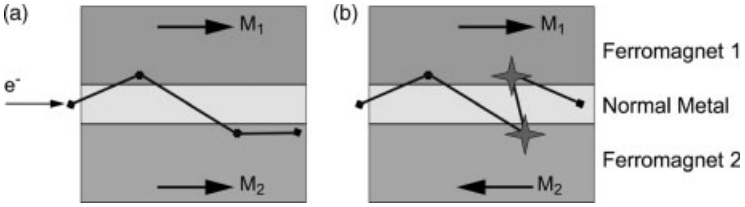


Figure 14.1 Illustration of the giant magnetoresistance (GMR) principle. For parallel alignment (a) of magnetizations M1 and M2, electron flow is subject to fewer resistive spin-flip scattering events than for antiparallel alignment (b).

apparent. The different resistance values for the high resistance state (R_{high}) and the low resistance state (R_{low}) can be used to define a magnetoresistance ratio (MR) as in Equation 14.1:

$$MR = \frac{(R_{\text{high}} - R_{\text{low}})}{R_{\text{low}}} \tag{14.1}$$

Typically, MR values for GMR devices are in the range of 5–10% for room-temperature operation.

By choosing different coercive fields for the two ferromagnets, it is possible to create a so-called *spin-valve* MRAM structure with a configuration similar to that shown in Figure 14.1. For example, ferromagnet 1 can be chosen to have a high coercivity, thus fixing its magnetization in a certain direction. Ferromagnet 2 can be chosen with a lower coercivity, allowing its magnetization direction to fluctuate. For a magnetic field sensor such as used in disk drive read heads, small changes in the magnetization angle of ferromagnet 2 induced by an external magnetic field can be sensed as changes in the resistance of the spin valve. Because the spin-valve sensitivity to external fields can be substantially better than inductive pickup, such devices have enabled dramatic shrinkage of the bit size in modern hard drives. An alternative use for the spin-valve structure is found if it is designed to utilize just two well-defined magnetization states of ferromagnet 2 (e.g., parallel or antiparallel to ferromagnet 1). Such spin-valve designs serve as a binary memory device, and have found application in rad-hard non-volatile memories as large as 1 Mb [6]. The drawbacks of this type of memory are:

- a relatively low magnetoresistance, providing only low signal amplitudes and thus longer read times
- a low device resistance, making for difficult integration with resistive CMOS transistor channels
- in-plane device formation which is more difficult to scale to small dimensions than devices formed perpendicular to the plane.

Solutions to these problems can all be found in the magnetic tunnel junction (MTJ) MRAM. The MTJ structure is similar to the GMR spin-valve in that it uses the property of electron spins aligning with the magnetic moment inside a ferromagnet. However, instead of passing current in-plane through a normal metal between

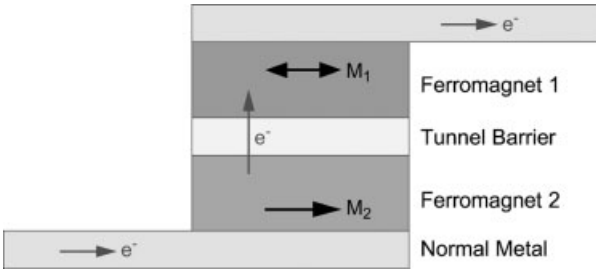


Figure 14.2 A simple magnetic tunnel junction structure. Ferromagnet 2 acts as an electron spin polarizer, and ferromagnet 1 as an electron spin filter, with magnetization either parallel or anti-parallel to the magnetization of ferromagnet 2. Parallel magnetizations generally result in a lower device resistance than anti-parallel magnetizations.

ferromagnets, the MTJ passes current perpendicular to the plane, through an insulating barrier separating two ferromagnets. An MTJ structure in its simplest form is shown in Figure 14.2. Here, one can envision the electric current impinging first on a ferromagnet which acts as a spin polarizer, then passing through the tunnel barrier and into a second ferromagnet which acts as a spin filter. The separation of polarizing and filtering functions is enabled by the physical thickness of the tunnel barrier, noting that the tunneling process preserves electron spin. The tunneling conductance will be proportional to the product of electron densities of states on each side of the barrier, and in general for ferromagnets there will be a larger density of states near the Fermi level for electrons polarized parallel to the magnetization of the ferromagnet as opposed to electrons polarized antiparallel. For polarizer and filter magnetizations aligned in the same direction, the density of states for spin-polarized electrons is large on both sides of the barrier, and the conductance of the structure is relatively high. For anti-parallel alignment of the polarizer and filter, the density of states available for spin-polarized electrons to tunnel into is somewhat reduced, and the conductance of the structure is relatively low. Proposed around 1974 [7], the first demonstrations of MTJs used Fe/Ge/Co multilayer stacks, but only showed appreciable MR (14%) at 4 K temperatures [8]. It was not until 1995 that improvements in materials processing techniques and the use of robust aluminum oxide tunnel barriers began to show reasonably large MR (18%) for MTJ devices at room temperature [9]. This breakthrough brought about huge investments from numerous companies, and ushered in a new era in the field of spintronics.

14.3 MTJ MRAM

The structure illustrated in Figure 14.2 can store binary information in the direction of magnetization within ferromagnet 1 (the “free layer”), provided that the magneti-

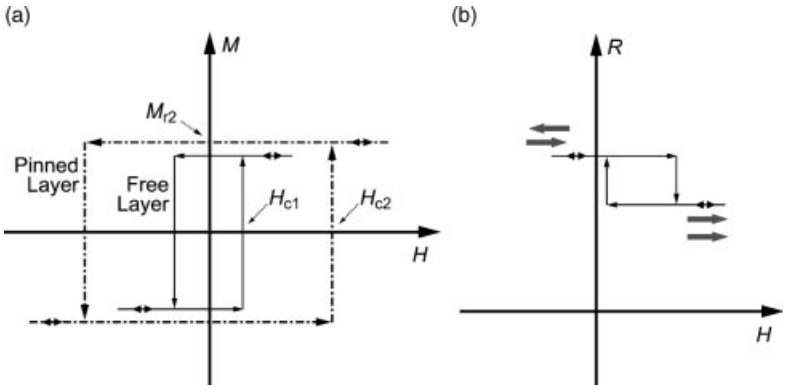


Figure 14.3 (a) Representative hysteresis curves of magnetization M versus applied field H, for a soft ferromagnet free layer and for a hard ferromagnet pinned layer in isolation. The coercive field H_{c2} is chosen large enough to keep the orientation of the pinned layer from switching while the free layer is being switched. M_{r2} represents the remanence from the pinned layer at zero applied field. (b) Resultant hysteresis of the MTJ resistance shown as a function of applied magnetic field. Due to the remanent magnetization from the pinned layer, the resistance loop is offset from the zero-applied field, and (as shown) can even result in but a single stable resistance at zero-applied field. The double arrows represent the magnetization state of the MTJ structure (anti-parallel or parallel).

zation within ferromagnet 2 (the "pinned layer") remains fixed in a predetermined direction. An asymmetry induced in the structure from device shape or intrinsic magnetic anisotropy can stabilize preferred orientations for the free layer to be one of either parallel to or anti-parallel to the pinned layer, thus maximizing the MR. A straightforward way to enable switching in the free layer without switching of the pinned layer is through the use of a material with a low coercive field H_c for the free layer, and a material with a high H_c for the pinned layer. This technique is illustrated in Figure 14.3a, with the hysteresis loops of a soft (low-H_c) free layer and a hard (high-H_c) pinned layer in isolation (i.e., not in the integrated MTJ stack structure). For operation at applied magnetic fields within the bounds set by H_c of the pinned layer, only the free layer will switch direction of magnetization. The hysteresis curve for the free layer demonstrates the necessary memory effect when the applied field is reduced to zero.

With the integrated multilayer structure of Figure 14.2, however, the hysteresis curves of the free and pinned layers in isolation are not straightforward predictors of the resistance states of the MTJ device. Because the pinned layer will maintain a remanence in a zero-applied field, there will be an offset imparted to the hysteresis loop of the free layer. (Note that there will be a similar offset of the pinned layer hysteresis loop imparted by the free layer's remanence, but for large enough H_{c2} there will be no effect on the device operation.) The effect of the pinned layer remanence on the magnetoresistive hysteresis loop R versus the applied field is illustrated graphically in Figure 14.3b. For a large remanence M_{r2}, the loop may shift so much that there is no longer a bistable memory for zero applied field. In principle, such an offset in memory product chips could be compensated by an external field

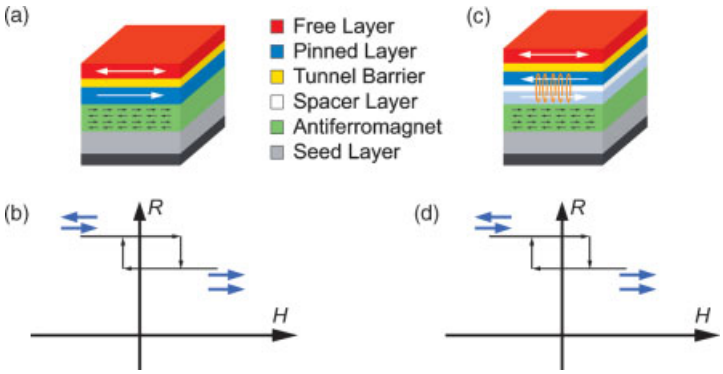


Figure 14.4 (a) Antiferromagnet-pinned reference layer structure with corresponding *R* versus *H* hysteresis loop (b). Also shown (c) is a flux-closed antiferromagnet-pinned reference layer structure with corresponding *R* versus *H* hysteresis loop (d) [2].

applied from a permanent magnet incorporated into the chip packaging. This is somewhat impractical, however, due both to packaging cost and to stringent requirements of across-chip uniformity.

Fortunately, clever manipulation of film properties has driven the evolution of several generations of MTJ structures, overcoming issues such as the offset field described above. Two such advances are illustrated in Figure 14.4. In Figure 14.4a, an antiferromagnet is exchange coupled to the pinned layer, thus providing a much larger effective coercive field for the pinned, or "reference" side of the tunnel junction [10]. With exceptional care to maintain a clean, smooth interface between the antiferromagnet and the pinned layer above it, one can obtain the strong exchange coupling between these films that is necessary to resist field switching. At least 1–1.5 nm of ferromagnetic pinned layer must still remain in the stack to act as an electron spin polarizer, but when coupled to the antiferromagnet it can be extremely well pinned even if the ferromagnet has a low H_c. By removing the need for a high-H_c ferromagnet in the pinned layer, this structure allows some additional flexibility in the choice of ferromagnet pinned layer material. One can optimize for maximum electron spin polarization for best magnetoresistance, and choose film qualities for low remanence and thus a lesser offset of the *R* versus *H* hysteresis curve. Correspondingly, Figure 14.4b illustrates a representative improvement in offset, for comparison with Figure 14.3b from the simpler stack structure.

Although there is much benefit in using the simple antiferromagnet (AF)-pinned structure of Figure 14.4a, best device operation often calls for reducing the *R* versus *H* hysteresis offset to an even smaller value. In this case, the flux-closed AF-pinned structure shown in Figure 14.4c can be tailored to give arbitrarily small offset fields. Here, a synthetic antiferromagnet (SAF) is formed from two ferromagnets separated by a thin spacer layer. For common spacer layers of 0.6–1.0 nm of Ru, one can obtain a strong antiparallel coupling between the two ferromagnets [11]. For reasonable

external fields, this coupling forces them to be antiparallel, and thus the thicknesses of the two ferromagnets can be balanced such that the external magnetic flux is negligible. The pinning of one of these ferromagnet layers with an antiferromagnet gives a high effective H_c while at the same time causing negligible offset to the R versus H hysteresis loop (Figure 14.4d).

Flux-closing the reference layer ferromagnet works remarkably well in practice, particularly with recent advances in materials deposition tooling which enable tight control over film thicknesses for multilayer film structures covering entire 200- to 300-mm wafers [13]. A cross-section transmission electron microscopy (TEM) image of such a flux-closed reference layer MTJ stack is shown in Figure 14.5. Some interesting features of the magnetics-related elements can be discerned from the TEM image, and these are discussed below.

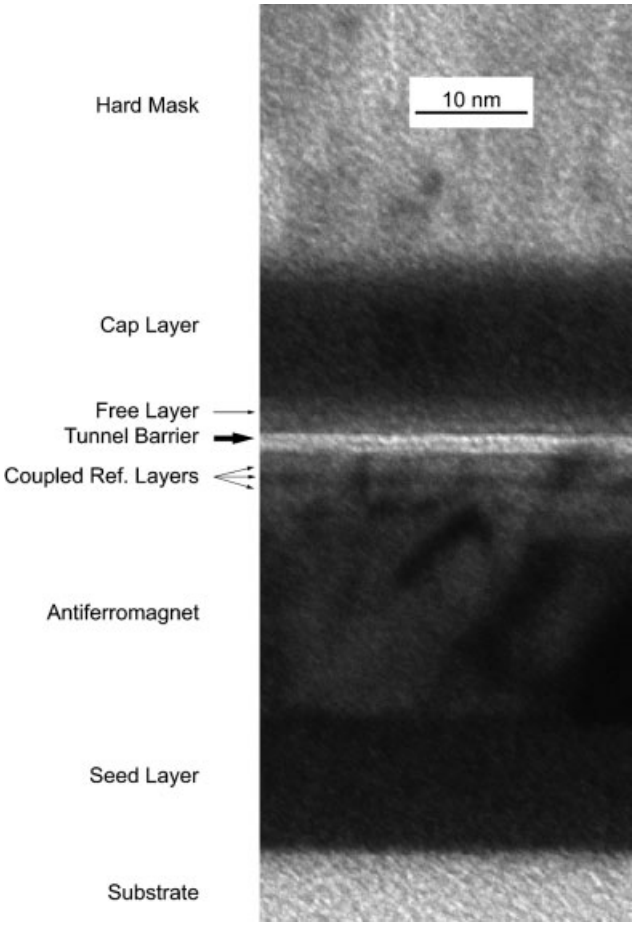


Figure 14.5 A transmission electron microscopy high-resolution, cross-sectional image of a MTJ stack with flux-closed, antiferromagnet-pinned reference layers.

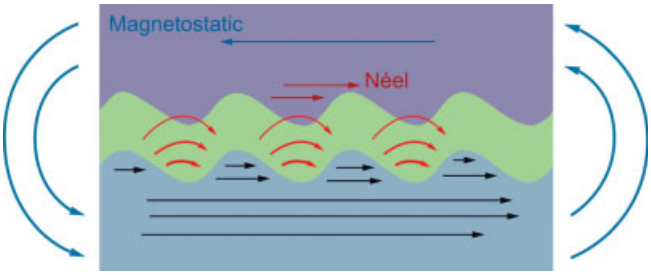


Figure 14.6 A schematic description of Néel coupling and how it relates to magnetostatic coupling. The rough-topped bottom film represents the pinned layer of Figure 14.4. Although exaggerated in the figure for clarity, an actual roughness greater than one atomic monolayer is cause for concern. The green intermediate layer represents the tunnel barrier, and the layer above is the free layer. Black arrows in the bottom film represent the internal magnetization of the pinned layer but, due to the rough surface, the magnetic poles are uncompensated in the region of the tunnel barrier. The resultant field from these poles creates a Néel field which favors parallel orientation of the free and pinned layers. The magnetostatic demagnetization field from the ends of the pinned layer favors antiparallel orientation of the free and pinned layers, but as this is non-local, it is less important in breaking the symmetry of devices with multiple layers [12].

14.3.1 Antiferromagnet

The antiferromagnet, which is generally a polycrystalline material such as FeMn, PtMn, or IrMn, is chosen and grown with several characteristics in mind:

- The interface roughness of the antiferromagnet must be sufficiently small that Néel coupling can be neglected (Figure 14.6), ensuring a smooth, pinhole-free tunnel barrier.
- The pinning strength must be large compared to the fields used to switch the free layer between its binary memory states.
- The blocking temperature of the antiferromagnet must be in a suitable range. In order to obtain an ideal pinning of the ferromagnet reference layer, the antiferromagnet/ferromagnet bilayer must be annealed above the blocking temperature T_B at which the exchange coupling between the films is zero. An applied magnetic field fixes the orientation of the ferromagnet, and then the bilayer is cooled. During cooling, the surface magnetization of the antiferromagnet aligns with the field-imposed ferromagnet magnetization. After cooling and removal of the field, exchange coupling across this interface keeps the ferromagnet pinned. Here, an antiferromagnet must be chosen with a blocking temperature T_B below approximately 300 °C in order to minimize material diffusion and tunnel barrier degradation. In addition, T_B must be sufficiently above the device operating temperatures, around 125 °C.
- The antiferromagnet must be able to withstand process temperatures of the ensuing circuit integration. Roughly, this translates into saying that the compo-

nents of the antiferromagnet should not dissociate and diffuse out of the layer for process temperatures below about 250 °C.

14.3.2 Reference Layer

The reference layer closest to the tunnel barrier must act as an effective spin polarizer, and so it must be of thickness at least of order the electron spin-flip scattering length. This implies that 1–1.5 nm is the minimum thickness of the layer closest to the tunnel barrier. For best flux closure and minimal offset to the free layer, the reference layer adjacent to the antiferromagnet will be of a similar thickness, although a perfect zero free-layer offset may dictate small differences in the thicknesses. An upper limit to the thickness is set by the additional surface roughening and resultant Néel coupling that thicker films will generate. Reference layer materials are chosen for their best spin polarization properties and compatibility with device-processing techniques (e.g., minimal corrosion and thermal stability). Films of CoFe of the order of 2 nm thickness are typically used, separated by the 0.6- to 1.0-nm exchange-coupling Ru layer.

14.3.3 Tunnel Barrier

Aside from the requirement of reasonable magnetoresistive properties, the tunnel barrier is chosen primarily for robustness. It must be extremely thin to ensure that spin polarization is maintained during electron transit across the barrier, and the barrier must be able to survive under billions of cycles of electrical bias during its lifetime, without developing pinholes or any substantial shift in resistance. Aluminum oxide has proven an extremely suitable candidate for such tunnel barriers, and is known to offer reasonable magnetoresistance for suitable magnetic pinned and free layers. Recent developments in tunnel barrier engineering show that magnesium oxide tunnel barriers can offer MR near 500% at room temperature, although MgO-barrier devices have not yet proven to serve as robust, manufacturable layers in large arrays with good magnetic switching characteristics [14]. Aluminum oxide barrier devices can display MR near 100%, but trade-offs in the choice of magnetic materials for best switching characteristics, and in the choice of operating point for best CMOS integration, generally result in an MR less than 50%. Such MR is suitable for maintaining distinct resistance groupings of millions of devices in modern MRAM arrays, and increasing the MR is advantageous primarily in that it can reduce the necessary signal integration time to read the state of a device. Such a reduction is not a terribly strong driver at this time, as the array read time is set as much by the circuit overhead as by the device signal-to-noise ratio. Increasing the MR to 500% would likely result in only a 10–20% reduction in read duration. One area in which MgO barriers may soon establish a strong foothold is in the formation of highly transparent tunnel barriers. As device sizes shrink, the lower resistance–area product afforded by MgO will enable the best match to CMOS drive transistors, and thus the highest

speed of operation. Today, even more highly transparent tunnel barriers are under development for a class of devices using electron spin current to switch the device state.

14.3.4 Free Layer

The free layer shown in the TEM is reasonably thin, rather like the underlying pinned layers. However, it does have a minimum thickness limit set by the spin filtering characteristics: for a thickness less than the approximate electron spin-flip scattering length, the magnetoresistance will begin to drop, and this again sets the thickness at around 1.5 nm or more. Thicker free layers require additional energy to switch, and so are undesirable for low-power operation. Of critical importance in the characteristics of the free layer is the need for well-defined magnetic states and well-behaved magnetic switching. As one cannot tailor the read or write circuitry to every individual device in megabit arrays of MRAM devices, it is critical that each device behave very much like all others in the array. Ill-defined magnetization states such as vortices, S-shapes, C-shapes, and multiple domains will add variability to the resistance measured by the circuitry, because electron spin polarization filtering may not be strictly parallel or antiparallel to the spin polarization imparted by the pinned layer. In addition, sensitivity of the film switching behavior to tunnel barrier and cap materials, or to device edge roughness or chemistry, can impart variability to the write operation of the individual bits in megabit arrays. NiFe alloys are preferred for good magnetic behavior with reasonable corrosion resistance. The addition of Co or Fe to the NiFe, or dusting with Co or Fe between the tunnel barrier and NiFe layer, can help to adjust the magnetic anisotropy and improve the MR. Layer thicknesses are typically in the 2- to 6-nm range for best low-power operation with good switching characteristics.

Several additional non-magnetic elements are visible in the TEM image, and these are discussed below.

14.3.5 Substrate

An ultra-smooth substrate is required as the starting point for smooth, uniform, and reliable tunnel barriers. Rough interfaces also result in increased Néel coupling, which is detrimental to device performance. Representative materials for the substrate are thermally oxidized silicon, or chemical-mechanical planarized (CMP) polished dielectrics such as silicon nitride, silicon oxide, or silicon carbide.

14.3.6 Seed Layer

An appropriate seed layer is required to obtain good growth conditions for the antiferromagnet, both to ensure a smooth top surface and to ensure good magnetic

pinning strength. Given the high stress in some of the films in the MTJ stack, this seed layer is also critical for ensuring good adhesion to the substrate. It may be formed from tantalum nitride or permalloy (NiFe), for example.

14.3.7
Cap Layer

The proper choice of a cap layer is necessary to protect the free layer during further device fabrication processing. It is essential as a barrier or "getter" for contaminants, keeping the free layer clean and magnetically well behaved. Often-used materials for this layer include ruthenium, tantalum, and aluminum. The choice of this material may also depend on its effect on the magnetic behavior of the free layer: certain cap materials can discourage smooth switching between free layer states, and can result in substantial "dead layers" which must be compensated for by a thicker free layer.

14.3.8
Hard Mask

A hard mask (as opposed to a "soft" photoresist mask) is used to enable patterning of the MTJ with industry-standard etch techniques. It also eases integration with the surrounding circuitry by providing a contact layer to connect the MTJ to wiring levels above. The hard mask material is largely chosen for its compatibility with subsequent processing in the fabrication route, and can be chosen from any number of metallic or dielectric materials.

The processing of MTJ structures to integrate them with CMOS circuitry is discussed in greater detail later in the chapter.

14.4
MRAM Cell Structure and Circuit Design

14.4.1
Writing the Bits

The mechanism for switching the state of the free layer in MRAM lends itself well to an array layout with a conventional planar semiconductor design and fabrication. A typical rectangular MTJ array layout, with word lines (WLs) arrayed beneath the devices and bit lines (BLs) arrayed atop the devices, is illustrated in Figure 14.7. Current driven along the WLs or BLs generates a magnetic field which imparts a torque on the magnetization of the device. In normal operation, the superposition of properly-sized "write" fields from both WL and BL will enable a switching event to occur in the free layer of the device at the intersection of the two lines. The write fields are chosen small enough so as not to exceed the coercivity of the pinned layer. Potential pitfalls from this scheme include write errors from half-selected devices (i.e., those subjected to only a WL or a BL field, but not both) and, worse, write errors

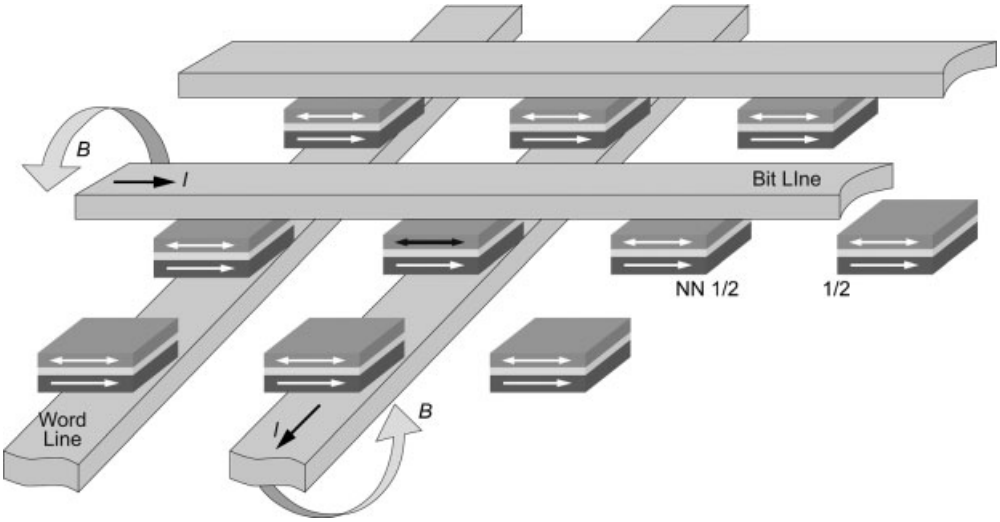


Figure 14.7 Schematic representation of a rectangular array of MTJ devices, with bit line and word line circuitry for writing the bits. Current-generated magnetic fields (B) from a given bit line and word line are sufficient only to switch the device at the intersection of the two wires. Write errors are typically worse for devices in the half-select state (MTJs labeled "1/2" in the figure), where a word line or a bit line is active, but not both. The situation is even worse for near-neighbor half-selected devices ("NN1/2" in the figure) where, for example, the device is in the column adjacent to the active word line, but is half-selected by the active bit line.

from near-neighbor half-selected devices (those subjected to a half-select field, but only one row or column away from another active line).

The diagrams in Figure 14.8 provide more details about the superposition of magnetic fields used to switch the active device. With the flux-closed antiferromagnet-pinned reference layer structure (see Figure 14.4c) forming the MTJ, the single-layer free layer is switched with characteristics first described by Stoner and Wohlfarth [15]. A simple case is that of an elliptical-shaped MTJ with shape anisotropy defining an easy axis (the major axis of the ellipse) and a hard axis (the minor axis of the ellipse). To switch the magnetization, a hard-axis field is applied to tilt the free layer magnetization away from the easy axis energy minimum, and an easy-axis field is applied to "set" the magnetization of the device in the desired easy-axis direction – parallel or antiparallel to the pinned layer. With this Stoner–Wohlfarth (S-W) switching, relatively small operating margins are illustrated by the closeness of the green and pink dots to the S–W boundary in Figure 14.8b. In addition to accounting for spreads in the switching characteristics between devices, one must also budget in extra operating window for thermal activation errors and the disturb effects of half-selects and near-neighbor field interaction. Circuit designers will try to tailor the operating window for at least 10 years of error-free operation. Without use of error-correction techniques, one generally aims for operating margins to keep the activation energy for a bit error to greater than 60 $k_B T$, where k_B is the Boltzmann constant and T is the temperature. This imposes extremely tight requirements on how uniform the array

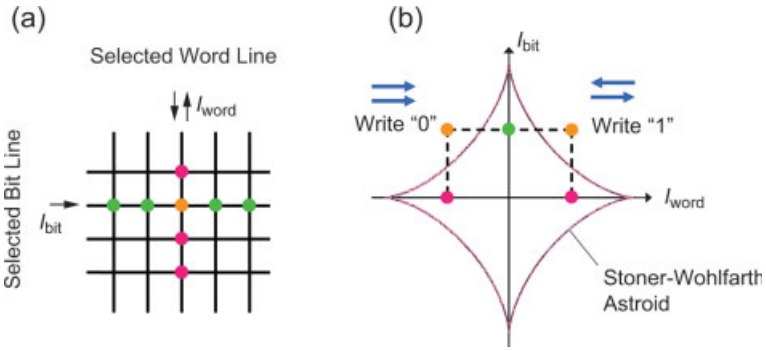


Figure 14.8 (a) Top-down schematic view of an MTJ array with rows and columns of bit lines and word lines with fields superposed to switch the device represented by a orange dot. Devices shown as green and pink dots are half-selected devices. Those green and pink devices adjacent to the orange device are near-neighbor half-selected devices. (b) A graph showing necessary bit line and word line current values needed to switch a desired device. The colored dots on the plot correspond to the devices represented by colored dots in Figure 14.8a. For suitable choice of word line and bit line currents, one can ensure switching of the desired device without switching half-selected devices.

must be in terms of switching, described in equation form by the array quality factor (AQF):

$$\mathrm{AQF} = \frac{H_{\mathrm{sw}}}{\sigma_{\mathrm{Hsw}}} \tag{14.2}$$

Here, H_{sw} is the average switching field of the devices and σ_{Hsw} is the standard deviation of the switching field distribution of all elements in the array. In rough terms, the AQF must be larger than about 30 in order to ensure a lifetime of 10 years, although some relief can be gained through the use of error correction techniques.

Toggle MRAM was invented to circumvent the difficulties faced by S–W MRAM in terms of the operating margin for half-selected bits [16]. As illustrated in Figure 14.9a, the structure has taken the flux-closed antiferromagnet-pinned reference layer structure (see Figure 14.4c) a step further by also flux-closing the ferromagnetic free layer. This is achieved by depositing a spacer layer atop the free layer ferromagnet, followed by a second ferromagnet. The spacer can be chosen (as in the pinned layer) to enhance antiparallel coupling, or the spacer can be chosen with zero or even with some parallel coupling characteristics to decrease the write field needed to switch the bit. The magnetizations of the two ferromagnets in the free layer will point in opposite directions, and their balance and proximity will flux-close the layers so there is little field seen emanating from the structure at a distance. The write operation of this toggle-mode structure is illustrated in Figure 14.9b. Noting the colors assigned to represent the magnetization of the free layers in Figure 14.9a (green for the top layer, red for the bottom layer), the plots at the top of Figure 14.9b show the relative orientation of the two magnetizations. Note that the initial state is such that the magnetization of the MTJ has easy (preferred) axis at 45° to the word and bit lines, rather than be aligned parallel to one of them as in S–W MRAM.

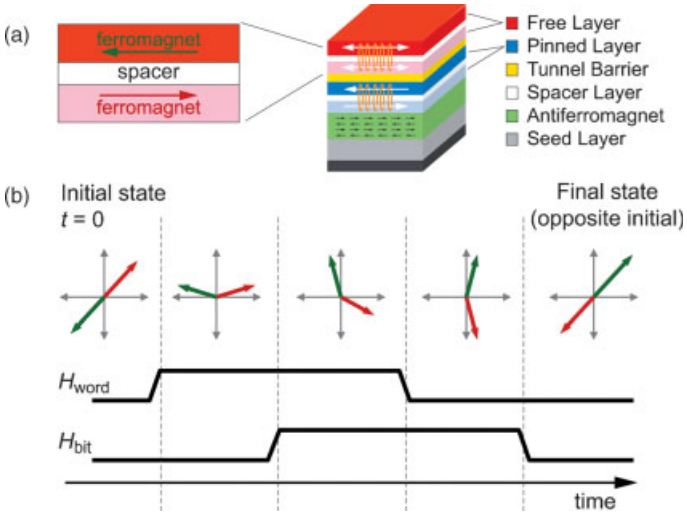


Figure 14.9 (a) Structure of the toggle-mode MTJ stack. (b) Time evolution of the free layer switching. See text for details.

Figure 14.9b illustrates the need for staggered timing of WL and BL write-field pulses. To switch the state of the free layers, a magnetic field is first applied from the WL along the positive y-direction. This magnetic field cants the magnetizations of both free layers as they try to align to the field. The antiparallel nature of the magnetic coupling between the free layers prevents the magnetizations from both fully lining up with the applied word field, as long as the field is not too large to overwhelm this antiparallel state. When the magnetizations are canted sufficiently, there is a net magnetic moment to the free layers, and this moment can be grabbed like a handle by the field now imparted by the BL. The BL applies a field in the positive x-direction, and the net moment of the two free layers follows this BL field. The WL field is then shut off, and the net moment continues to rotate around towards the applied BL field. As the BL field is shut off, the free layer magnetizations relax into their energetically favorable antiparallel configuration, but now with magnetizations exactly opposite to those at the start.

The name "toggle-mode device" is derived from the characteristic that cycling the WLs and BLs in this manner will always switch the state of the device. To set a bit in a particular state, a read operation must be performed to determine if a write "toggle" operation is required. Aside from this drawback, and the additional complexity of the magnetic stack, there are several advantages to the toggle-mode structure:

- As alluded to above, the write operating margins can be substantially larger than for devices with S–W switching. Rather than a S–W astroid boundary, the toggle-mode devices exhibit an L-shaped boundary that does not approach the WL or BL axes. The potential for half-select errors is dramatically reduced, and the requirement on AQF is approximately halved.

- In principle, shape anisotropy is not required to ensure that the bit has only two preferred states for binary memory. One can utilize the intrinsic anisotropy of the ferromagnetic free layers to define two such states. This allows the use of circular MTJ devices for the smallest memory cell size.
- The flux-closed nature of the free layers greatly reduces dipole fields emanating from the free layer. Such fields can affect the energetics of nearby devices, resulting in variability of switching characteristics, depending on the states of such devices. Thus, with flux-closed free layers, nearby devices can be packed in closer proximity for improved scaling.

14.4.2 Reading the Bits

The array structure illustrated in Figure 14.7 is often termed a "cross-point cell" (XPC) structure. More specifically, XPC refers to the case where the MTJ devices are located at the cross-points of the BLs and WLs, and are directly connected to the BLs and WLs above and below the MTJ stack. This structure offers an extremely high packing density for the lowest cost memory. The write mechanism is reasonably straightforward as described above, as long as the MTJ resistance is not so low that it shunts the write currents. More troublesome is that the read mechanism suffers from a reduced signal-to-noise ratio in this XPC structure. In order to read the resistance state of a XPC bit, a bias is applied between a desired BL and WL, and the resistance measured. However, due to the interconnected nature of the XPC structure, not only the resistance of the cross-point device is measured – there are parallel contributions of resistance from many other devices along "sneak paths" that include traversing additional sections of BL and WL. Due to the resulting loss of signal, the device must be read much more slowly to allow for integration to improve the signal-to-noise ratio. Device read times can be substantially longer for such XPC structures, making this type of memory far less desirable than one which can be read as fast as DRAM, for example.

The solution to the problem of sneak paths is to insert an isolation mechanism which ensures that read currents will only traverse a single MTJ device. For example, this can be achieved by placing a diode in series with each MTJ. Although this seems simple when drawn as a circuit schematic on paper, it is actually more straightforward to place a field effect transistor (FET) in series with each MTJ, and assign a second WL to control the read operation. The "FET cell" circuit structure is shown schematically in Figure 14.10, with separate WLs for the write and read operations. The BL is used for both read and write operations.

Figure 14.11 illustrates the implementation of the circuit structure shown in Figure 14.10, suitable for a densely packed array of MTJs. Structural additions to standard CMOS circuitry include:

- the via contact VJ between the bit line and the top of the MTJ stack
- the MTJ device

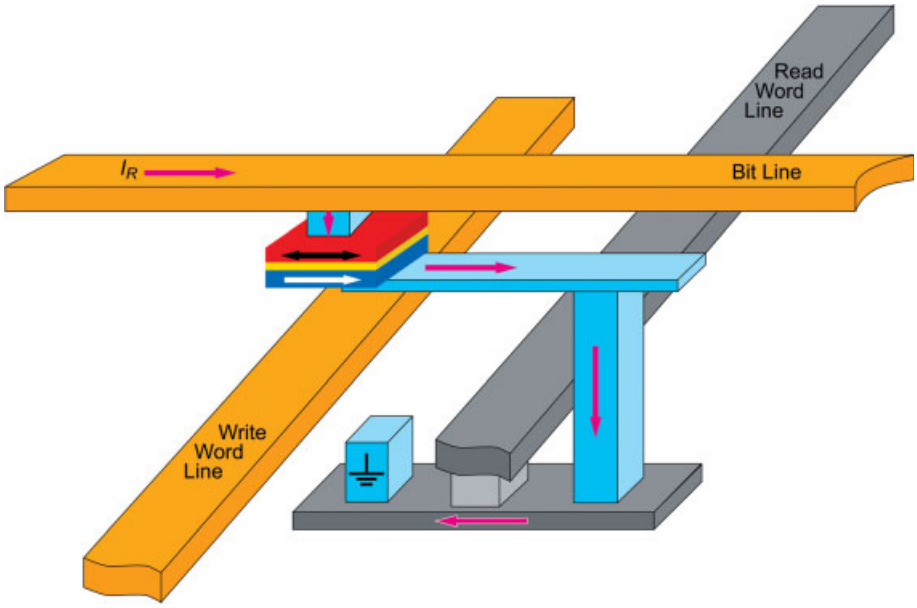


Figure 14.10 Field-effect transistor (FET) cell circuit topology, showing individual word lines for reading and for writing. A FET located in the silicon beneath the MTJ is used to switch on only the device being read, thus preventing leakage of read currents (purple arrows) through nearby MTJ devices. Additional conductor elements in this structure (compared to Figure 14.7) include a contact between the bit line and the top of the MTJ, a local metal strap (MA) connecting the base of the MTJ with a via chain that connects to the underlying FET.

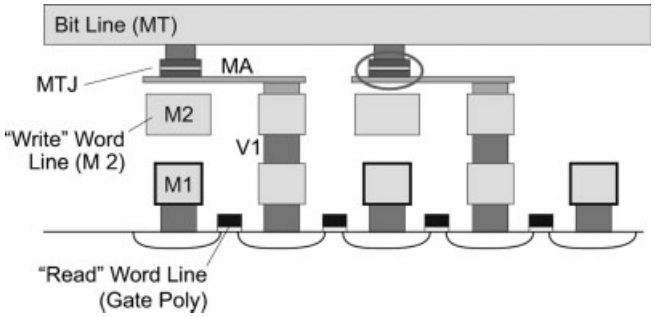


Figure 14.11 A cross-section of the FET cell topology, with two adjacent cells shown atop the silicon CMOS front-end of line (FEOL) structure. The oval encloses the critical components for MRAM implementation. As cell size is determined primarily by the MTJ and via chain above the via V1, two FETs can be used for each MTJ in order to achieve lower resistance and some redundancy. Thus, the FET gates on either side of a V1 via chain will be connected to the same "read" word line. Wires formed in the first level of metallization (M1) (outlined in bold) form a grid at a reference potential. M2 denotes the second level of metallization. The reader is referred to Ref. [17] for further details on such structures.

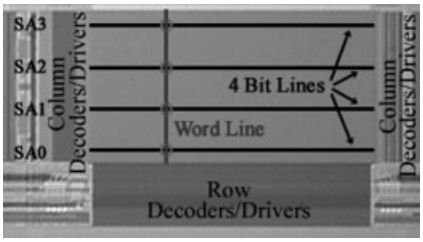


Figure 14.12 Photograph of a 128 Kb subarray, showing locations of the sense amplifiers (SA), the row and column decoders and drivers, and the concurrent activation of four bit lines with one word line for a ×4 organization of the block. A single MTJ cell is indicated by a circle at the intersection of a word line (WL) and bit line (BL).

- the local metal strap (MA) between the bottom of the MTJ stack and the via to M2
- the via VA between the MA strap and the M2 wiring, which serves to isolate the MTJ from the write WL while providing connection to the underlying FET structure for reading.

A slightly higher packing density may be achieved with a mirror-cell design, where adjacent bits mirror each other. The simple unmirrored design of Figure 14.11 is preferable to minimize any across-array non-uniformity due to inter-level misalignment and inter-cell magnetic interference. Megabit and larger MRAM memories are formed from multiple subarrays, with size determined largely by the resistance of the BLs and WLs. There is a desire always to keep applied voltage low, for CMOS compatibility and best array efficiency. The required current to generate the necessary MTJ switching fields then sets a maximum length on the BL or WL, depending on the resistive voltage drop. Bootstrapped write drivers can be used to allow smaller write drivers with improved write current control [18]. A 16 Mb MRAM under development at IBM utilizes 128 Kb subarrays (see Figure 14.12), with 512 WLs and 256 BLs of active memory elements.

The read operation is performed with sense amplifiers that compare the desired bit to a reference cell. The reference cell uses two adjacent MTJs fixed in opposite states in a configuration that acts like an ideal mid-point reference between the R_{high} and the R_{low} states [18]. Four BLs are activated in a given cycle, and are uniformly spaced along the height of the array to reduce magnetic interference between activated BLs during the write operation, and to minimize distance from the activated BLs to the sense amplifiers during a read operation. Additional reference BLs are located within the array, with one set shared by sense amplifiers 0 and 1, and one set shared by sense amplifiers 2 and 3.

The array driving circuitry for MRAM memories is commonly standardized to an asynchronous SRAM-like interface for easy interchangeability in battery-backed SRAM applications. The IBM 16 Mb chip uses a ×16 architecture that is prevalent in mobile and handheld applications with packaging intended for simple direct replacement of SRAM chips. As shown in Figure 14.13, the 16 Mb chip measures 79 mm^2 with individual memory cells of 1.42 μm^2, for an array efficiency of almost

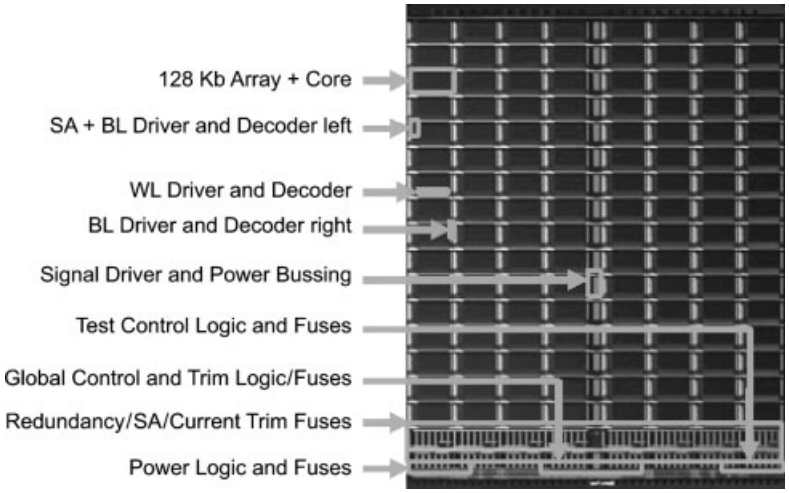


Figure 14.13 A photograph of the 16 Mb MRAM chip, showing locations of 128 Kb array cores (eight columns of 16 rows) and support circuitry [18].

30%. The array efficiency may be improved by using more metal layers and by eliminating some of the developmental test mode structures used in this chip. A reduction of the standby current for power-critical applications is achieved through the extensive use of high threshold, long-channel FET devices and careful grounding of inactive terminals in the arrays and in the write driver devices [18].

Redundant elements are included in the chip to allow the correction of defective array elements. Such redundancy is implemented with fuse latches and address comparators in a manner consistent with industry-standard memory products. The CMOS base technology is quite mature, so the focus of the redundancy is on the MRAM features. Single-cell failures or partial WL failures (from MRAM reference cell defects) are considered the most likely defects. The redundancy architecture favors replacement of WLs to capture the partial WL fails from MRAM reference cell defects. Redundancy domains are implemented at a high level in the block hierarchy so as to span several blocks and be capable of effectively fixing any random defects [18].

14.4.3 MRAM Processing Technology and Integration

The implementation of MRAM hinges on complex magnetic film stacks and several critical steps in back-end-of-line (BEOL) processing. Cell size is presently limited by the size of the MTJ devices and driving wires, and older, mature CMOS front-end-of-line (FEOL) technology can be used without limiting performance. Fabrication of the FET-cell circuit, from the CMOS FEOL through the MRAM BEOL, can encompass several hundred process steps, resulting in the fully functional structure shown in Figure 14.14. The MRAM-critical portion of the circuit is a relatively small part of the entire configuration. After the last standard CMOS step (the M2 wire completion),

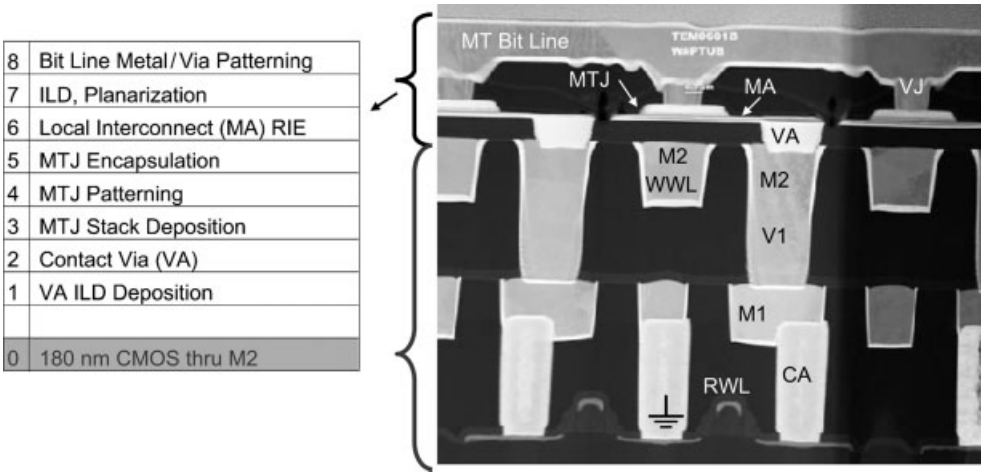


Figure 14.14 Cross-section of a product cell, showing the integration of MRAM with CMOS, and the process steps used for the MRAM-specific layers. ILD is the interlayer dielectric.

there remains the need to pattern the shallow vias, the MTJs, the local interconnects, and at least one level of wiring with contact to MTJs and the functional circuitry below. Even for simple functional circuits, five or more photomask levels are required to complete the MRAM-centric portion of the structure.

14.4.3.1 Process Steps

In conjunction with the steps outlined in Figure 14.14, below is a discussion of the important considerations for the process steps in the fabrication of the MRAM-specific levels.

1. *VA contact via and ILD:* The VA via provides a path for read current to flow from the local (MA) metal strap down through a via chain to the underlying read transistor. The most critical aspect of this module is that it must form a substrate which is sufficiently smooth for good magnetic stack growth.

2. *Magnetic film stack deposition:* Arguably the most essential technological advance in enabling MTJ MRAM was the development of tooling for the large-area deposition of extremely uniform films with well-controlled thickness. Such tooling has proven suitable for the deposition of magnetic, spacer, and tunnel barrier films with sub-Ångström uniformity across 200 mm and even 300 mm wafers [13]. The critical aluminum oxide tunnel barrier is generally formed by depositing a thin aluminum layer, followed by exposure to an oxidizing plasma [19].

3. *Tunnel junction patterning:* A commonly used and straightforward approach to patterning the MTJs is with the use of a conducting hard mask. This is later utilized as a self-aligned stud bridging the conductive MT wiring to the active magnetic films in the device. A thick hard mask, however, introduces additional

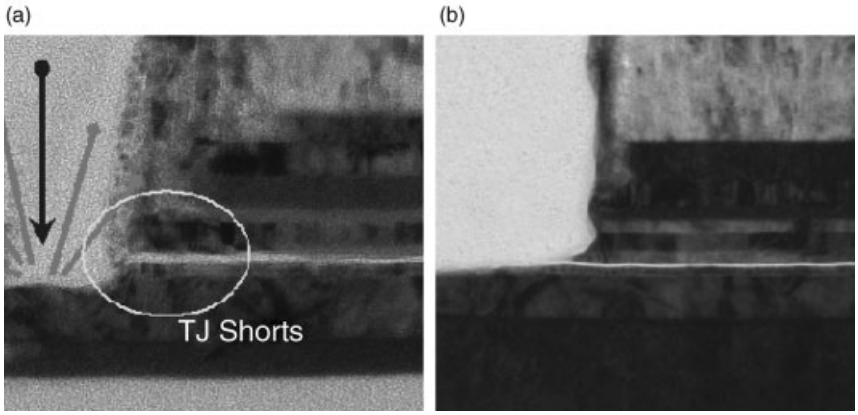


Figure 14.15 (a) Transmission electron microscopy image of the edge of a MTJ after etching to define the free magnetic layers. The etch has progressed to a depth just past the oxide tunnel barrier (the lightest contrast film in the stack). The dark arrow represents incoming sputtering ions; the lighter lines represent the path of atoms sputtered from the surface of the device being etched, many of which result in redeposits on vertical device surfaces. The consequences of sputter-etch redeposits on the sidewall of a MTJ device can be seen as a short-circuited tunnel barrier and a poorly defined edge with thick redeposits. (b) Improvements in the etch conditions can result in a much cleaner sidewall and the elimination of residues that would short-circuit the tunnel junction.

difficulties in that it can shadow the etch being used to pattern the magnetic devices. Such shadowing can add an element of variability into the size of the devices, and may also result in metal redeposits on the sidewall of the device structure. As illustrated in Figure 14.15, sidewall redeposits are particularly troublesome for commonly used MRAM stack materials because the materials do not readily form volatile RIE byproducts that provide some isotropic character to the etch. Directional physical sputtering is the main mechanism for etching of the stack materials [20]. Because the difficulties in etching the magnetic stack materials often outweigh the benefits of a simpler process integration scheme, it is often preferable to use a thinner hard mask for less etch shadowing, and an additional via level (VJ in Figure 14.14) to connect the top of the MTJ with the bit line wiring.

4. *MTJ encapsulation:* Silicon nitride and similar compounds are desirable for their adhesion to the MA and MTJ metal surfaces, and for strong interfacial bonds that inhibit migration of metal atoms along the dielectric/metal interfaces. Such metal migration is one well-documented cause of MTJ thermal degradation, and can limit processing temperatures in patterned MTJ devices to below 300 °C [21]. The use of tetra-ethyl-orthosilicate (TEOS) as a precursor in the deposition of silicon oxide films [22] is known to offer the benefits of a relatively inert depositing species which can readily diffuse into spaces adjacent to high-aspect ratio structures, even at temperatures below 250 °C.

5. *MA patterning:* For suitable thickness of seed and reference layers, the series resistance of layers remaining after MTJ etch is small enough to impart negligible

dilution of the MTJ MR signal. This simplifies the processing, as a dedicated film need not be created for the MA strap, and the reference or seed layers of the magnetic stack can perform double duty. As in the MTJ etch, the MA etch may be subject to the problem of non-volatile etch byproducts redepositing along the hard mask sidewalls.

6. *ILD deposition and planarization and wiring:* After the MA metal strap has been patterned, an interlayer dielectric is deposited in which to house the counter-electrode wiring layers VJ and MT. The counterelectrode wiring is formed with well-established semiconductor-industry Damascene techniques.

As alluded to in Figure 14.11, the MRAM-specific elements form but a small portion of the entire integrated circuit. For rapid characterization of these MRAM-specific elements, there is no need to perform a fully CMOS-integrated wafer build; rather, it is sufficient to perform a subset of the process integration steps to focus only on the critical magnetics issues [23].

14.5 MRAM Reliability

One of the strong selling points of MRAM is its reliability: write endurance is expected to be essentially infinite, the magnetics are intrinsically rad-hard, and its non-volatile memory storage can eliminate soft errors in many applications. As in any new technology struggling for successful commercialization, there are certain aspects of the new technology that are unproven and require demonstration of reliability. Areas of potential reliability risk include [24].

14.5.1 Electromigration

Electromigration in the write WLs and BLs, resulting from high write current density. Current pulses of 10 mA are typical for conservative wire cross-sections of 0.2 μm^2, corresponding to a current density of 5 MA cm^{-2}. This alone represents a serious challenge to the reliability in the array, and can potentially be worsened by local disruptions to the quality and thickness of wire material. The VJ vias of Figure 14.14, or direct connection between the BL and the MTJ hard mask in the thick hard mask integration scheme discussed above, can impact the BL wiring electromigration resistance.

Electromigration issues can potentially be improved through the use of bidirectional switching currents, which fit neatly into toggle-mode MRAM operation, but cost in terms of array efficiency. One promising method for reducing electromigration stress is through the use of ferromagnetic liners in a U-shape around the BLs and write WL. These liners serve to focus the magnetic field onto the MTJs in the desired row or column, and can increase the effective field by as much as a factor of 2 for a given current [25]. The use of ferromagnetic liners around the BL is illustrated in

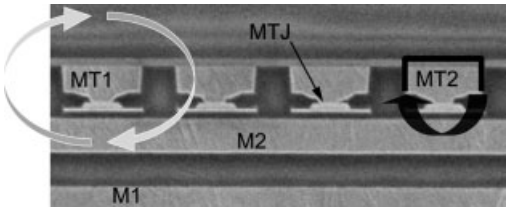


Figure 14.16 A cross-sectional image of a product array, from a viewpoint perpendicular to that of Figure 14.14. The arrows around bit line wire MT1 suggest the magnetic field configuration generated by a current through wire MT1; it is loosely contained, with only moderate magnitude at the MTJ free layers. Conversely, the wire MT2 exhibits an enhanced field magnitude due to its localization by the ferromagnetic film (lines) surrounding the copper MT2 wire.

Figure 14.16. Similar, but inverted, structures can be formed around the write WL (M2 in Figure 14.16) to enhance the field from that wire. The potential reduction in necessary current to obtain a required switching field can dramatically reduce electromigration issues. Not only do ferromagnetic liners offer potential reduction in current density, but they also improve electromigration performance relative to conventional copper processes. By reducing the interface diffusion of copper atoms, ferromagnetic cladding on the top surface of the MT wire enhances electromigration reliability to an extent similar to that seen in the industry by advanced Ta/TaN or CoWP capping processes [26].

One added benefit of the ferromagnetic liner field focusing is the reduction of any near-neighbor disturb effects. Because the field is better focused on devices along the desired WL and BL, adjacent devices are less likely to be switched by near-neighbor fields, or the combination of near-neighbor fields and thermal activation.

14.5.2
Tunnel Barrier Dielectrics

These are subject to reliability concerns because of the extremely thin nature of the barrier and related susceptibility to pinholes or dielectric breakdown. Aluminum oxide tunnel barriers have so far proven quite robust. Time-dependent dielectric breakdown (TDDB) and time-dependent resistance drift (TDRD) have been examined in 4 Mb arrays and found to exceed requirements for a 10-year lifetime [27]. The voltage stresses on the tunnel barrier are relatively modest, as the read operation takes place at 100–300 mV because the MR is higher for a lower voltage. The write operation is performed with one side of the MTJ floating, so there is no significant voltage stress on the MTJ during the higher-power write pulse.

14.5.3
BEOL Thermal Budget

The BEOL thermal budget for MRAM devices ($<$250–300 °C) is significantly lower than for conventional semiconductor fabrication processes ($\sim$400 °C), in order to

prevent degradation to the MTJs. This can affect the intrinsic quality of dielectrics being used in the BEOL, and can also worsen seam and void formation around the topographical features being encapsulated. A low thermal budget also prevents the use of certain post-processing passivation anneals, and packaging materials and processes. The move to lead-free solder with increased solder reflow temperatures represents a further challenge for MRAM.

14.5.4
Film Adhesion

This is a serious concern with the multiple new materials being introduced into the integrated process. The novel etch and passivation techniques being used also may leave behind poorly adherent layers which cannot be subjected to harsh wet cleans without MTJ exposure and degradation. Delamination risks must be mitigated through specially developed dry and wet cleans, the use of materials with tuned stress, and the choice of materials with compatible thermal expansion.

14.6
The Future of MRAM

As of July, 2006, MRAM products such as the 4 Mb memory shown in Figure 14.17 have been available from Freescale Semiconductor [28]. The market space targeted by Freescale includes networking, security, data storage, gaming, and printer data logging and configuration storage. From a customer viewpoint, this product means fewer part counts, a higher level of performance, higher reliability, greater environmental friendliness, and a lower cost solution than their current approaches, such as battery-backed SRAM.

Progressing downwards from the available 180 nm technology, future generations of MRAM are expected to utilize the same magnetic infrastructure with only evolutionary improvements, to below the 90 nm node. However, constraining the scaling are the following concerns:

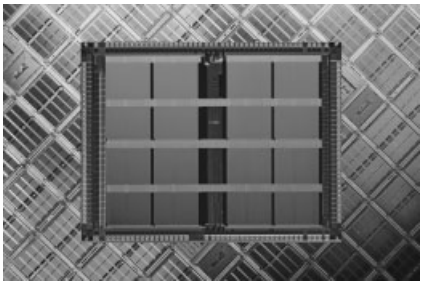


Figure 14.17 Photograph of a MR2A16A 4 Mb MRAM chip atop a wafer filled with such chips, presently available from Freescale semiconductor. (Illustration courtesy of G. Grynkewich and Freescale Semiconductor.).

- *Near-neighbor interactions:* When packing devices closer together, magnetic fields emanating from a given device can affect the switching behavior of devices nearby, and this can also be dependent on the given device's free layer state. In addition, the write wires for switching a given device will perturb neighboring devices to a greater extent as the latter come closer. It remains unclear how well these effects can be suppressed with the use of flux-closed MTJ layers and ferromagnetic cladding of write wires. Additional techniques such as enhanced-permeability dielectric (EPD) encapsulating films may be required to overcome these problems [29].
- *Increased switching fields:* As devices are scaled to smaller volumes, the anisotropy field must be increased to compensate and maintain activation energy greater than 60 $k_B T$ [30]. Write fields will scale to be of similar magnitude to the anisotropy field, and will increase superlinearly with inverse device size. As with the MTJs, the write wires must scale to a smaller footprint, making it more difficult to accommodate the increasing switching fields. In addition, ferromagnetic cladding of the wires becomes less effective because of the bending energy of the flux inside the cladding as the wire corner radius sharpens. EPD device encapsulations will help in this regard.
- *Device-to-device variability:* Process-induced line-edge roughness will become a more substantial fraction of the total device width, so that edge irregularities may become more effective at pinning the domains so they do not switch smoothly. The total device area and aspect ratio will also exhibit larger spreads, both from line-edge roughness and from variability in lithography. A reduced aspect ratio for tighter packing density will also decrease AQF, as the anisotropy field is more sensitive to shape for devices with a smaller aspect ratio [30].

Each of these concerns is not a fundamental limitation, but rather a practical limitation that can most likely be overcome with sufficient – albeit perhaps prohibitively expensive – investment in materials development and processing techniques. Hard physical limits do not appear to set in until superparamagnetism becomes important – that is, for device sizes below 20 nm [30], a dimension substantially below the limits suggested by the aforementioned practical issues.

Even with the practical limits to scaling conventional MRAM, one can expect to see revolutionary modifications to standard MRAM cell such that MRAM will be available with far greater densities, lower cost, and faster operation. Beyond the scope of this chapter are the impressive developments and exciting new proposals in the areas of:

- thermally assisted MRAM for reduced power requirements [31]
- spin-momentum transfer (SMT) MRAM for scaling to advanced process nodes and extremely small active memory devices [32]
- domain-wall memory for very high density serial storage [33]
- embedded MRAM as a replacement for embedded flash and low-density on-chip SRAM, for high-performance microprocessor cache memory and other ASIC applications [34].

In summary, this chapter has provided an overview of the rapid developments in MRAM technology over the past decade. Many major hurdles for MRAM product development have been surmounted in the face of funding limits set by competition with the huge silicon industry. Now that MRAM devices have grasped a toe-hold in the marketplace, new applications will be identified and MRAM development will proceed at an even faster pace over the next decade. Perhaps soon we will again see magnetic RAM in spacecraft!

Acknowledgments

The author would like to thank: W.J. Gallagher for the figures and editorial assistance; IBMs Materials Research Laboratory (MRL) for process development and fabrication; P. Rice, T. Topuria, E. Delenia, and B. Herbst for the TEM imaging; J. DeBrosse, T. Maffitt, C. Jessen, R. Robertazzi, E. O'Sullivan, D.W. Abraham, E. Joseph, J. Nowak, Y. Lu, S. Kanakasabapathy, P. Trouilloud, D. Worledge, S. Assefa, G. Wright, B. Hughes, S.S.P. Parkin, C. Tyberg, S.L. Brown, J.J. Connolly, R. Allen, E. Galligan for various contributions; and M. Lofaro for the advances in CMP. It is also acknowledged that the studies summarized here were supported in part by the Defense Microelectronics Activity (DMEA) and built on prior investigations conducted with Infineon (now Qimonda) within the MRAM Development Alliance, and also on earlier MRAM studies at IBM that were supported in part by DARPA.

References

1 DeBrosse, J. personal communication.
2 Gallagher, W.J. and Parkin, S.S.P. (2006) Development of the magnetic tunnel junction MRAM at IBM: From first junctions to a 16-Mb MRAM demonstrator chip. *IBM Journal of Research and Development*, **50**, 5–23. Sincere thanks also to John DeBrosse, John Barth, Chung Lam, and Ron Piro.
3 Jones, J. (1976) Coincident current ferrite core memories. *Byte*, **1**, 6–22.
4 Hanaway, J.F. and Moorehead, R.W. (1989) Space Shuttle Avionics System, NASA SP-504 available at, http://klabs.org/DEI/Processor/shuttle/sp-504/sp-504.htm.
5 (a) Binasch, G. *et al.* (1989) *Physical Review B-Condensed Matter*, **39**, 4828–4830;(b) Baibich, M.N. *et al.* (1988) Giant magnetoresistance of (001)Fe/(001)Cr magnetic superlattices. *Physical Review Letters*, **61**, 2472;(c) Heiliger, C., Zahn, P. and Mertig, I. (2006) Microscopic origin of magnetoresistance. *Materials Today*, **9**, 46–54.
6 (a) Kaakani, H. (March 10–17 2001) Radiation Hardened Memory Development at Honeywell. IEEE Aerospace Conference, Big Sky, MT, vol. 5, pp. 2273–2279;(b) Katti, R.R. (2002) *Journal of Applied Physics*, **91**, 7245.
7 Slonczewski, J. IBM internal report.
8 (a) (1975) Juliere (CNR-France) – first MTJ demonstration Fe/Ge/Co, DR/R∼14% at 4.2K;(b) Julliere, M. (1975) Tunneling between ferromagnetic films. *Physics Letters A*, **54**, 225–226.
9 Miyazaki, T. and Tezuka, N. (1995) Giant magnetic tunneling effect in Fe/Al_2O_3/Fe junction. *Journal of Magnetism and Magnetic Materials*, **139**, L231–L234.

10 (a) Berkowitz, A.E. and Takano, K. (1999) Exchange anisotropy – a review. *Journal of Magnetism and Magnetic Materials*, **200**, 552–570;(b) Nogues, J. and Schuller, I.K. (1999) Exchange bias. *Journal of Magnetism and Magnetic Materials*, **192**, 203–232.

11 (a) Stiles, M.D. (2006) Exchange coupling in magnetic multilayers, in *Nanomagnetism: Ultrathin Films, Multilayers and Nanostructures. Contemporary Concepts of Condensed Matter Science*, Volume **1** (eds D. Mills and J.A.C. Bland), Elsevier, New York, pp. 51–77. (b) Parkin, S.S.P. and Samant, M.G. (2003) Magnetic random access memory with thermally stable magnetic tunnel junction cells, U.S. Patent 6,518,588;(c) Parkin, S.S.P., More, N. and Roche, K.P. (1990) Oscillations in exchange coupling and magnetoresistance in metallic superlattice structures: Co/Ru, Co/Cr, and Fe/Cr. *Physical Review Letters*, **64**, 2304–2306; (d) Slaughter, J.M., Dave, R.W., DeHerrera, M., Durlam, M., Engel, B.N., Janesky, J., Rizzo, N.D. and Tehrani, S. (2002) Fundamentals of MRAM Technology. *Journal of Superconductivity: Incorporating Novel Magnetism*, **15**, 19–25; (e) Parkin, S.S.P. *et al.* (1999) Exchange-biased magnetic tunnel junctions and application to nonvolatile magnetic random access memory. *Journal of Applied Physics*, **85**, 5828–5833.

12 Schrag, B.D. *et al.* (2000) Néel 'orange-peel' coupling in magnetic tunneling junction devices. *Applied Physics Letters*, **77**, 2373–2375.

13 (a) C-7100EX Sputter Deposition Tool from Canon Anelva Corporation, Tokyo, Japan (see http://www.canon-anelva.co.jp/english); (b) Timaris Sputter Deposition Tool from Singulus Technologies, Kahl, Germany (see http://www.singulus.de).

14 (a)Parkin, S.S.P. *et al.* (2004) Giant tunnelling magnetoresistance at room temperature with MgO (100) tunnel barriers. *Nature Materials*, **3**, 862–867;(b) Hayakawa, J., Ikeda, S., Lee, Y.M., Matsukura, F. and Ohno, H. (2006) Effect of high annealing temperature on giant tunnel magnetoresistance ratio of CoFeB/MgO/CoFeB tunnel junctions. *Applied Physics Letters*, **89**, 232510–232512.

15 Stoner, E.C. and Wohlfarth, E.P. (1948) A mechanism of magnetic hysteresis in heterogeneous alloys. *Philosophical Transactions of the Royal Society*, **A240**, 599–642.

16 (a) Savtchenko, L., Engel, B.N., Rizzo, N.D., DeHerrera, M.F. and Janesky, J.A. (2003) Method of writing to scalable magnetoresistance random access memory element, U.S. Patent 6,545,906; (b)Durlam, M. *et al.* (2003) A 0.18 μm 4 Mb toggling MRAM, *IEDM Technical Digest*, 995;(c) Worledge, D. (2004) Spin flop switching for magnetic random access memory. *Applied Physics Letters*, **84**, 4559–4561;(d) Worledge, D.C. (2006) Single-domain model for toggle MRAM. *IBM Journal of Research and Development*, **50**, 69–79.

17 Reohr, W., Hoenigschmid, H., Robertazzi, R., Gogl, D., Pesavenot, F., Lammers, S., Lewis, K., Arndt, C., Lu, Y., Viehmann, H., Scheuerlein, R., Wang, L.-K., Trouilloud, P., Parkin, S., Gallagher, W. and Mueller, G. (2002) Memories of Tomorrow. *IEEE Circuits & Devices*, **18**, 17–27.

18 (a) Maffitt, T.M., DeBrosse, J.K., Gabric, J.A., Gow, E.T., Lamorey, M.C., Parenteau, J.S., Willmott, D.R., Wood, M.A. and Gallagher, W.J. (2006) Design considerations for MRAM. *IBM Journal of Research and Development*, **50**, 25–39;(b) Gogl, D. *et al.* (2005) A 16 Mb MRAM featuring bootstrapped write drivers. *IEEE Journal of Solid State Circuits*, **40**, 902–908.

19 Zhu, J.G. and Park, C. (2006) Magnetic tunnel junctions. *Materials Today*, **9**, 36–45.

20 Pearton, S.J. *et al.* (2000) Dry etching of MRAM structures. *Materials Research Society Symposium Proceedings*, **614**, F10.2.1–F10.2.11.

21 Samant, M.G., Luning, J., Stohr, J. and Parkin, S.S.P. (2000) Thermal stability of

IrMn and MnFe exchange-biased magnetic tunnel junctions. *Applied Physics Letters*, **76**, 3097–3099.

22 Crowell, J., Tedder, L., Cho, H., Cascarano, F. and Logan, M. (1990) Model studies of dielectric thin film growth: chemical vapor deposition of SiO_2. *Journal of Vacuum Science & Technology*, **A8**, 1864–1870.

23 Gaidis, M.C., O'Sullivan, E.J., Nowak, J.J., Lu, Y., Kanakasabapathy, S., Trouilloud, P.L., Worledge, D.C., Assefa, S., Milkove, K.R., Wright, G.P. and Gallagher, W.J. (2006) Two-level BEOL processing for rapid iteration in MRAM development. *IBM Journal of Research and Development*, **50**, 41–54.

24 Hughes, B. (2004) Magnetoresistive random access memory (MRAM) and reliability. Proc. IEEE 42nd Annual International Reliability Physics Symposium, Phoenix, AZ, pp. 194–199.

25 Durlam, M. *et al.* (2003) A 1-Mbit MRAM based on 1T1MTJ bit cell integrated with copper interconnects. *IEEE Journal of Solid-State Circuits*, **38**, 769–773.

26 (a) Gajewski, D.A., Meixner, T., Feil, B., Lien, M. and Walls, J. (2004) Electromigration of MRAM-customized Cu interconnects with cladding barriers and top cap. IEEE Integrated Reliability Workshop Final Report, pp. 90–92; (b) Walls, J. *et al.* (2004) Improved electromigration resistance of copper interconnects using multiple cladding layers, unpublished; see www.IP.com Document ID IPCOM000009315D.

27 (a)Åkerman, J. *et al.* (2004) Demonstrated reliability of 4-Mb MRAM. *IEEE Transactions on Device Materials Reliability*, **4**, 428–435;(b)Åkerman, J. *et al.* (2005) Reliability of 4-Mbit toggle MRAM. *Materials Research Society Symposium Proceedings*, **830**, 191–200.

28 Engel, B.N., Åkerman, J., Butcher, B., Dave, R.W., DeHerrera, M., Durlam, M., Grynkewich, G., Janesky, J., Pietambaram, S.V., Rizzo, N.D., Slaughter, J.M., Smith, K., Sun, J.J. and Tehrani, S. (2005) A 4-Mbit toggle MRAM based on a novel bit and switching method. *IEEE Transactions on Magnetics*, **41**, 132; http://www.freescale.com/mram.

29 Pietambaram, S.V., Rizzo, N.D., Dave, R.W., Goggin, J., Smith, K., Slaughter, J.M. and Tehrani, S. (2007) Low-power switching in magnetoresistive random access memory bits using enhanced permeability dielectric films. *Applied Physics Letters*, **90**, 143510.

30 Cowburn, R.P. (2003) Superparamagnetism and the future of magnetic random access memory. *Journal of Applied Physics*, **93**, 9310.

31 (a) Abraham, D.W. and Trouilloud, P.L. (May 7, 2002) Thermally assisted magnetic random access memory, U.S. Patent 6,385,082;(b) Prejbeanu, I.L., Kula, W., Oundadjela, K., Sousa, R.C., Redon, O., Dieny, B. and Nozieres, J.-P. (2004) Thermally assisted switching in exchange-biased storage layer magnetic tunnel junctions. *IEEE Transactions on Magnetics*, **40**, 2625; (c) Redon, O., Kerekes, M., Sousa, R., Prejbeanu, L., Sibuet, H., Ponthennier, F., Persico, A. and Nozières, J.P. (May 21–24 2005) Thermo-assisted MRAM for low power applications. Proceedings 1st International Conference on Memory Technology and Design (ICMTD-2005), Giens, France, pp. 113–114.

32 (a) Stiles, M.D. and Miltat, J. (2006) Spin transfer torque and dynamics, in *Spin Dynamics in Confined Magnetic Structures III: Topics in Applied Physics 101* (eds B. Hillebrands and A. Thiaville), Springer, Berlin, pp. 225–308;(b) Huai, Y., Albert, F., Nguyen, P., Pakala, M. and Valet, T. (2004) Observation of spin-transfer switching in deep submicron-sized and low-resistance magnetic tunnel junctions. *Applied Physics Letters*, **84**, 3118; (c)Fuchs, G.D. *et al.* (2004) Spin transfer effects in nanoscale magnetic tunnel junctions. *Applied Physics Letters*, **85**, 1205.

33 (a) Parkin, S.S.P. (2004) Shiftable magnetic shift register and method of using the same, U.S. Patent 6,834,005; (b) Parkin, S.S.P. (2005) System and method for writing to a magnetic shift register, U.S. Patent 6,898,132.

34 Iyer, S.S., Barth, J.E. Jr., Parries, P.C., Norum, J.P., Rice, J.P., Logan, L.R. and Hoyniak, D. (2005) Embedded DRAM: Technology platform for the Blue Gene/L chip. *IBM Journal of Research and Development*, **49**, 333–350.

15
Phase-Change Memories

Andrea L. Lacaita and Dirk J. Wouters

15.1
Introduction

15.1.1
The Non-Volatile Memory Market, Flash Memory Scaling, and the Need for New Memories

During the past decade, the impressive growth of the market for portable systems has been sustained by the availability of successful semiconductor non-volatile memory (NVM) technologies, the key driver being the Flash memories. In the past 15 years, the scaling trend of these charge-based memories has been straightforward. The cell density of NOR Flash, which is adopted for code storage, has doubled every one to two years, following Moore's law; the memory cell size is 10–12 F^2, where F is the technology feature size. The NAND Flash, which is optimized for sequential data storage, has been aggressively scaled and, nowadays, has a cell size of about 4.5 F^2. However, further scaling of both NOR and NAND Flash is projected to slow down, due mainly to the tunnel oxide (NOR), which cannot be further thinned down without impairing data retention, and to electrostatic interactions between adjacent cells (NAND).

Moreover, as the scaling proceeds, the number of electrons stored on the floating gate and present in the device channel decreases. As few electrons are involved, effects such as the random telegraph noise arising from trapping processes are expected to cause threshold instabilities and reading errors [1], while the requirements on retention become even more challenging. At the 32 nm node, the maximum acceptable leakage over a 10-year period will be less than 10 electrons per cell [2]. All of these difficulties, arising from the fundamental limitation of the charge storage concept, are calling for novel approaches to non-volatile storage at the nanoscale.

In recent years a number of different alternative memory concepts have been explored. Most notably, memories based on switchable resistors are considered

Nanotechnology. Volume 3: Information Technology I. Edited by Rainer Waser

ISBN: 978-3-527-31738-7

promising; among these, the phase-change memory (PCM) technology is attracting growing interest.

15.1.2 PCM Memories

PCM-based memory devices were first proposed by J.F. Dewald and S.R. Ovshinsky who, during the 1960s, reported the observation of a reversible memory switching in chalcogenide materials [3, 4]. Chalcogenides are semiconducting glasses made from the elements of Group VI of the Periodic Table, such as sulfur, selenium and tellurium, and many of these demonstrate the desired material properties for possible use in PCM applications. Two different chalcogenide material systems may be discriminated, based on their switching properties [5]:

- *Threshold-switching* in so-called "stable" glasses that show negative differential resistance and a bistable behavior, requiring a minimum "holding voltage" to sustain the high-conductive state. The typical materials are three-dimensionally cross-linked chalcogenide alloy glasses.
- *Memory-switching* in "structure reversible films" that may form crystalline conductive paths. A typical composition is $Te_{81}Ge_{15}X_4$ close to the Ge-Te binary eutectic, with X being an element from Group V or VI (e.g., Sb). The latter materials also show threshold switching to initiate the high conduction in the glass state, followed by an amorphous to crystalline phase transition which stabilizes the high-conductive state.

A non-volatile and reprogrammable phase-change (256 bit) memory array based on chalcogenide materials originally was reported by R.G. Neale, D.L. Nelson and Gordon E. Moore as far back as 1970 [6]. In these memories the memory element is basically a resistor made from a chalcogenide material and, depending on whether the chalcogenide layer is amorphous or crystalline, the device resistance would be either high (RESET state) or low (SET state) [7]. Programming of the phase state is carried out by current-induced Joule heating: either the material is heated above the melting temperature, followed by fast quenching in the amorphous state; or the element is heated to a high temperature below the melting point, allowing crystallization of the amorphous material. However, the operation characteristics of these memories were still poor (e.g., 25 V, 250 mA, 5 μs for programming in the RESET state). Indeed, such a high programming power requirement led to the suggestion that these prototype memories should be called "Read-Mostly Memory".

Chalcogenide phase-change materials were instead successfully adopted in xerography, where the photoconductive properties of arsenic-selenide (As-Se) were exploited, and in optical recording, spurred on by the development of Ge-Te glasses capable of undergoing rapid crystalline-amorphous phase transformations [8, 9]. In particular, rewriteable optical media (e.g., CDs, DVDs) became a huge field of application. In the case of CDs, the selective crystallization/amorphization is induced

by an external laser beam and not by Joule heating, while the binary information is read out by exploiting the change in optical reflectivity between the amorphous and the crystalline state, rather than the difference in electrical resistivity.

The advancements in the materials used for optical disks, coupled with significant technology scaling and a better understanding of the fundamental electrical device operation, eventually triggered the development of solid-state memory technology, which led initially to the Ovonic Unified Memory (OUM™) concept based on the use of the $Ge_2Sb_2Te_5$ chalcogenide compound [10, 11]. Since early 2000, the different semiconductor industries have considered the exploitation of the same concept for large-sized, solid-state memories [12–14]. Phase-change memories are known by different names. For example, the former OUM name was superseded by the terms PCM and phase-change RAM (PRAM). Today, PCM are considered promising candidates eventually to become the mainstream non-volatile technology, this being due to their large cycling endurance [15, 16], fast program and access times, and extended scalability [17, 18].

15.2
Basic Operation of the Phase-Change Memory Cell

15.2.1
Memory Element and Basic Switching Characteristics

The vertical OUM PCM memory element in the so-called Lance-like structure is shown schematically in Figure 15.1. The active phase-change material ($Ge_2Sb_2Te_5$; GST) is sandwiched between a top metal contact and a resistive bottom electrode (also called the heater). The programming current flows vertically from the bottom

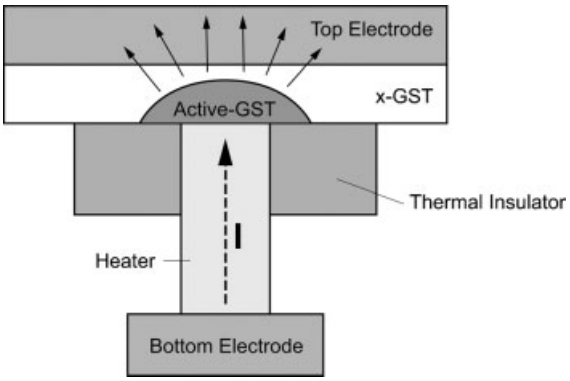


Figure 15.1 Schematic of the OUM™ vertical phase-change memory element. Due to the typical bias polarity, current flows vertically from the bottom electrode through the heater, through the $Ge_2Sb_2Te_5$ (GST) layer and to the top electrode. The current concentration near the (narrow) heater/GST contact results in local heating of the GST in a semispherical volume where the amorphous/crystalline phase change occurs. Amorphization of this region stops the low-resistive current path and results in an overall large resistance.

electrode through the heater, through the GST layer and to the top electrode. The current concentration near the (narrow) heater/GST contact results in a local heating of the GST in a semi-spherical volume where the amorphous/crystalline phase change occurs. Amorphization of this area stops the low-resistive current path and results in an overall large resistance.

The thermal and electrical switching characteristics of a vertical OUM PCM memory element are shown in Figure 15.2, with temperature evolution in the GST region above the heater contact in response to current pulses shown graphically in Figure 15.2a [12]. In order to form the amorphous phase, a 50- to 100-ns current pulse heats up the region until GST reaches its melting temperature (620 °C). The subsequent swift cooling, along the falling edge of the current pulse, freezes the undercooled molten material into a disordered, amorphous phase below the glass transition temperature. In order to recover the crystalline phase, Joule heating from another current pulse, with a lower amplitude (resulting in temperatures above the crystallization temperature but below the melting temperature), is used to speed-up the spontaneous amorphous-to-crystalline transition: the crystalline phase builds up in about 100 ns by a combination of nucleation and growth processes.

The typical current–voltage (*I*–*V*) curve of a cell for both states is shown in Figure 15.2b [19]. As the electrical resistivity of the two phases differs by orders of magnitude, at low bias, the resistance of the two memory states ranges from few kΩ (low resistance = *ON* or *SET state*) to some MΩ (high resistance = *OFF* or *RESET state*). Reading is accomplished by biasing the cell and sensing the current flowing through it; for example, a few hundreds of millivolts across the cell in the SET state generates 50–100 μA. This current is able to load the bit-line capacitances

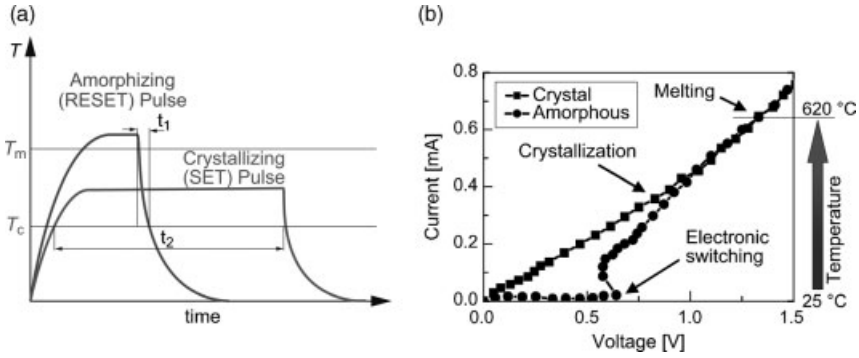


Figure 15.2 (a) Thermal-induced phase change of the material, either by melting and subsequent quenching in the amorphous phase, or by heating in the solid state inducing crystallization of the amorphous state. (Figure reproduced from Ref. [12]). (b) Current–voltage (*I*–*V*) curves for both the crystalline and amorphous states. (Figure reproduced from Ref. [19]). The high current levels required for the Joule heating can be obtained at low voltages, even for the amorphous state, on the basis of the electronic threshold switching phenomenon, which strongly increases the conductivity in the amorphous material above a certain threshold voltage.

of a memory array, making possible a reading operation in 50 ns. The same bias across the cell in the RESET state is not able to generate enough current to trigger the sensing amplifier, thus resulting in the evaluation of a "0".

It should be noted that the *I–V* curve in the high-resistance, amorphous state is quite peculiar. As the bias reaches a certain voltage (the threshold switching voltage) a "snap-back" takes place and the conductance abruptly "switches" to a high conductive state (see Figure 15.2b). The *I–V* curve of the crystalline GST does not feature threshold switching, and approaches the *I–V* of the amorphous state in the high current zone.

The occurrence of this "threshold switching" is a very important characteristic of PCM material. Indeed, without such a switching mechanism, which allows large currents to flow in the amorphous material at low voltages (~few volts), very high voltages (~100 V) would be required to switch the material to the "on" state, thus making electronic programming effectively non-practical.

The ratio of the threshold switching voltage and the thickness of the amorphous zone is usually referred to as the *critical threshold switching field*; for GST this quantity ranges between 30 and 40 V μm^{-1}. The critical threshold switching field can be taken as a guideline to compare different materials; for example, the lower the switching field the lower the switching voltage for the same thickness of the amorphous layer. However, as shown in Figure 15.3, even if the threshold voltage does scale with the memory resistance, which in turn depends on the amorphous layer thickness, the line does not cross the origin [20]. The concept of threshold switching field should, therefore, be handled with some care.

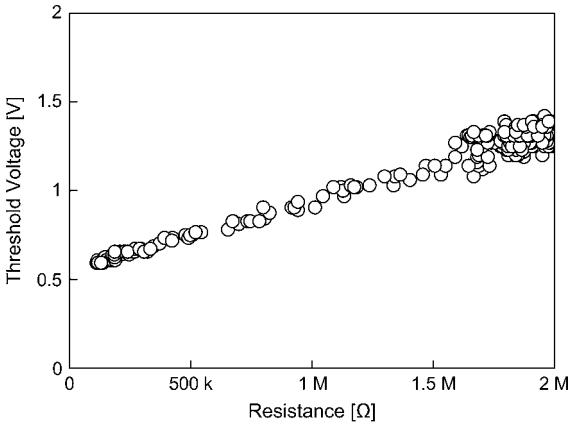


Figure 15.3 Experimental dependence of the threshold voltage on the low field resistance of the amorphous state. The threshold voltage scales with the device resistance, and therefore with the width of the amorphous zone. However, the line does not cross the origin, which highlights that a minimum voltage value of ~0.50 V, close to the holding voltage value, is required for switching to occur. (Figure adapted from Ref. [20]).

15.2.2
SET and RESET Programming Characteristics

The programming characteristic of a PCM cell [20] – that is, the dependence of the cell resistance R as a function of the programming current – is shown in Figure 15.4. The open symbols in Figure 15.4 refer to the resistance obtained when driving a cell from the RESET state. During the measurement procedure, a 100-ns programming pulse is applied and the cell resistance after programming is read at 0.2 V. Before the subsequent measurement, the cell is brought again into the initial reference RESET state by using a proper current pulse. The measurement cycle is then restarted, driving the cell with a new 100-ns programming current pulse with a different amplitude.

During this procedure, three distinct regions can be recognized:

- For programming pulses below 100 μA, the ON-state conduction is not activated and the very small current does not provide any phase change.
- In the 100 to 450 μA range, the resistance decreases following the crystallization of the amorphous GST, reaching the minimum resistance in the SET state, as denoted by R_{set}.
- Above 450 μA, the programming pulse melts some GST close to the interface with the bottom electrode, leaving it in the amorphous phase.

The solid symbols in Figure 15.4 also show the R–I characteristics obtained for the same cell, but starting from the SET state. The resistance value changes only when the current exceeds 450 μA and the chalcogenide begins to melt. The current is therefore denoted as the *melting current*, I_{melt}. From thereon the curve overlaps to the R–I of the RESET state. For programming pulses above 700 μA, the resistance of the cell reaches an almost constant value. It transpires that the PCM cell can be switched between the two SET and RESET states using current pulses of 400 and 700 μA, respectively, these pulses

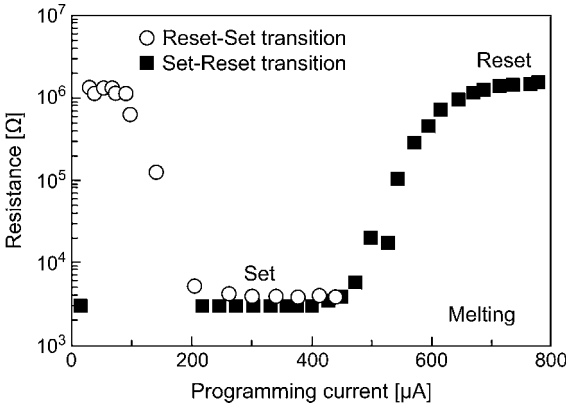


Figure 15.4 PCM programming characteristics, i.e., R as a function of the programming current I_P. Program pulses are applied to RESET cell (reset-set transition) or to a SET cell (set-reset transition). (Figure reproduced from Ref. [20]).

being independent of the initial cell state (resistance). Therefore, the cell can be rewritten with no need for any intermediate erase. The minimum current capable of bringing the cell into the full RESET state (700 μA in Figure 15.4) is denoted as reset current, I_{reset}.

The orders of magnitude difference between the cell resistance in the SET and RESET states makes the PCM memory ideally suitable for a multibit operation. In this scheme, the resistance of the cell may be set between the two extreme values, thus placing more than two levels per cell. This approach may become a viable option to further reduce the cost per bit of PCM devices.

15.3
Phase-Change Memory Materials

15.3.1
The Chalcogenide Phase-Change Materials: General Characteristics

The requirements for phase-change materials include easy glass formation during quenching from the melt, as well as congruent crystallizing compositions to avoid phase segregation during crystallization. Melting temperatures should be low to limit the switching power, whereas for non-volatility a good stability of the amorphous phase at application temperatures is required. It follows that the activation energy[1)] for crystallization of the amorphous state should be high enough to enable long data retention times. On the other hand, crystallization rates, at least at elevated temperatures, should be high enough to allow for a rapid amorphous to crystal transition, preferably in the range of a few tens of nanoseconds.[2)]

Such materials have now been under investigation for many years for their applications in DVD-RAM and DVD-R/W optical disk storage systems. Typically, metal alloys containing chalcogenide elements [by definition, elements of Group VI of the Periodic Table (O, S, Se, Te, Po)], and often referred to as "chalcogenide materials", are used. Chalcogenide elements are of interest as Se and Te compounds are easy glass-formers, because of their relatively high melt viscosities [22]. Compositions searched for are those that form a stable state in the solid phase ("polymorphic transformations"; i.e., where long-range diffusion is not required) [23].

The two typical chalcogenide material "families" used in PCM are both based on compositions of Ge, Sb and Te: (i) the pseudo-binary GeTe-Sb_2Te_3 compositions; and (ii) compositions based on the $Sb_{70}Te_{30}$ "eutectic" compound (see the Ge-Sb-Te ternary phase diagram in Figure 15.5) [24, 25].

1) The so-called "crystallization temperature" is not a uniquely defined material property, and varies depending on the time window of observation. "Activation energy" is therefore a better defined and more relevant physical parameter.

2) Recent discussions have indicated that the requirement of a rapid crystallization actually contradicts with easy glass formation, and rapid-crystallizing chalcogenides should be categorized rather as bad glass-formers based on their low glass transition to melt temperature (T_G/T_m) ratio compared to other easy glass formers such as SiO_2 [21].

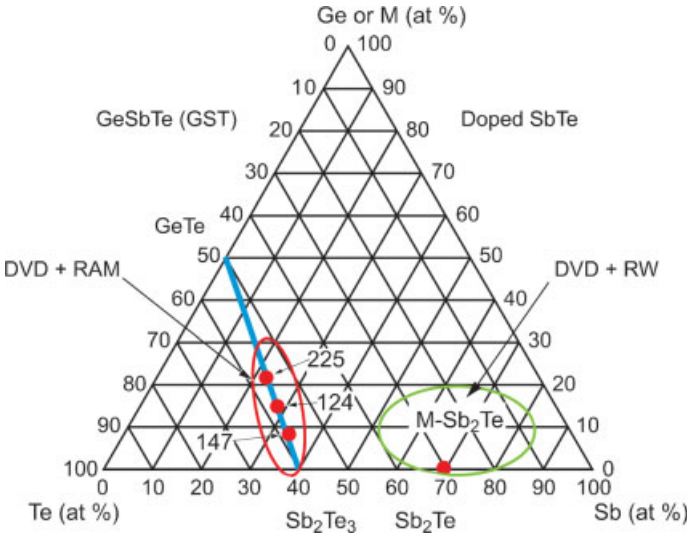


Figure 15.5 The Ge-Sb-Te (GST) ternary phase diagram, indicating the two classes of commonly used phase-change recording materials – that is, stoichiometric compositions along the GeTe-Sb_2Te_3 tie-line and compositions near the "eutectic" Sb_2Te. (Figure modified from Refs. [24, 25]).

15.3.1.1 The Pseudo-Binary GeTe-Sb_2Te_3 Compositions

The stoichiometric compositions around the GeTe and Sb_2Te_3 tie line are known as pseudo-binary compositions. These include the most widely used material $Ge_2Sb_2Te_5$, and they are used in the ovonic unified memory (OUM) [11], together with other compositions such as $Ge_1Sb_2Te_4$ [(1,2,4) material] and $Ge_1Sb_4Te_7$ [(1,4,7) material]. All of these materials are nucleation-controlled; that is, nucleation is dominant over growth [25], and are widely used in DVD-RAM applications. Along the tie line, the properties change from GeTe with high crystallization temperature (i.e., high stability) but slow crystallization speed, to Sb_2Te_3 that has a high crystallization speed but a low stability [26].

15.3.1.2 Compositions Based on the $Sb_{70}Te_{30}$ "Eutectic" Compound[3)]

These compositions are more generally indicated as doped SbTe (M-SbTe) compounds. Variants include doping with In: $In_x(Sb_{70}Te_{30})_{1-x}$, doping with Ag and In: $Ag_xIn_y(Sb_{70}Te_{30})_{1-x-y}$ (so-called "AIST"), and doping with Ge: $Ge_x(Sb_{70}Te_{30})_{1-x}$ + Sb. These materials are so-called fast-growth materials [28]: the growth starting from the crystal regions surrounding the amorphous zone is the dominant crystallization mechanism rather than nucleation of new crystals inside the amorphous. The benefits of these materials are possibly faster switching (<20 ns), a better

3 It should be noted that the $Sb_{70}Te_{30}$ "eutectic" material composition (sometimes quoted as Sb_2Te, or more exactly as $Sb_{69}Te_{31}$) is actually *not* an eutectic but an azeotropic minimum [27]; that is, it fulfills the basic requirement of a congruent crystallizing material.

high-temperature retention, and a lower threshold field for conductivity switching in the amorphous phase (10–20 V μm^{-1} instead of 30–40 V μm^{-1}). On the other hand, cycle endurance and resistance ratio would be smaller [24].

15.3.1.3 Other Material Compositions

Some other material compositions, based on selenium rather than tellurium compounds, have more recently been investigated for possible application in phase-change memories, including antimony selenide (Sb_xSe_{1-x}; main attributes, lower T_m and faster crystallization speed) [29], and indium selenide (In_2Se_3; main advantage wider resistivity range) [30].

15.3.1.4 N- or O-Doped GST

Both, nitrogen and oxygen doping of GST have been used mainly to control the resistivity of the material.

N-doping has been used successfully used to increase the crystalline GST-resistivity (from 2 to 200 mΩ·cm for N concentration from 0 to 7 atom%), and furthermore results in a smaller grain size and an increased crystallization temperature (+50 °C) [31].

O-doping of GST is reported to increase the resistance ratio (from 100 to 1000), and to improve the high-temperature retention with an increase of the activation energy from 3.6 eV to 4.4 eV [32].

15.3.2 Material Structure

15.3.2.1 Long-Range Order: Crystalline State in GST and Doped Sb-Te

GST is characterized by two different crystal structures in the crystalline state – that is, a lower temperature (higher resistivity) fcc cubic state, and a higher temperature (lower resistivity) hexagonal state (see Figure 15.6) [24]. It should be noted at this point that the temperatures at which the crystallization and phase transitions occur are not fixed but rather depend on the heating rate. In a fast phase-change operation the amorphous GST probably crystallizes in the metastable fcc phase. The existence of the two different states may however influence the long-term, low-temperature stability and eventual resistance in the SET state. However, it has been reported that, for nano-sized structures, the temperature of the fcc to hexagonal phase moves up (from ~360 °C to >450 °C for 65-nm patterns). The possible inhibition of the hexagonal state formation, together with the higher resistivity of the cubic state, may be a beneficial side effect of the scaling [34]. On the other hand, doped Sb-Te shows only one, low-resistive, hexagonal state.

15.3.2.2 Short-Range Order in Crystalline versus Amorphous State

Recent investigations have clarified the situation regarding the crystal-amorphous phase transition in these chalcogenide materials.

It has been noted [9, 35, 36] that, in contrast to the covalent semiconductors, in which amorphization does not changes the local ordering, in chalcogenide materials

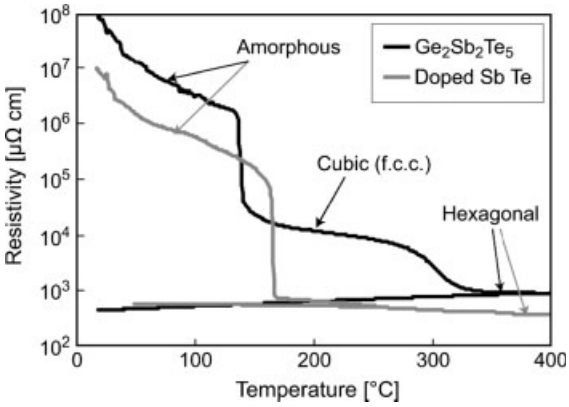


Figure 15.6 Resistive traces measured during temperature ramp of as-deposited amorphous GST and doped SbTe films, evidencing the occurrence of two different crystal states for GST but only one state for doped-SbTe. (Figure reproduced from Ref. [24]).

the amorphization induces a substantial increase in the local ordering and an important change in the resultant physical properties, such as an increased energy gap. This order–disordered transition is mainly due to a flip of the Ge atoms from an octahedral position into a tetrahedral position without rupture of strong covalent bonds (see Figure 15.7) [35]). Therefore, this class of materials is characterized by two competing structures with similar energy but different local order and different physical properties.

It is assumed that the nature of such a transformation ensures not only the large changes of physical properties (e.g., reflectivity and conductivity), but also the rapid performance and repeatable switching over millions of cycles.

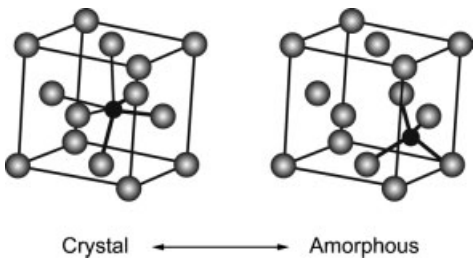


Figure 15.7 Fragments of the local structure of GST around Ge atoms in the crystallized (left) and amorphous (right) states. Upon heating the sample by a short intense pulse (above the melting point, T_m) and subsequent quenching, the Ge atoms flip from the octahedral to tetrahedral-symmetry position. Note that the stronger covalent bonds remain intact upon the umbrella-flip structural transition rendering the Ge lattice random. Exposure to light that heats the sample above the glass-transition temperature (T_g) – but below T_m – reverses the structure. (Figure reproduced from Ref. [35]).

15.3.3
Specific Properties Relevant to PCM

The majority of PCM materials investigated to date were developed for (re)writable optical discs, and existing knowledge of them, as well as experience of their reliability in products, should allow their rapid introduction into CMOS integration technology. However, their use in PC-RAM application may include various pitfalls:

- There exist certain *important operational differences* between DVD and PCM applications, that require the tuning/optimization of a number of specific material parameters for PCM that are not important for optical applications (e.g., resistivity in the on and off state, rather than reflectivity changes) (see Table 15.1). In addition, the amorphous phase should possess the particular property of threshold switching. The main material parameters for PCM applications are listed in Table 15.1.
- More importantly, however, the operating differences may have a major influence in the operation stability and repeatability. Indeed, in DVD applications, programming is achieved by laser pulse power coupling, and reading by reflectivity change. Data programming and storage relies on average material properties, such as reflectivity and absorption, that are not very sensitive to local variations. For example, they may be caused by small crystalline particles embedded in the amorphous region or vice versa, due to incomplete/inhomogeneous nucleation or amorphization. On the other hand, PCM relies on programming by Joule heating; that is, by current conduction through the device. Furthermore, the SET operation requires threshold current switching in the amorphous phase, which is a filamentary process. Also, the reading is based on a resistance (i.e., current) measurement. As the current conduction is greatly affected by the existence of local inhomogeneities, programming and reading may become highly sensitive to non-uniform/incomplete crystallization or amorphization. For example, low-resistive current paths in the incomplete amorphized state, or amorphous

Table 15.1 The major important material parameters for PCM, optimization direction, and value for GST material.

Symbol	Parameter	Optimization	GST 225 value	Reference(s)
T_m	Melting temperature	Minimal (low RESET power)	621 °C	[37]
T_c	Crystallization Temperature	Maximal (good retention)	155 °C	[37]
E_a	Activation energy	Maximal (good retention)	2.6–2.9 eV	[15, 38]
ρ_c	Resistivity crystalline state	"High" for low program current – "low" for fast read	$\sim$350 Ω μm	[18]
ρ_α	Resistivity amorphous state	High for good Resistance Ratio	$\sim$0.3 MΩ μm	[18]
E_c	Critical threshold switching field	Low for low SET program voltage	30–40 V–μm^{-1}	[24]

current-blocking regions in the incomplete crystallized state, may jeopardize proper programming or reading. In particular, the filamentary process of current switching during the SET state may induce strong inhomogeneous heating due to uncontrolled location of the filament that may results in only partial crystallization [20].

15.4
Physics and Modeling of PCM

From the basic device operation described in Section 15.2 it transpires that material-phase transitions (amorphization and crystallization dynamics) and conductance (threshold) switching mechanisms in the amorphous phase are the key processes involved in PCM. The physics of these mechanisms are outlined in greater detail below. The results of modeling studies implementing detailed microscopic descriptions of these effects have contributed greatly to an understanding of the subject, and have also supported the basis of design optimization of these devices.

15.4.1
Amorphization and Crystallization Processes

The different phase transitions of the PCM material during programming are illustrated graphically in Figure 15.8 [39]. Here, the crystalline material serves as an ideal starting point. During the programming transition from a SET to a RESET (high-resistance) state, the material is heated, begins to melt at solidus temperature,

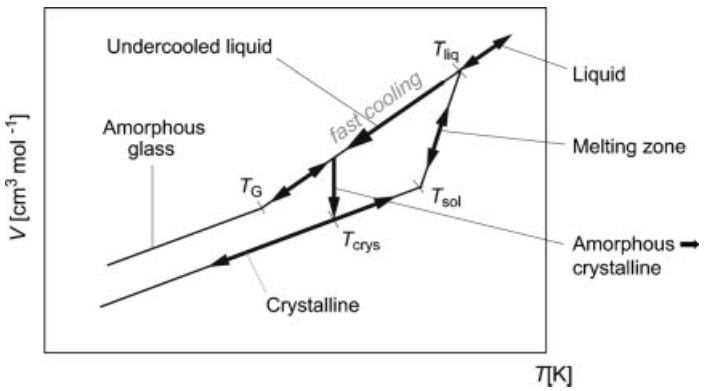


Figure 15.8 Schematic of molar volume (corresponding to 1/density) changes of phase-change material with temperature, showing the different possible phase transitions and critical temperatures. When heating a crystalline phase, melting will start at the solidus temperature T_{sol}, and will be complete at the liquidus temperature T_{liq}. Slow cooling will crystallize the material again following the same transition line; however, fast cooling results in an undercooling liquid that will be quenched in an amorphous glass below the glass transition temperature T_G. Heating the amorphous phase above T_G will result in crystallization at the crystallization temperature T_{crys}.

T_{sol}, and is completely molten at the melting temperature, T_m. When the cooling rate is higher than 10^9 K s^{-1} [40], the material does not begin to solidify at T_m, but remains an undercooled liquid. Below the glass transition temperature, T_G, the material freezes into the amorphous state.

The reverse situation is that, during the RESET to SET transition, the current pulse heats the material above T_G but below T_m, where it will begin to crystallize. There is no uniquely defined crystallization temperature, and even at relatively low temperatures (100–200 °C) crystallization may occur over long time scales (perhaps up to years). (These processes govern the basic long-term temperature retention of the RESET state, and will be discussed in Section 15.6.) For the SET programming, high-temperature rapid (<100 ns) crystallization processes are required, an understanding of which is based mainly on the general physical models for nucleation and growth. Different models have been proposed, for example by Peng *et al.* [41] and by Kelton [42], by which the temperature-dependent nucleation and growth rate can be calculated. Calculations based on these models using the different material properties of, for example GST and AIST, indeed confirm the nucleation, respectively growth-dominated crystallization mechanisms, that have been observed experimentally in these materials [43].

15.4.2
Band-Structure and Transport Model

The development of a comprehensive and quantitative framework to support the design and optimization of these devices is a challenging task. Today, PCM physics is extremely well developed, and a model should be capable of coupling a description of carrier transport in both crystalline and amorphous phases, together with the heat equation and phase-transition dynamics. The starting point here is to describe the electrical properties, thus aiming to reproduce correctly the electronic switching effect in the amorphous chalcogenide alloy.

Recently, it has been shown that the adoption of a semiconductor-like picture for both the amorphous and crystalline phases is quite effective, and may successfully account for the experimental *I*–*V* curves [33]. Moreover, this also allows the handling of simulations within the frame of codes already widely adopted by the semiconductor industries. Optical absorption data have shown that both crystalline and amorphous GST have gaps of 0.7 and 0.5 eV, respectively [33]. Moreover, there is no doubt that the carrier dynamics in the crystalline GST can be treated according to Bloch's theorem, as in a crystalline semiconductor. In contrast, investigations on amorphous compounds have demonstrated the existence of states with variable transport properties [44]. During the 1960s, it became common practice to describe these materials as shown in Figure 15.9 (left), where "mobility edges" separate fully conductive bands from the low mobility states. This picture can be easily translated in terms of a semiconductor-like framework. By assuming that low-mobility localized states behave like trapping centers, and that more conductive levels resemble delocalized states, a band structure can be defined and the amorphous GST modeled as a "very defective" crystalline semiconductor [33] (Figure 15.9, right). The material

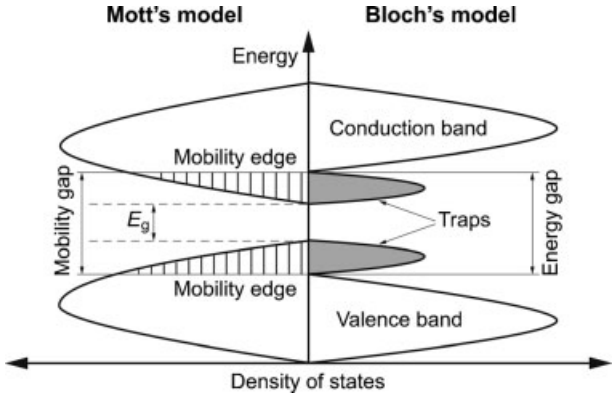


Figure 15.9 Comparison between the classical Mott and Davis's picture for the amorphous band diagram, and the scheme recently proposed by Pirovano *et al.* [33]. (Figure from Ref. [19]).

parameters, such as energy gap, trap densities and density of states for both phases, adopted in the numerical simulations, are listed in Table 15.2. It should be noted that the crystalline GST is p-type, due to a large density of vacancies (10% of the lattice sites), while the amorphous GST is characterized by a large density of donor/acceptor-like defects: the so-called Valence Alternation Pairs. This semiconductor-like picture is able to account for the peculiar conduction of the amorphous state and for the threshold switching [19].

The physics involved in the threshold switching remains a subject of debate. Since Ovshinsky first reported threshold switching [4], different models have been proposed, with many groups supporting the idea that switching is essentially a thermal effect and that the current in an amorphous layer rises above due to the creation of a hot filament [45, 46]. Later, Adler showed that the effect is not thermal (at least in thin chalcogenide films), in agreement with Ovshinsky's original picture. In their pioneering studies [47, 48], Adler and colleagues showed that a semiconductor resistor may feature switching, without any thermal effect. The condition for the threshold snap-back to occur is the presence of a carrier generation depending on

Table 15.2 Electronic parameters for both crystal and amorphous phases. (From Ref. [33]).

Property	GST crystalline	GST amorphous
E_{gap} [eV]	0.5	0.7
N_C [cm^{-3}]	2.5×10^{19}	2.5×10^{19}
N_V [cm^{-3}]	2.5×10^{19}	10^{20}
Vacancies [cm^{-3}]	5×10^{20}	—
C_3^+ [cm^{-3}]	—	10^{17}–10^{20}
C_1^- [cm^{-3}]	—	10^{17}–10^{20}
μ_n-μ_p [$cm^2 V^{-1} s^{-1}$]	0.1–23.5	5–200
F_C [$V cm^{-1}$]	3×10^5	3×10^5

field and carrier concentration (e.g., impact ionization) competing with a Shockley–Hall–Read (SHR) recombination via localized states.

The numerical model reported in Ref. [33] implements Adler's picture accounting for avalanche impact ionization in the amorphous and SHR recombination via the localized defects.

The schematic dependence of the band structure along a cross-section of a PCM device is shown in Figure 15.10, where the wide-gap region corresponds to the amorphous GST. At low bias, the quasi Fermi levels in the amorphous GST are close to their equilibrium position. As both the carrier density and their mobility is low (the average hole mobility is about $0.15\ cm^2\ V^{-1}\ s^{-1}$ [33]), the conduction regime is ohmic. By increasing the voltage, the applied field approaches the avalanche critical field of 3×10^5 [10], significantly increasing the carrier generation. The quasi Fermi levels thus split and move close to the band-edges (Figure 15.10, lower diagram). Carrier recombination mainly takes place in the region, close to the anode, where the electron Fermi level approaches the conduction band. At large bias, all defects available for recombination are full, and recombination may no longer be able to balance the exponentially rising generation rate. The system reacts by reducing the voltage drop in order to maintain the balance between recombination and generation, leading to the electronic switching. Hence, the snap-back takes place and, after switching, the GST is still amorphous but highly conductive. Generation is sustained by the large density of free carriers. According to this picture, the minimum voltage required for the switching to occur is of the order of the split between quasi Fermi levels (i.e., the

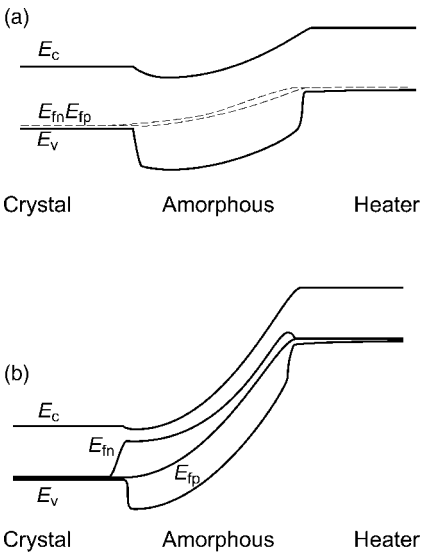


Figure 15.10 Band diagrams along the cross-section of a PCM cell in the RESET state (according to Ref. [33]). (a) At low bias, quasi-Fermi levels are close to the equilibrium value. (b) Close to threshold switching; generation by impact ionization is properly balanced by recombination though trap levels. When recombination saturates, generation finds a new stable working point reducing the voltage across the device. (Figure from Ref. [19]).

energy gap). This argument may justify why both, the holding voltage, V_H, and the asymptotic value at low R in Figure 15.2b, approaches approximately 0.5 V.

Although this picture so far has been successful in accounting for the experimental findings, it should be accepted with a degree of caution, and further investigations are needed to better assess the material properties. The quantitative description of impact ionization in these materials, as well as the role of the interfaces or of Poole–Frenkel mechanisms, deserve further investigation as many of the details still lack direct experimental verification. However, recent industrial interest in PCMs may lead to new experimental efforts and to the fabrication of devices purposely designed to test the validity of these key assumptions.

15.4.3
Modeling of the SET and RESET Switching Phenomena

The above conduction model was then coupled to heat equation and to phase-transition dynamics (nucleation and growth). When implemented in a three-dimensional (3-D) semiconductor device solver, it highlighted substantial differences in the two phase transitions [20]. The temperature maps during the SET–RESET transition, and the resulting phase distributions obtained with increasing current pulses with a plateau of 150 ns, are illustrated in Figure 15.11 [20]. In this figure, all of the pictures refer to a Lance PCM device in which a cylindrical metallic heater is in contact with the GST layer. The current flows almost uniformly across the polycrystalline GST, thus resulting in a roughly hemispherical shape of the final a-GST volume. As the programming current increases above melting, the volume left in the amorphous state increases.

On the other hand, Figure 15.12 [20] shows the calculated final phase distribution for a RESET to SET transition with programming current of 130 μA (a) and 160 μA (b) , respectively. Figure 15.12a shows that the current first sparks by electronic threshold

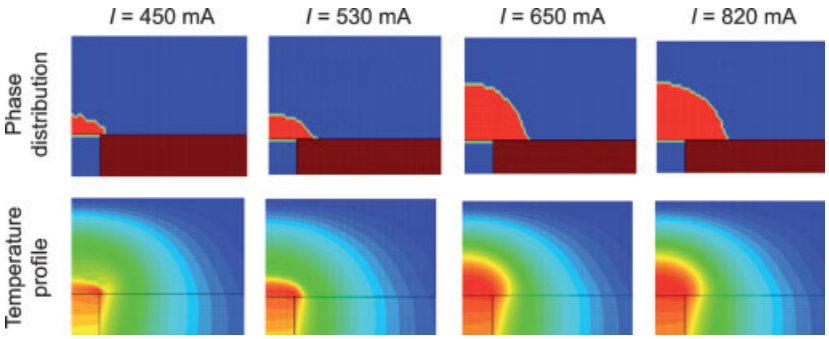


Figure 15.11 Homogeneous heating of the crystalline GTS region (blue) during SET to RESET transition (bottom row), resulting in homogeneous amorphous regions (red) at the end of a programming pulse (top row). Figures from left to right correspond to increasing the peak current value. A larger amorphous region will correspond to a higher resistance level. (Figure from Ref. [20]).

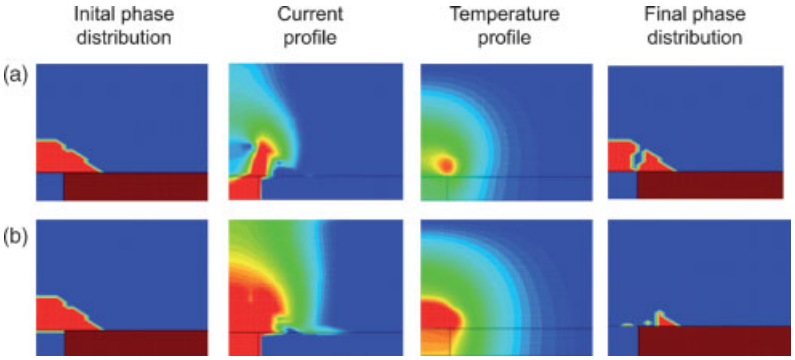


Figure 15.12 Programming operation from amorphous to crystal state (only half of the cell is shown because of cylinder-symmetry). (a) Top row: $I = 130\,\mu A$: electronic switching will occur first in the regions where the amorphous thickness is minimum. After the switching event, current spikes will increase the temperature only locally, and crystallization may only be induced in these hot zones resulting in a non-uniform final phase distribution with a major part of the amorphous zone remaining. (b) Bottom row: $I = 160\,\mu A$: only at higher voltages, eventually the hot filament extends in the whole the active area, leading to an homogenization of temperature and of the transformed volume. (Figure from Ref. [20]).

switching in the weakest a-GST region, locally triggering Joule heating and consequent crystallization processes. In case (a), the pulse amplitude/time is not sufficient to provide a complete crystalline path, and thus the active region features a residual a-GST layer causing the large measured R. In case (b), a further increase in the applied voltage eventually extends the hot filament in the whole of the active area. The initial amorphous volume has been almost completely crystallized by the programming pulse, resulting in a sufficiently low resistance. It should be noted that the localized phase transition is directly related to the details of the electronic switching mechanism, and thus can be reproduced only by a self-consistent model describing both electrothermal and phase-change dynamics.

Figures 15.11 and 15.12 also suggest that, by changing the programming pulse, the value of the cell resistance can be reliably placed in between the largest and the minimum SET value, thus opening the way to a multi-level operation. For example, four levels – each with different resistance values – might be programmed per cell, thus reducing the cost per bit.

15.4.4
Transient Behavior

V_{TH} and R are two key parameters of the memory cell. Figure 15.13 illustrates the time dependence of these parameters as measured soon after the current pulse programming the cell in the *RESET* state [19, 49]. The first fast component of the transient is referred to as *recovery*. On the longer time scale, in the so-called drift regime, the V_{TH} and R transients follow a slower power law. The recovery sets the minimum time needed after programming before reading (if the cell is read soon after being programmed in the *RESET* state, the read value might erroneously be "1").

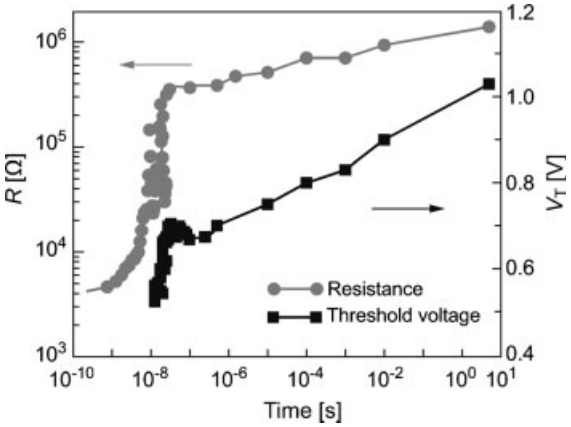


Figure 15.13 Low field resistance and threshold voltage for the amorphous phase as a function of time after reset programming operation. In about 50 ns, both low field resistance and threshold voltage are recovered, after which they continue to increase due to the drift phenomenon. (Figure from Ref. [19]).

Drift is instead a limit for multi-level operation, as the resistance of an intermediate level during ten years (3×10^8 s) might cause the bit to be erroneously decoded.

Different models have been proposed to justify these effects. Recovery is likely due to charge transients. After quenching, the newly formed amorphous region is full of trapped carriers, and some nanoseconds are required for these carriers to be released by trapping states and for the Fermi levels to recover the equilibrium value (see Figure 15.10a). During this transient stage, V_{TH} and R change from the *set* to the corresponding *RESET* values. Drift physics is more controversial, however, it having been suggested that drift might be due to mechanical stress release following the crystalline-to-amorphous phase transition. The resulting band-gap widening may reduce the mobile carrier density, thus contributing to the charge conduction. Another possible explanation links the effect to changes of the electronic states [50]. Variation of the density of states close to the band-edge already observed in other chalcogenide compounds [51] as the amorphous evolves towards to a more regular microscopic structure. In order for multi-level storage to be implemented in PCM memories, this effect should be fully understood and minimized.

15.5 PCM Integration and Cell Structures

15.5.1 PCM Cell Components

PCM elements may be organized in a memory matrix with or without adding a selection device to each phase-change element. Indeed, a raw cross-point matrix may

be conceived [52], but this would suffer read errors due to leakage through neighboring "on" elements. Moreover, as important program disturbs may occur in half-select regimes, each PCM memory cell should contain both the phase-change element and a selection device. The choice and dimensions of the selection element is determined by cell size constraints and the RESET program current. A MOSFET transistor is the most evident choice for memory integration in CMOS technology (Figure 15.14) [53]). The need for Source and Drain contacts, such a "one transistor-one resistor" (1T1R) cell would require a minimum area of 8–10 F^2 (where F is the minimum feature size of the technology). However, MOSFETs have limited current drivability, so that minimum size transistors cannot be used, and cell sizes are much larger, up to $\sim$40 F^2 in 0.18 μm technology with 0.6 mA reset current [13].

An alternative is the use of a bipolar p-n-p select device. In that case, there are still two (Base and Emittor) contacts per cell, while the Collector is a collective substate contact. As the bipolar current drivability is much higher, the overall cell size is smaller (only $\sim$10 F^2 in 0.18 μm technology with 0.6 mA reset current [13]). The trade-off is of course a more complex integration scheme for fabricating these bipolar devices in a CMOS technology, restricting this solution for stand-alone memory applications only.

In order to obtain the smallest cell area, a diode selector may be used [17], as a diode would only need one contact and can handle large currents (a self-rectifying device would be the most ideal case, so that a raw cross-array could have the same functionality with minimum cell size). However, in the diode selection regime, both bitlines and wordlines have to carry relatively high currents, leading to a partitioning of the memory as well as to larger X-decoders. Moreover, isolation is also less perfect and parasitic resistances in the full signal path are important. Both, series resistances and leakage currents, contribute to read error, while diode integration may also result in the formation of a parasitic bipolar transistor.

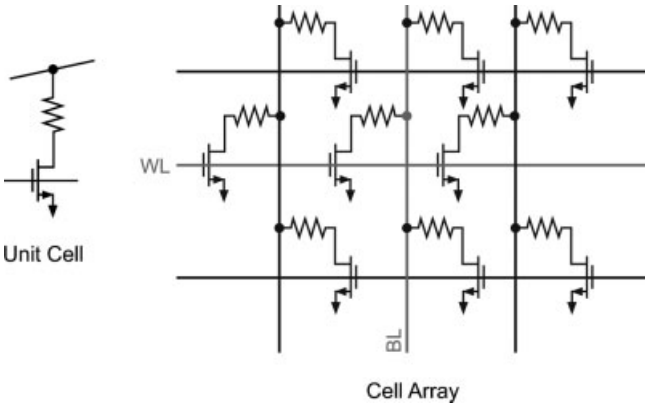


Figure 15.14 Schematic of 1T1R cell structure and memory array matrix. (Adapted from Ref. [53]).

15.5.2
Integration Aspects

Besides formation of the selection transistor, PCM fabrication requires the integration of a phase-change element in a CMOS technology. The memory element may be fabricated after the transistor processing (the so-called front-end-of-line processing; FEOL), and either before (e.g., in-between Si contacts and first metal interconnect layer) or after the first steps of the interconnect (e.g., on top of Metal 0 or Metal 1 interconnect levels). This latter scheme is the back-end-of-line or BEOL processing. A schematic state-of-the-art process flow is shown in Figure 15.15 [54, 55].

The PCM material is typically deposited by sputtering (physical vapor deposition; PVD) from a multi-element target with the desired composition. The as-deposited phase is either amorphous (for room- or low-temperature deposition), or crystalline if deposited above the crystallization temperature. In either case, due to the temperature budget of the following BEOL processing (up to 400–500 °C), the material will be fully crystallized after the integration. In a number of cell concepts, the conformal deposition of the PCM material and/or the ability to fill small pores is important [14], and these requirements would ideally call for a conformal deposition technique such as metalorganic chemical vapor deposition (MOCVD), rather than PVD. Critical elements for the integration are a good adhesion of the PCM material on the underlying substrate structure (typically a patterned metal electrode in a SiO_2 matrix), the material out-diffusion/inter-reaction/oxidation during high-temperature steps [56], and dry-etch patterning of the PCM [57, 58].

Furthermore, suitable electrode materials are needed for both the "heater" contact (where the metal will be in contact with the hot/molten PCM) and the (cold) top electrode. Material stability, low contact resistance, and good adhesion are important parameters. Poor electrode contact properties indeed have been identified as being responsible, for example, for "first fire effects" (see Section 15.6). Typical electrode materials are standard conductive barrier materials available in Si processing such as

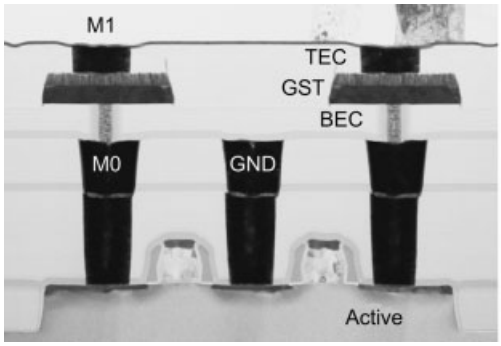


Figure 15.15 Cross-sectional scanning electron microscopy image showing the cell integration scheme (from Ref. [54]), and schematic process flow for PCM cell formation (from Ref. [55]). In this case, the PCM cell is integrated between the M0 and M1 interconnect levels.

TiN (e.g., Ref. [31]), although W is also often used [59]. The top electrode typically defines the PCM area and can be used as (part of) a hard mask during the PCM patterning.

Stability and controllability of the different process steps of the integration technology are crucial for the preparation of large-density memory arrays. The most important array characterization technique is therefore the distribution of the ON and OFF program state resistances. Tight distributions are needed to maximize the sensing window and to avoid bit errors. By process optimization, excellent distributions have been obtained on 256 Mb PCM after full integration [60].

15.5.3
PCM Cell Optimization

In the basic OUM PCM cell, the PCM and top electrode are planar layers deposited on a plug-type, bottom heater contact. That part of PCM material effectively involved in the phase switching is basically a hemispherical volume on top of the heater. To reduce the heating power (or program current), it is important to try to confine the dissipated heat as much as possible. While many different cell structures have been proposed in literature (see Figure 15.16), the optimization of heat confinement is in fact based on two simple principles: (i) by concentrating the volume where effective Joule heating takes place; and/or (ii) by improving the thermal resistance to reduce the heat loss to the surroundings.

15.5.3.1 **Concentrating the Volume of Joule Heating**

The Joule heating volume can be confined by pushing the current through a small cross-section with high current density. One obvious way to do this is to reduce the contact area of the heater contact with the PCM material, for example by minimizing the heater plug diameter (as in the small "sub-litho" contact heater cell [31]), by using only a conductive liner as heater (as in the edge-contact cell [61]), or by filling the plug with isolating dielectric material (as in the μ-trench cell [13, 62] and ring-contact heater cell [63]). The main advantage of using only the conductive liner is that at least one dimension of the heater area is controlled by the liner thickness, and not by the lithography.

Another way of confining the Joule heating volume is by structuring the PCM material to a narrow cross-section (see bottle-neck cell [64], line heater cell [24], self-heating pillar cell [65]). Finally, further improvement can be made by increasing the resistivity of the PCM material, for example by N or O doping [31, 32, 53]. In a different approach, a highly resistive TiON layer is made between the TiN electrode and the PCM material, in which layer the Joule heating will be concentrated resulting in lower program power [66]. However, such an approach may negatively influence the contact resistance to the PCM.

15.5.3.2 **Improving the Thermal Resistance**

In the so-called confined cell structure [14], the PCM is deposed in a pore etched back in the heater. This not only concentrates the Joule heating region but also surrounds a

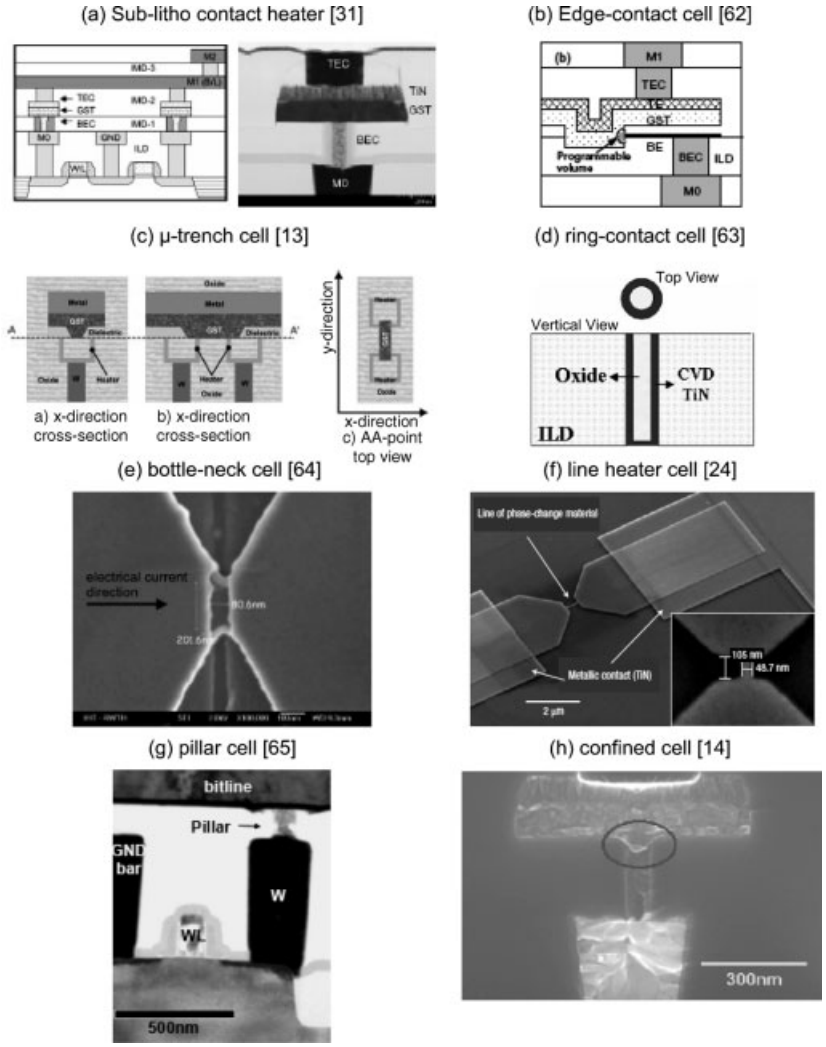


Figure 15.16 Different PCM cell structures for reducing the program power.

large part of this volume by a dielectric layer with reduced thermal conductivity. The drawback here is the topography which, ideally, would require a conformal deposition of the PCM. The alternative is to structure the PCM material, rather than the heater, leading to a plug or pillar. This approach is based on the same principle [18].

Another way to improve the thermal resistance is by increasing the PCM thickness (as this limits the heat flow to the top electrode heat sink) [14]. However, this option should be traded-off with the threshold voltage required for electronic switching during SET programming. One of the benefits of the horizontal line cell [24] is also the setting apart of the heated zone from a metal heater contact and a capping with thermal insulating dielectric.

Apart form changing only the cell structure, it is clear also that the correct material selection (use of different dielectric materials, and especially the use of porous dielectric materials) can improve the thermal heat confinement [53]. It should be noted, however, that the improved thermal isolation should not avoid the rapid quenching from the melt during RESET, otherwise the device cannot be programmed to the OFF-state.

15.6 Reliability

15.6.1 Introduction

As for any other non-volatile memory technology, reliability is one of the major concerns. The main specific reliability issues of PCM are: (i) data retention of the RESET, affected by the (limited) stability of the amorphous state; (ii) endurance, limited by the occurrence of stuck at RESET (open) or stuck at SET (short)-type defects; and (iii) program and read disturbs – that is, the stability of the amorphous phase due to repeated, though limited, thermal cycling caused by reading or programming neighboring cells.

While this section is based on reliability tests on the cell level (probing "intrinsic" reliability), for large memories the effect of reliability tests (e.g., of a temperature bake or of a large number of SET/RESET programming cycles) on the resistance distributions should also be evaluated to screen eventual "extrinsic" failures. However, until now only a few (preliminary) results on array reliability statistics have been reported [15, 67].

15.6.2 Retention for PCM: Thermal Stability

The most important requirement for a non-volatile memory is the ability to retain the stored information for a long time, the typical specification being 10 years (at a minimum of 85 °C). As the SET state is stable from a thermodynamic point of view, it has no problem of data retention. On the other hand, the RESET state, corresponding to the amorphous phase, is instead meta-stable and may crystallize following a dynamic which is heavily dependent on temperature. The retention of an amorphous state is, therefore, critical.

The retention performance of PCM technology is addressed by performing accelerated measurements at high temperatures. Figure 15.17 shows the typical failure time under isothermal conditions and with no applied bias, measured at several temperatures ranging from 150 to 200 °C [15]. The failure time is defined as the time required by a fully amorphized cell to lose the stored information. The resistance value for failure has been instead defined as the geometric average between set and reset resistances. The data clearly show an Arrhenius behavior

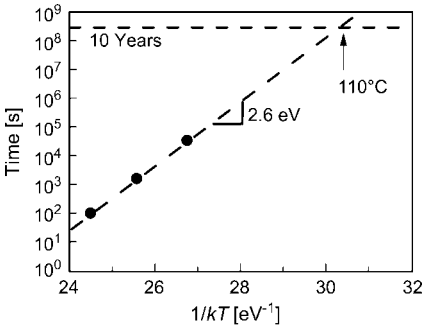


Figure 15.17 Experimental crystallization time of the RESET state as a function of temperature, shown in the Arrhenius plot [15]. A maximum temperature of 110 °C can be tolerated to guarantee 10 years' data retention.

with an activation energy of 2.6 eV, which extrapolates to a data retention capability of 10 years at 110 °C.

A wide range of activation energies have been reported. In general, quite high activation energies have been obtained from $E_a \sim 2.9$ eV for recent fully integrated cells [38] up to 3.5 eV for intrinsic material characterization of GST [17]. Even if such high activation energies are favorable for long retention, however, the physics underlying these experimental values is still not completely understood. The details of the material and integration processes apparently have a large effect, as activation energy was found to be dependent on the presence of capping layers (from a low 2.4 eV for uncapped GST to 2.7 eV for ZnS-SiO_2-capped GST [68]). Furthermore, material doping increases the activation energy (up to 4.4 eV has been reported for O-doped GST [32]).

In principle, the crystallization process during the accelerated retention tests should be described by the same theoretical models for crystal nucleation and growth as used to account for crystallization at much higher temperature (but at much shorter times) during the SET programming pulse (see Section 15.4). Although the conditions for the time window vary over many orders of magnitude (<100 ns during SET, but $>10^2$–10^3 s during retention tests), it has been shown [69] that – remarkably! – the models are indeed able accurately to describe the crystallization processes under both conditions. Moreover, the crystallization statistics also significantly impact the data retention measurements in high-temperature accelerated tests.

15.6.3 Cycling and Failure Modes

PCM cells have been shown to have an intrinsic long programming endurance – that is, up to 10^{12} SET–RESET program cycles [13, 17] (Figure 15.18), which is much superior with respect to Flash technology. In fact, cell endurance has been shown to depend heavily on the interface quality of the heater-GST system and on the possible interdiffusion between GST and adjacent materials. A non-optimized fabrication technology results in devices showing the so-called "first fire" effect [47], namely a

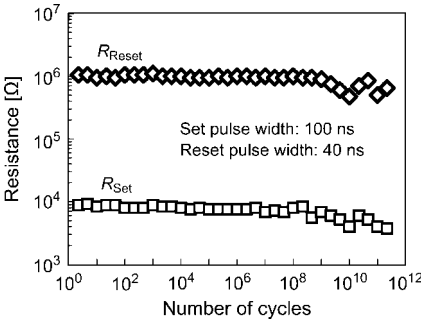


Figure 15.18 Program cycling of a PCM element. (From Ref. [13]).

higher initial programming pulse required for the first cycle of virgin devices. The same devices usually feature poor characteristics in terms of stability during cycling characterization, usually ending in a physical separation of the chalcogenide alloy from the heater (stuck at RESET).

A second failure mode has been also observed (usually called "short-mode failure" or "stuck at SET"), where the devices remain permanently in the highly conductive condition. This phenomenon requires an auxiliary physical mechanism that either forbids the phase-change transition of the GST or creates a conductive parallel path that shunts the cell electrodes. Both cases require a chemical modification of the chalcogenide alloy, suggesting that the interdiffusion of chemical species from adjacent materials plays a role. A careful definition of the materials belonging to the device active region is mandatory in order to achieve good reliability performance.

Finally, current density, as well as the high temperatures (>600 °C) reached in the active region during programming, must be considered as accelerating factors for the previously proposed failure mechanisms (i.e., poor quality interface and contamination). This explains the strong decrease in endurance as a function of the "overcurrent", or, equivalently with energy per pulse larger than required for switching [17]. The data in Figure 15.19 show that endurance at a constant programming current (~700 μA) scales inversely to the reset pulse [70]. The longer the pulse,

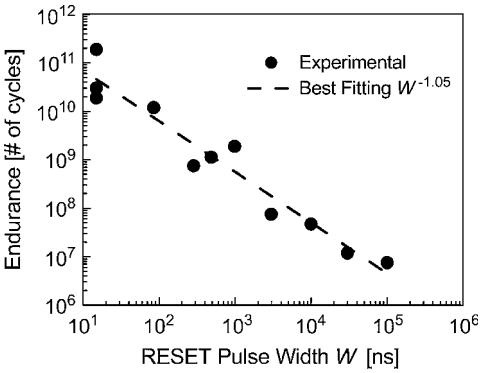


Figure 15.19 Cycling capability as function of the reset pulse window (programming energy per pulse; from Ref. [70]).

the larger the energy released per pulse, and the faster the interface degradation, while the overall energy released up to the bit failure remains almost constant.

15.6.4 Read and Program Disturbs

For most memory technologies, one important concern is the ability of the cell to retain data in the face of spurious voltage transients caused by reading and programming in the memory matrix. For PCM cells such disturbs are not directly induced by voltage pulses but rather by "thermal spikes" that can trigger the crystallization of the metastable amorphous state.

A first failure mechanism can be caused by multiple read accesses of a cell: the small current flowing through the device can induce a localized heating able to accelerate the spontaneous amorphous to crystalline transition. In a second failure mechanism, repeated programming operations on a cell can induce an unwanted heating of the adjacent bits (thermal cross-talk) that can lose the data stored.

Read disturb tests have been described in literature [15, 71], indicating that the repetitive reading of a cell in the high-resistance state with a current below 1 μA allows for a 10-year bit preservation. Such a current is one order of magnitude larger than the current flowing through the cell in standard reading conditions, thereby confirming the robustness to read-disturbs of PCM also in continuous reading, worst-case conditions. Results from program disturb tests on demonstrators have shown that cross-talk is not an issue down at the 90-nm technology node [15]. Thermal simulations, furthermore, confirm program disturb immunity up to at least the 45-nm technology node [18].

15.7 Scaling of Phase-Change Memories

A new memory technology, to be competitive with the existing Flash, must feature a small cell size combined with an excellent scalability beyond the 45-nm technology node. In this section, the scaling potential of the PCM memory is addressed. The main aspects are: (i) scaling of the thermal profiles in order to avoid thermal disturbs at shrinking cell separation; (ii) scaling of the program (RESET) current (and voltage), not only to reduce the program energy/cell, but also because it affects the cell size through the dimensions of the select transistor; and (iii) conservation of the basic material characteristics down to very small dimensions. These aspects will have a crucial impact on the scalability perspectives of PCM technology, and are still to be verified.

15.7.1 Temperature Profile Distributions

As programming of the PCM cell is based on strong heating up to high temperatures (>600 °C, above the PCM material melting point), yet on the other hand the

amorphous RESET state can become unstable at much lower temperatures (<200 °C), it is crucial for the correct operation of a PCM memory that the heating remains very localized and does not affect neighboring cells. Whilst this has been proven for current technologies down to 90 nm (see Section 15.6), it may be less evident to maintain the high temperatures localized in much further scaled technologies, beyond the 45-nm node.

As far as all the linear dimensions are reduced isotropically, the temperature distribution profile indeed does scale. This property transpires from the heat equation:

$$\kappa \cdot \nabla^2 T = g = \rho J^2 \tag{15.1}$$

where κ is the thermal conductivity and g is the heat generation per unit volume. The latter is proportional, via the electrical resistivity, ρ, to the square of the current density, J.

Let us now assume that a new device is fabricated, by shrinking all the linear dimensions by a factor α, but keeping the same boundary conditions (e.g., $T(0) = T_0$) and material properties (e.g., κ, ρ). In the new device it is:

$$\kappa \cdot \nabla'^2 T' = g' = \rho J'^2$$

Since for all the spatial coordinates it is $x' = x/\alpha$, we obtain:

$$\kappa \cdot \nabla'^2 T' = \kappa\alpha^2 \cdot \nabla^2 T' = g' = \rho J'^2$$

That is:

$$\nabla^2 T' = \rho \frac{J'^2}{\alpha^2} \tag{15.2}$$

Provided that $J' = \alpha J$, Equation 15.1 and 15.2 coincide, leading to the same temperature profile, but on a spatial scale uniformly compressed by a factor α.

It follows that if two cells in the original device, at a distance d, do not suffer from cross-talk, then in the isotropically scaled device two cells at a distance d/α will be immune to thermal disturbs. The argument holds as far as $J' = \alpha J$, which will be demonstrated in the following section.

In a more aggressive scaling scheme, only the contact area of the cell is scaled down, without changing so much the thickness of the different layers. In such an anisotropically approach, aiming to more drastically reduce the programming current, cross-talk immunity is no longer granted. However, simulations results show that, without any specific care or materials, thermal disturbs are not expected to slow down cell scaling until the 45-nm node [18].

15.7.2
Scaling of the Dissipated Power and Reset Current

The highest power and programming current is required during the RESET operation, where locally the PCM material must be heated above the melting point.

By using simplified models, the RESET power can be calculated as follows. Even if the programming current pulses last only some tens of nanoseconds, the temperature rise ΔT of the hot spot, where the PCM material eventually melts down, may be computed at steady state. This assumption holds true since the thermal transients in such a small region are characterized by a nanosecond time constant, much faster than the typical current pulse width.

At steady state, the power dissipated by Joule heating (P_J) balances the heat loss ($P_{HL} = \Delta T / R_{TH}$), where R_{TH} is the thermal resistance to the thermal sinks at room temperature (i.e., top and bottom metal layers). On the other hand, the power P_J is proportional to the current, I^2, via the electrical resistance, $P_J = R \cdot I^2$. Using this simple model, the temperature rise ΔT_M, needed to reach the melting temperature, can be written as: $\Delta T_M = P_{J,M}\, R_{TH} = R_{TH}\, R \cdot {I_M}^2$, where I_M is the melting current. It follows:

$$I_M^2 = \Delta T / (R_{TH} \cdot R) \tag{15.3}$$

In the frame of the isotropic scaling rule[4)], as the technology scales, the cell surface area decreases as F^2, but also the distances to the heat sinks decrease as F. The thermal resistance will therefore linearly increase with the scaling factor: $R_{TH} \sim \alpha$ or, equivalently, $R_{TH} \sim F^{-1}$. As for the thermal resistance, the electrical resistance of the PCM cell also increases linearly with geometry scaling ($R = \rho \cdot \text{length}/\text{Area} \sim F^{-1}$). It follows from Equation 15.3 that the melting current and the programming current, which is proportional to the latter, scales as F. The smaller the feature size, the smaller the programming current: $I_{reset} \sim F$ (or $I_{reset} \sim 1/\alpha$). Note that the current density $J = I/A$, will scale as α, as assumed above in deriving Equation 15.2 while discussing the immunity to thermal disturbs.

Although the scaling result is independent of the adopted cell architecture, this does not mean that cell architecture is not important. At a fixed technological node, the cell architecture should be optimized, by accurate design of the geometry and material engineering, to minimize the programming current and the dissipated power (examples of different cell optimizations were provided in Section 15.5).

A more aggressive reduction of the programming current may be obtained by scaling the contact area but not the other dimensions (e.g., the PCM thickness). This choice will mainly affect the thermal and electrical resistances (R and R_{TH} scale faster, i.e., $\sim F^{-2}$ instead of F^{-1}). It follows that $I_{reset} \sim F^2$ (or, $\sim 1/\alpha^2$). The scaling properties of the PCM cell are summarized in Table 15.3[18]. The more aggressive scaling will however pose some problems of manufacture, as the aspect ratio of some cell features (e.g., the thickness to cell size of the PCM material) will increase. Moreover, as discussed above, at this point thermal disturbs may begin to enter the game.

The scalability of the reset current has been addressed experimentally by measuring several test devices with different contact areas [17, 18]. An example of resulting values is given in Figure 15.20 [18]. From this figure it is clear that the reset current follows the reduction of the contact area, and values as low as 50 μA have been

4) Scaling can be indicated either by a dependence on technology feature size, F, or by a dependence on a scaling factor α, with $\alpha \sim 1/F$.

Table 15.3 Scaling rules of PCM cell. (Adapted from Ref. [18]).

	Isotropic	**Aggressive**
Parameters	Scaling factor	Scaling factor
Heater contact area A_{cell}	$1/\alpha^2$	$1/\alpha^2$
Vertical dimensions d	$1/\alpha$	1
Electrical/thermal resistances R	α	α^2
Power dissipation P_{cell}	$1/\alpha$	$1/\alpha^2$
Current I	$1/\alpha$	$1/\alpha^2$
Voltage V_{cell}	1	1
Current density J	α	1

α = scaling factor, $\alpha \sim 1/F$ with F the feature size.

achieved with a com1plete device functionality. The reset current reduction shown in Figure 15.20 can be mainly ascribed to the increase of the heater thermal resistance, R_{TH} caused by the reduction of the contact area.

Data published recently show that I_{reset} is indeed scaling in between $\sim$F and $\sim F^2$ (Figure 15.21) [19]. A scaling behavior stronger than F^2 has been indeed reported by Cho *et al.* [72]. Such a steep dependence, which is well beyond the F^2 theoretical limit, is a strong indication that other cell parameters related to the cell structure and/or material characteristic have been changed.

15.7.3
Voltage Scaling

The required program voltage is determined by the threshold for the electronic switching of the amorphous phase (SET). This voltage value scales with PCM thickness, but is also strongly material dependent; for example, the threshold field is reported to be larger for GST (225) material (threshold field $\sim$30–40 V μm^{-1}) than for the fast-growth doped SbTe materials (threshold field $\sim$14 V μm^{-1}) [24].

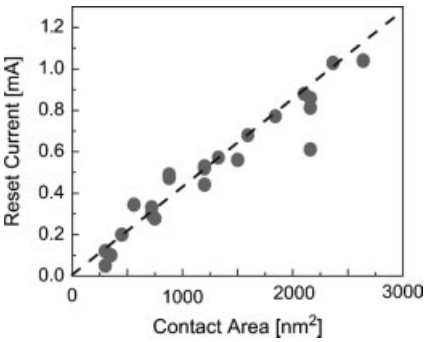


Figure 15.20 Reset current versus contact area (from Ref. [18]). Experimental values (dots), together with scaling trend line (dashed line).

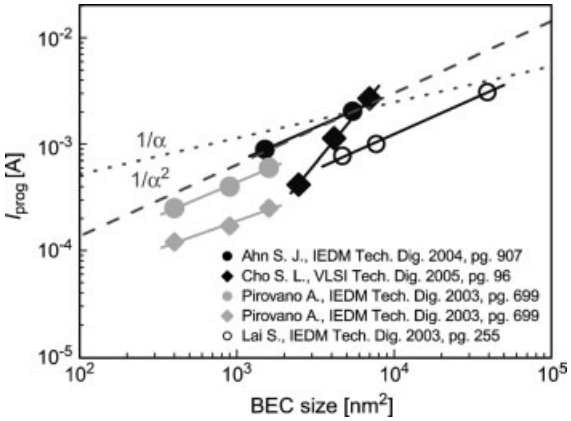


Figure 15.21 Plot of published experimental values of RESET program current versus bottom electrode contact (BEC) size, compared with the $1/\alpha$ and $1/\alpha^2$ scaling law ($\alpha =$ scaling factor, $\alpha \sim 1/F$ with F the feature size). The data are related to different cell structures. (Figure from Ref. [19]).

Furthermore, in Figure 15.3 there was shown to be a minimum value of the threshold switching voltage determined by material (bandgap) and/or contact characteristics. This is why, in Table 15.3, the voltage drop across the cell was considered not dependent on feature scaling. This limitation may pose a constraint on the maximum voltage which should be sustained by the bipolar selectors or by the gate oxide of the MOSFETs across the unselected cells of the array. Beyond the 45-nm technology node, an accurate design of the selecting device will therefore become mandatory.

15.7.4 Cell Size Scaling

In order to assess the scaling of PCM cell size, let us consider a typical MOS-select transistor PCM cell layout (Figure 15.22). The cell size can be calculated as $\sim(\mathrm{W} + 1\mathrm{F}) \times (4.5\ \mathrm{F})$, where W is the width of the access transistor. For a minimal device, $W = 1$ F, the cell size would be $\sim 9\ \mathrm{F}^2$. However, this is the ideal case as in practice, $W/L \gg 1$ to obtain the required drive current to reset the cell through the access transistor. For $W/L > 1$, splitting the gate (the so-called "dual gate concept") is favorable to minimize the cell area [13], resulting in a cell area of $\sim 6\mathrm{W} \times (1/2\ W + \mathrm{F})$.

Taking the I_{reset} values from the ITRS roadmap (predicting, in agreement with the scaling laws derived in Section 15.7.2 above, a scaling according $I_{\mathrm{reset}} \sim \alpha^{1.5}$) [73], as well as the predicted scaling of the drive current with the technology node into account, the minimum access transistor size and the corresponding PCM cell size can be predicted as a function of the technology node (Figure 15.23). From these values it transpires that, at the 45-nm node, a cell size reduction to $<15\ \mathrm{F}^2$ would become possible (with $I_{\mathrm{reset}} \sim 100\ \mu\mathrm{A}$ and $\mathrm{W} \sim 2.7$ F).

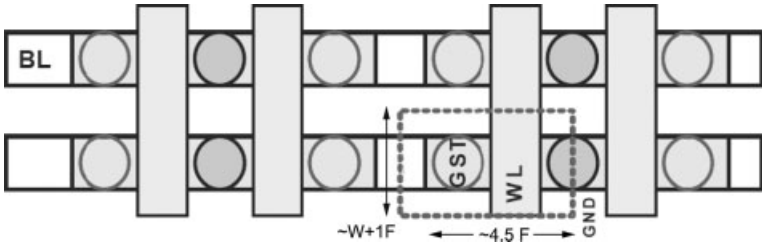


Figure 15.22 Layout schematic of PCM cell as shown in Figure 15.15. BL = bit line (M1); WL = word line (Poly-Si gate line); GND = ground connection (M0). The cell area (dashed rectangle) is approximately $(W + 1\,F)\cdot(4.5\,F)$, where W is the transistor width (in number of Fs) and F the technology feature size. (Adapted from Ref. [53]).

Further room may be obtained by using alternative solutions for ultra-scaled MOS transistors, such as ultra-thin body fully depleted on SOI or Double Gate (or FinFET). These devices are expected to have an improved driving current capability up to $2\,mA\,\mu m^{-1}$ (forecasted to be 2010) [73]. Long-term solutions may also involve the adoption of vertical MOS structures that take advantage of having only a single contact (and hence a reduced area on silicon) and an improved W (by about a factor of 3) with respect to planar solutions.

Figure 15.24 shows the projection of the PCM scaling trend. The PCM cell size will take full advantage of technology scaling, as no intrinsic limitations are expected to halt further scaling. On the other hand, both NOR and NAND scaling are expected to slow down due to scaling limitations (due to high program drain voltage for NOR and electrostatic field coupling for NAND). As a consequence, the PCM cell size will reach the NOR cell size at about the 45-nm node, and may even reach the NAND cell size at about the 32-nm node. Moreover, the memory is ideally suited to store more than two

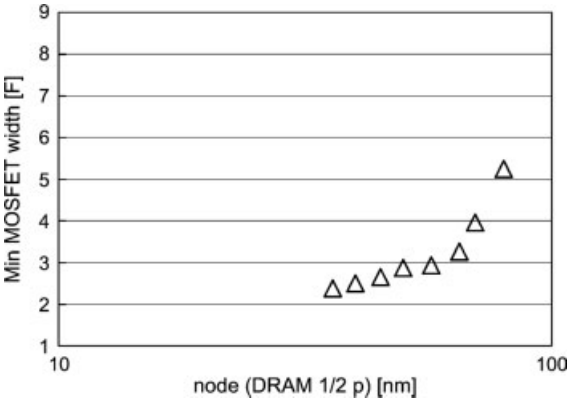


Figure 15.23 Predicted width W (in multiples of feature size F) of the PCM cell MOS select transistor required to drive the required reset current [73] as function of technology node (defined as DRAM half pitch).

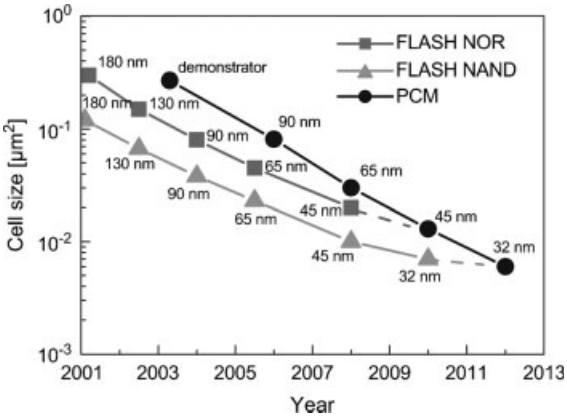


Figure 15.24 Scaling trends for NAND, NOR and PCM (with bipolar select transistor). The phase-change memory technology is expected to reach the same Flash-NAND size at the 32-nm technology feature size. (Figure from Ref. [19]).

levels per cell. This approach may become a viable option to further reduce the cost per bit of PCM devices, if reliable multi-level programming becomes feasible.

15.7.5
Scaling and Cell Performance: Figure of Merit for PCM

A low programming current can indeed be reached by tightly confining the current flow and the corresponding heat generation. However, this usually results in a high cell resistance, and there is an upper limit to the acceptable SET resistance value. The larger the resistance, the lower the cell current during the read operation, and the longer the time needed to charge the bit-line capacitance. In practice, for a read operation to occur within 50 ns the SET resistance should be kept at less than 50 kΩ (i.e., corresponding to a minimum read current in the SET state of a few μAs), but this constraint leads to a trade-off between programming current, and to the introduction of a new figure of merit, the product $R_{set} \cdot I_{melt}$, to compare different devices [19]. Figure 15.25 shows I_{melt} versus R_{set} as derived from published results, where the constant $R_{set} \cdot I_{melt}$ lines are highlighted by the dashed lines. Different cell architectures and material systems correspond to different $R_{set} \cdot I_{melt}$ values. The data clearly show the potential of PCM cells to reach programming currents of few tens of μAs.

15.7.6
Physical Limits of Scaling

To date, very few data are available regarding size effects on PCM behavior. However, from optical memory measurements, phase-change mechanisms seem scalable to at least 5 nm [74]. On the other hand, for both nanosized elements or films thinner than 20 nm, shifts in the crystallization behavior of GST (225) have been observed [75]. Yet, it is unclear if these are intrinsic effects or they are related to

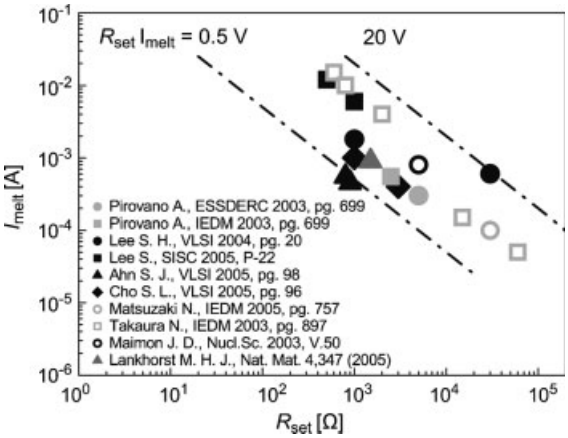


Figure 15.25 I_{melt} versus R_{set} as derived from published results. The constant $R_{set}{\cdot}I_{melt}$ lines are highlighted by the dashed lines. Different cell architectures and material systems correspond to different $R_{set}{\cdot}I_{melt}$ values, proposed as a figure of merit for PCM cells. The data clearly show the potential of PCM cells to reach programming currents of a few tens of μA. (Figure from Ref. [19]).

the preparation method (patterning damage and porosity of thinner films), or if they would inhibit phase switching. What is clear is that both the fabrication technology and behavior of nanosized PCM structures require further exploration in the near future.

15.8 Conclusions

During recent years, PCM have evolved from an interesting new concept to a viable memory technology, based on the use of improved and faster switching materials (<100 ns) and on an improved understanding of phase transformation and electronic switching processes in chalcogenide materials. PCM further shows excellent reliability properties, such as good data retention (10 years at 110 °C) and very high endurance (up to 10^{12} cycles, compared to $<10^6$ for FLASH), while the optimization of cell design has resulted in a drastic reduction of the program current (down to few hundred μA). In addition, integration technology has matured to a point where large demonstrator circuits (up to 256 Mb, in 100-nm technology) have already been built. Perhaps more importantly, the scaling potential of PCM has been assessed, and is expected to result in small cell sizes (in the range of 10 F^2 or smaller) for technologies of 45 nm and below, this being concomitant with a further reduction in program currents. For these reasons, PCM is expected eventually to replace the no-longer-scaling Flash technologies.

References

1 Kurata, H., Otsuga, K., Kotabe, A., Kajiyama, S., Osabe, T., Sasago, Y., Nerumi, S., Tosami, K., Kamohara, S. and Tsuchiya, O. (2006) The impact of random telegraph signals on the scaling of multilevel Flash memories. *IEEE Symp. VLSI Circuits, Tech. Digp*, pp. 140–141.

2 Shin, Y. (2005) Non-volatile memory technologies for beyond 2005. *IEEE Symp. VLSI Circuits, Tech. Digp*, pp. 156–159.

3 Dewald, J.F., Pearson, A.D., Northover, W.R. and Peck, W.F. (1962) *Journal of the Electrochemical Society*, **109**, 243.

4 Ovshinsky, S.R. (1968) Reversible electrical switching phenomena in disordered structures. *Physical Review Letters*, **21**, 1450–1453.

5 Ovshinsky, S.R. and Fritzsche, H. Amorphous semiconductors for switching, memory, and imaging applications. *IEEE Trans. Elec.* Dev., Vol. ED-20, No. 2, February 1973, pp. 91–105.

6 Neale, R., Nelson, D. and Moore, G. (1970) Nonvolatile and reprogrammable, the read-mostly memory is here. *Electronics*, **43**, 56–60.

7 Maimon, J., Hunt, K., Rodgers, J., Burcin, L. and Knowles, K. Circuit demonstration of radiation hardened chalcogenide non-volatile memory. Proceedings of Aerospace Conference, Vol. 5, pp. 5_2373–5_2379.

8 Yamada, N., Ohno, E., Nishiuchi, K., Akahira, N. and Takao, M. (1991) Rapid-phase transitions of GeTe-Sb2Te3 pseudobinary amorphous thin films for an optical disk memory. *Journal of Applied Physiology*, **69**, 2849–2857.

9 Kolobov, A.V., Fons, P., Frenkel, A.I., Ankudinov, A.L., Tominaga, J. and Uruga, T. (2004) Understanding the phase-change mechanism of rewritable optical media. *Nature Materials*, **3**, 703–708.

10 Wicker, G. (1996) A comprehensive model of submicron chalcogenide switching devices, Ph. D. Dissertation, Wayne State University, Detroit, MI.

11 Wicker, G. (1999) Nonvolatile, high density, high performance phase change memory. in Proceedings SPIE Conference on Electronics and Structures for MEMS, Vol. 3891, SPIE, The International Society for Optical Engineering, Bellingham, WA, pp. 2–9.

12 Lai, S. and Lowrey, T. (2001) OUM-A 180 nm nonvolatile memory cell element technology for stand alone and embedded applications, *IEEE International Electron Devices Meeting Technical Digest*, pp. 803–806.

13 Pellizzer, F., Pirovano, A., Ottogalli, F., Magistretti, M., Scaravaggi, M., Zuliani, P., Tosi, M., Benvenuti, A., Besana, P., Cadeo, S., Marangon, T., Morandi, R., Piva, R., Spandre, A., Zonca, R., Modelli, A., Varesi, E., Lowrey, T., Lacaita, A., Casagrande, G., Cappelletti, P. and Bez, R., Novel μtrench phase-change memory cell for embedded and stand-alone non-volatile memory applications, 2004 Symposium on VLSI Technology Digest of Technical Papers, pp. 18–19.

14 Hwang, Y.N., Lee, S.H., Ahn, S.J., Lee, S.Y., Ryoo, K.C., Hong, H.S., Koo, H.C., Yeung, F., Oh, J.H., Kim, H.J., Jeong, W.C., Park, J.H., Horii, H., Ha, Y.H., Yi, J.H., Koh, G.H., Jeong, G.T., Jeong, H.S. and Kim, K. (2003) Writing current reduction for high-density phase-change RAM, *IEEE International Electron Devices Meeting Technical Digest*, pp. 893–896.

15 Pirovano, A., Redaelli, A., Pellizzer, F., Ottogalli, F., Ielmini, D., Lacaita, A.L. and Bez, R., Reliability study of phase-change nonvolatile memories. IEEE Transactions on Device and Materials Reliability, Vol. 4, No. 3, September 2004, pp. 422–427.

16 Kim, K. *et al.* (2005) Reliability investigations for manufacturable high density PRAM. *IRPS Tech. Dig.*, 157–162.

17 Lai, S. (2003) Current status of phase change memory and its future, *IEEE International Electron Devices Meeting Technical Digest*, pp. 255–258.

18 Pirovano, A., Lacaita, A.L., Benvenuti, A., Pellizzer, F., Hudgens, S. and Bez, R. (2003) Scaling analysis of phase-change memory technology. *IEDM Technical Digest*, 699–702.

19 Lacaita, A.L. Progress of phase-change non volatile memory devices, presented at European Phase Change Ovonic Science (Joint E*PCOS-IMST Workshop), Grenoble, France, May 29–31, 2006; http://www.epcos.org.

20 Lacaita, A.L., Redaelli, A., Ielmini, D., Pellizzer, F., Pirovano, A., Benvenuti, A. and Bez, R. (2004) Electrothermal and phase-change dynamics in chalcogenide-based memories, *IEEE International Electron Devices Meeting Technical Digest*, pp. 911–914.

21 Wuttig, M., Klein, M., Kalb, J., Lecner, D. and Spaepen, F.,Ultrafast data storage with phase change media: from crystal structures to kinetics, Presented at the 5th European Phase Change Ovonic Science symposium (Joint E*PCOS-IMST Workshop), Grenoble, France, May 29–31, 2006; http://www.epcos.org.

22 Hudgens, S. and Johnson, B. (2004) Overview of phase-change chalcogenide nonvolatile memory technology. *MRS Bulletin*, **29**, 829–832.

23 Libera, M. and Chen, M. (1990) Multilayered thin-film materials for phase-change erasable storage. *MRS Bulletin*, **15**, 40–45.

24 Lankhorst, M.H.R., Ketelars, B.W.S.M.M. and Wolters, R.A.M. (2005) Low-cost and nanoscale non-volatile memory concept for future silicon chips. *Nature Materials*, **4**, 347–352.

25 Borg, H., Lankhorst, M., Meinders, E. and Leibbrandt, W. (2001) Phase-change media for high-density optical recording. Materials Research Society Symposium Proceedings, Materials Research Society, Vol. 674, V1.2.1–V1.2.10.

26 Miao, S.S., Shi, L.P., Zhao, R., Tan, P.K., Lim, K.G., Li, J.M. and Chong, T.C. Temperature dependence of phase change random access memory cell. Extended Abstracts, 2005 International Conference on Solid State Devices and Materials (SSDM), Kobe 2005, pp. 1052–1053.

27 Moffatt, W.G. *The Handbook of Binary Phase Diagrams*, Genum Publishing Companyp, p. 7/91.

28 Khulbe, P.K., Hurst, T., Horie, M. and Mansuripur, M. (2002) Crystallization behavior of Ge-doped eutectic Sb 70 Te 30 films in optical disks. *Applied Optics*, **41**, 6220–6229.

29 Yoon, S.-M., Lee, N.-Y., Ryu, S.-O., Choi, K.-J., Park, Y.-S., Lee, S.-Y., Yu, B.-C., Kang, M.-J., Choi, S.-Y. and Wuttig, M. Lower power and higher speed operation of phase-change memory device using antimony selenide (Sb_xSe_{1-x}), Extended Abstracts of the 2005 International Conference on Solid State Devices and Materials (SSDM), Kobe 2005, pp. 1050–1051.

30 Lee, H. and Kang, D.-H. Indium selenide based phase change memory, Extended Abstracts of the 2004 International Conference on Solid State Devices and Materials (SSDM), Tokyo, 2004, pp. 646–647.

31 Horii, H., Yi, J.H., Park, J.H., Ha, Y.H., Baek, I.G., Park, S.O., Hwang, Y.N., Lee, S.H., Kim, Y.T., Lee, K.H., Chung, U.-In. and Moon, J.T. A novel cell technology using N-dopes GeSbTe films for Phase change RAM, 2003 Symposium on VLSI Technology Digest of Technical Papers, pp. 177–178.

32 Matsuzaki, N., Kurotsuchi, K., Matsui, Y., Tonomura, O., Yamamoto, N., Fujisaki, Y., Kitai, N., Takemura, R., Osada, K., Hanzawa, S., Moriya, H., Iwasaki, T., Kawahara, T., Takaura, N., Terao, M., Matsuoka, M. andMoniwa, M. (2005) Oxygen-doped GeSbTe phase-change memory cells featuring 1.5-V/100-μA standard 0.13-μm CMOS operations, *IEEE International Electron Devices Meeting Technical Digest*, pp. 738–741.

33 Pirovano, A., Lacaita, A.L., Benvenuti, A., Pellizzer, F. and Bez, R. (2004) Electronic switching in phase-change memories.

IEEE Transactions on Electron Devices, **51**, 452–459.

34 Raoux, S., Rettner, C.T. and Jordon-Sweet, J.L., Crystallization behavior of phase change nanostructures. Presented at the European Phase Change Ovonic Science (EPCOS 2005) Symposium, 3–6 September, 2005, Cambridge, UK; http://www.epcos.org.

35 Kolobov, A.V., Fons, P., Tominaga, J., Frenkel, A.I., Ankudinov, A.L. and Uruga, T. (2005) Local structure of Ge-Sb-Te and its modification upon the phase transition. *Journal of Ovonic Research*, **1**, 21–24.

36 Welnic, W., Pamungkas, A., Detemple, R., Steimer, C., Bluegel, S. and Wuttig, M. (2006) Unraveling the interplay of local structure and physical properties in phase-change materials. *Nature Materials*, **5**, 56–62.

37 Kalb, J., Spaepen, F. and Wuttig, M. (2003) Calorimetric measurements of phase transformations in thin films of amorphous Te alloys used for optical data storage. *Journal of Applied Physiology*, **93**, 2389–2393.

38 Pellizzer, F., Benvenuti, A., Gleixner, B., Kim, Y., Johnson, B., Magistretti, M., Marangon, T., Pirovano, A., Bez, R. and Atwood, G., A 90 nm phase change memory technology for stand-alone non-volatile memory applications, 2006 Symposium on VLSI Technology Digest of Technical Papers, pp. 122–123.

39 Chen, H.S. (1980) Glassy metals. *Reports on Progress in Physics*, **43**, 353–432.

40 Wei, J. and Gan, F. (2003) *Thin Solid Films*, **441**, 292–297.

41 Peng, C., Cheng, L. and Mansuripur, M. (1997) Experimental and theoretical investigations of laser-induced crystallization and amorphization in phase-change optical recording media. *Journal of Applied Physiology*, **62**, 4183.

42 Kelton, K.F. (1991) Crystal nucleation in liquids and glasses. *Solid State Physics*, **45**, 75.

43 Gille, T., Goux, L., Lisoni, J., De Meyer, K. and Wouters, D.J. (2006) Impact of material crystallization characteristics on the switching behavior of the phase change memory cell. in Chalcogenide-Based Phase-Change Materials for Reconfigurable Electronics, Materials Research Society Symposium Proceedings 918E, (eds A.H. Edwards, P.C. Taylor, J. Maimon and A. Kolobov), Warrendale, PA, paper no. 0918-H06-02-G07-02.

44 Mott, N.F. and Davis, E.A. (1967) *Electronic processes in non-crystalline materials*, Clarendon Press, Oxford.

45 Popescu, C. (1975) The effect of local non-uniformities on thermal switching and high field behaviour of structures with chalcogenide glasses. *Solid-State Electron*, **18**, 671–681.

46 Owen, A.E., Robertson, J.M. and Main, C. (1979) The threshold characteristics of chalcogenide-glass memory switches. *Journal of Non-Crystalline Solids*, **32**, 29–52.

47 Adler, D., Henisch, H.K. and Mott, S.D. (1978) The mechanism of threshold switching in amorphous alloys. *Reviews of Modern Physics*, **50**, 209–220.

48 Adler, D., Shur, M.S., Silver, M. and Ovshinsky, S.R. (1980) Threshold switching in chalcogenide-glass thin films. *Journal of Applied Physiology*, **51**, 3289–3309.

49 Ielmini, D., Lacaita, A.L., Mantegazza, D., Pellizzer, F. and Priovano, A. (2005) Assessment of threshold switching dynamics in phase-change chalcogenide memories, *IEEE International Electron Devices Meeting Technical Digest*, pp. 877–880.

50 Pirovano, A., Lacaita, A.L., Pellizzer, F., Kostylev, S.A., Benvenuti, A. and Bez, R. (2004) Low-field amorphous state resistance and threshold voltage drift in chalcogenide materials. *IEEE Transactions on Electron Devices*, **51**, 714–719.

51 Kunghia, K., Shakoor, Z., Kasap, S.O. and Marshall, J.M. (2005) Density of localized electronic states in a-se from electron time-of-flight photocurrent measurements. *Journal of Applied Physiology*, **97**, 033706-1–033706-11.

52 Chen, Y., Chen, C.F., Chen, C.T., Yu, J.Y., Wu, S., Lung, S.L., Liu, R. and Lu, C. (2003) An access-transistor-free (0T/1R) non-volatile resistance random access memory (RRAM) using a novel threshold switching, self-rectifying chalcogenide device, *IEEE International Electron Devices Meeting Technical Digest*, pp. 905–908.

53 Kim, K., Jeong, G., Jeong, H. and Lee, S., Emerging memory technologies, Proceedings of the IEEE 2005 Custom Integrated Circuits Symposium, 18–21 September 2005, pp. 423–426.

54 Kim, Y.T. *et al*. (2004) (Samsung) Extended Abstracts of the 2004 International Conference on Solid State Devices and Materials, Tokyo D-3-2, pp. 244–245.

55 Yi, H., Ha, Y.H., Park, J.H., Kuh, B.J., Horii, H., Kim, Y.T., Park, S.O., Hwang, Y.N., Lee, S.H., Ahn, S.J., Lee, S.Y., Hong, J.S., Lee, K.H., Lee, N.I., Kang, H.K., Chung, U. andMoon, J.T. (2003) Novel cell structure of PRAM with thin metal layer inserted GeSbTe, *IEEE International Electron Devices Meeting Technical Digest*, pp. 901–904.

56 Alberici, S.G., Zonca, R. and Pashmakov, B. (2004) Ti diffusion in chalcogenides: a ToF-SIMS depth profile characterization approach. *Applied Surface Science*, **231/232**, 821–825.

57 Yoon, S.-M., Lee, N.-Y., Ryu, S.-O., Park, Y.-S., Lee, S.-Y., Choi, K.-J. and Yu, B.-G. (2005) Etching characteristics of $Ge_2Sb_2Te_5$ using high-density helicon plasma for the nonvolatile phase-change memory applications. *Japanese Journal of Applied Physics*, **44**, L869–L872.

58 Pellizzer, F., Spandre, A., Alba, S. and Pirovano, A. (2004) Analysis of plasma damage on phase change memory cells, 2004 IEEE International Conference on Integrated Circuit Design and Technology, p. 227.

59 Takaura, N., Terao, M., Kurotsuchi, K., Yamauchi, T., Tonomura, O., Hanaoka, Y., Takemura, R., Osada, K., Kawahara, T. andMatsuoka, H. (2003) A GeSbTe phase-change memory cell featuring a tungsten heater electrode for low-power, highly stable, and short-read-cycle operations, *IEEE International Electron Devices Meeting Technical Digest*, pp. 897–900.

60 Song, Y.J., Ryoo, K.C., Hwang, Y.N., Jeong, C.W., Lim, D.W., Park, S.S., Kim, J.I., Kim, J.H., Lee, S.Y., Kong, J.H., Ahn, S.J., Lee, S.H., Park, J.H., Oh, J.H., Oh, Y.T., Kim, J.S., Shin, J.M., Park, J.H., Fai, Y., Koh, G.H., Jeong, G.T., Kim, R.H., Lim, H.S., Park, I.S., Jeong, H.S. and Kim, Kinam, Highly reliable 256 Mb PRAM with advanced ring contact technology and novel encapsulating technology, 2006 Symposium on VLSI Technology Digest of Technical Papers, pp. 118–119.

61 Ha, Y.H., Yi, J.H., Horii, H., Park, J.H., Joo, S.H., Park, S.O., Chung, U.-In. and Moon, J.T., An edge contact type cell for phase change RAM featuring very low power consumption, 2003 Symposium on VLSI Technology Digest of Technical Papers, pp. 175–176.

62 Pirovano, A., Pellizzer, F., Redaelli, A., Tortorelli, I., Varesi, E., Ottogalli, F., Tosi, M., Besana, P., Cecchini, R., Piva, R., Magistretti, M., Scaravaggi, M., Mazzone, G., Petruzza, P., Bedeschi, F., Marangon, T., Modelli, A., Ielmini, D., Lacaita, A.L. and Bez, R. (2005) Trench phase-change memory cell engineering and optimization, Proceedings ESSDERC 2005, pp. 313–316.

63 Ahn, S.J., Hwang, Y.N., Song, Y.J., Lee, S.H., Lee, S.Y., Park, J.H., Jeong, C.W., Ryoo, K.C., Shin, J.M., Park, J.H., Fai, Y., Oh, J.H., Koh, G.H., Jeong, G.T., Joo, S.H., Choi, S.H., Son, Y.H., Shin, J.C., Kim, Y.T., Jeong, H.S. and Kim, K., Highly reliable 50 nm contact cell technology for 256 Mb PRAM, 2005 Symposium on VLSI Technology Digest of Technical Papers, pp. 98–99.

64 Haring-Bolívar, P., Merget, F., Kim, D.-H., Hadam, B. and Kurz, H. Lateral design for phase change random access memory cells with low-current consumption. Presented at the 3rd European Phase Change Ovonic Science Symposium (EPCOS 2004),

Balzers, Principality of Liechtenstein, September 4–7, 2004; http://www.epcos.org.

65 Happ, T.D., Breitwisch, M., Schrott, A., Philipp, J.B., Lee, M.H., Cheek, R., Nirschl, T., Lamorey, M., Ho, C.H., Chen, S.H., Chen, C.F., Joseph, E., Zaidi, S., Burr, G.W., Yee, B., Chen, Y.C., Raoux, S., Lung, H.L., Bergmann, R. and Lam, C., Novel one-mask self-heating pillar phase change memory, 2006 Symposium on VLSI Technology Digest of Technical Papers, pp. 120–121.

66 Kang, D.-H., Ahn, D.-H., Kwon, M.-H., Kwon, H.-S., Kim, K.-B., Seok Lee, K. and Cheong, B.-ki. (2003) Lower voltage operation of a phase change memory device with a highly resistive TiON layer. *Japanese Journal of Applied Physics*, **42**, 2382–2386.

67 Bedeschi, F., Resta, C., Khouri, O., Buda, E., Costa, L., Ferraro, M., Pellizzer, F., Ottogalli, F., Pirovano, A., Tosi, M., Bez, R., Gastaldi, R. and Casagrande, G., An 8 Mb demonstrator for high-density 1.8 V phase-change memories. 2004 Symposium on VLSl Circuits Digest of Technical Papers, pp. 442–445.

68 Friedrich, I., Weidenhof, V., Njoroge, W., Franz, P. and Wuttig, M. (2000) Structural transformations of $Ge_2Sb_2Te_5$ films studied by electrical resistance measurements. *Journal of Applied Physiology*, **87**, 4130.

69 Redaelli, A., Ielmini, D., Lacaita, A.L., Pellizzer, F., Pirovano, A. and Bez, R. (2005) Impact of crystallization statistics on data retention for phase change memories, *IEEE International Electron Devices Meeting Technical Digest*, pp. 742–745.

70 Ottogalli, F., Priovano, A., Pellizzer, F., Tosi, M., Zuliani, P., Bonetalli, P. and Bez, R., Phase-change memory technology for embedded applications. Proceedings, 34th European Solid-State Device Research Conference (ESSDERC 2004), Leuven, Belgium, 21–23 September 2004, pp. 293–296.

71 Osada, K., Kawahara, T., Takemura, R., Kitai, N., Takaura, N., Matsuzaki, N., Kurotsuchi, K., Moriya, H. and Moniwa, M., Phase change RAM operated with 1.5-V CMOS as low cost embedded memory, Proceedings, IEEE 2005 Custom Integrated Circuits Symposium, 18–21 September 2005, pp. 431–434.

72 Cho, S.L. *et al.* (2005) Symposium on VLSl Circuits Digest of Technical Papers, p. 96.

73 International Technology Roadmap for Semiconductors (ITRS), 2005 edition, Process Integration, Devices, and Structures, accessible through: http://www.itrs.net.

74 Wright, D., Aziz, M.M., Armand, M., Senkander, S. and Yu, W. Can we reach Tbit/sq.in. storage densities with phase-change media? Presented at the 3rd European Phase Change Ovonic Science symposium (EPCOS 2004), Balzers, Principality of Liechtenstein, September 4–7, 2004; http://www.epcos.org.

75 Raoux, S., Rettner, C.T., Jordan-Sweet, J.L., Deline, V.R., Philipp, J.B. and Lung, H.-L. Scaling properties of phase change nanostructures and thin films. Presented at the 5th European Phase Change Ovonic Science symposium (EPCOS 2006), May 29–31, 2006 Grenoble, France; http://www.epcos.org.

16
Memory Devices Based on Mass Transport in Solid Electrolytes

Michael N. Kozicki and Maria Mitkova

16.1
Introduction

As standard semiconductor nanoelectronic devices approach their scaling limits (see Chapters 00 on "FET", 11 on "Flash", and 12 on "DRAM"), concepts beyond traditional purely charge-based functionality offer additional opportunities. One such paradigm goes by the name "nanoionics". Whereas nanoelectronics involves the movement of electrons within their nanostructured settings, nanoionics concerns materials and devices that rely on ion transport and chemical change at the nanoscale. Rising interest in nanoionics has been fuelled by the wide range of demonstrated and potential applications so that the field has been equated in significance by some with nanoelectronics [1].

It is impossible to discuss nanoionics without introducing the basic principles of electrochemistry. As the name suggests, electrochemistry deals with the relationship between electricity and chemical change. In many respects, batteries are the prime example of the application of electrochemical principles; the movement of ions and the change in their oxidation state within the cell is used to release electrical energy over time. However, since ions not only carry charge but also have a significant mass, ion transport can be seen as a means to *move material* in a controlled manner. For example, a metal atom that becomes *oxidized* at one location can be moved as a cation through an electrolyte by an electric field. On receiving an electron at another location, the displaced ion is *reduced* and becomes a neutral metal atom again. In this situation, *the net change in the system is the redistribution of mass* – material is removed from one location and deposited at another using energy from an external power source. The world of electronics has benefited from such "deposition electrochemistry" for many decades. Electroplating, in which metal ions in a liquid solution are reduced to create a uniform metal film, is used in printed circuit boards and packages, and in the processes used to make copper interconnect within integrated circuits. In such cases, physical dimensions, such as electrode spacing, are typically quite large and the electric fields relatively small. The term nanoionics is

Nanotechnology. Volume 3: Information Technology I. Edited by Rainer Waser

ISBN: 978-3-527-31738-7

applied when electrochemical effects occur in materials and devices with interfaces, for example electrodes or electrochemically different material phases, that are closely spaced – perhaps by a few tens of nanometers, or less. In this size regime, the functionality of ionic systems is quite different from the macro-scale versions, but in a highly useful manner. For example, internal electric fields and ion mobilities are relatively high in nanoionic structures and this, combined with the short length scales, results in fast response times. In addition, whereas deposition electrochemistry and many batteries use liquids as ion transport media, nanoionics can take advantage of the fact that a variety of solid materials are excellent electrolytes, largely due to effects which dominate at the nanoscale. This allows nanoionic devices based on *solid electrolytes* to be more readily fabricated using techniques common to the integrated circuit industry, and also facilitates the marriage of such devices with mainstream integrated electronics. Indeed, *in-situ* changes may be controlled by the integrated electronics, leading to electronic-ionic system-on-chip (SoC) hybrids.

In this chapter, we describe the basic electrochemistry, materials science and potential applications in information technology of mass-transport devices based on solid electrolytes and nanoionic principles. The electrodeposition of even nanoscale quantities of a noble metal such as silver can produce localized *persistent* – but *reversible* – changes to macroscopic physical or chemical characteristics; such changes can be used to control behavior in applications that go well beyond purely electronic systems. Of course, electrical resistance will change radically when a low resistivity electrodeposit (e.g., in the tens of $\mu\Omega \cdot$ cm or lower) is deposited on a solid electrolyte surface which has a resistivity many orders of magnitude higher. This resistance change effect has a variety of applications in memory and logic. Here, emphasis will be placed on low-energy, non-volatile memory devices which utilize such resistance changes to store information.

16.2 Solid Electrolytes

16.2.1 Transport in Solid Electrolytes

The origins of solid-state electrochemistry can be traced back to Michael Faraday, who performed the first electrochemical experiments with Ag_2S and discovered that this material was a good ion conductor [2]. Subsequently, greater emphasis was placed on liquid electrolytes and their use in plating systems and battery cells, until the 1960s and 1970s when a significant rise in interest was noted in solid-state electrochemistry. This renewed attention was spurred in part by the development of novel batteries which had a particularly high power-to-weight ratio due to the use of solid electrolyte, mainly beta-alumina, which is an excellent conductor of sodium ions [3]. Even though these solid materials were clearly different from their liquid counterparts, many of the well-known principles developed in the field of liquid electrochemistry were found to be applicable to the solid-state systems. One major difference between most

solid and liquid ion conductors is that in solids, the moving ions are of only one polarity (cations or anions) and the opposite polarity species is fixed in the supporting medium. This has a profound effect on the types of structure that can be used for mass transport (this subject will be outlined later in the chapter). The solid electrolyte family currently includes crystalline and amorphous inorganic solids, as well as ionically conducting polymers. In general, the best solid electrolytes have high ionic but low electronic conductivity, chemical and physical compatibility with the electrodes used, thermodynamic stability over a wide temperature range, and the ability to be processed to form continuous mechanically stable thin film structures.

The mobile ions in a solid electrolyte sit in potential wells separated by low potential barriers, typically in the order of a few tenths of an electron-volt (eV), or less. The ions possess kinetic energy, governed by Boltzmann statistics, and so at finite temperature will constantly try (with around 10^{12} attempts per second) to leave their low-energy sites to occupy energetically similar sites within the structure. Thermal diffusion will result from this kinetic energy, driving ions down any existing concentration gradient until a uniform concentration is achieved. Subsequent movement of the ions produces no net flux in any particular direction. The application of an electric field to the electrolyte effectively reduces the height of the barriers along the direction of the field, and this increases the probability that an ion will hop from its current potential well to a lower energy site (see Figure 16.1) [4]. An ion current therefore results, driven by the field. It should be noted that, unlike electrons, ions in a solid are constrained to move through a confining network of narrow channels. These pathways may be a natural consequence of order in the material, as in the case of the interstitial channels present along certain directions in crystalline materials, or they may be a result of long-range disorder, as in amorphous (glassy) and/or nanoscopically porous materials. Glassy electrolytes, typically metal oxides, sulfides or selenides, are of particular interest as they can contain a wider variety of routes for cation transport than purely crystalline materials. This is a major reason for the interest in these materials – and for Group VI glasses in particular – and why they feature heavily in this chapter.

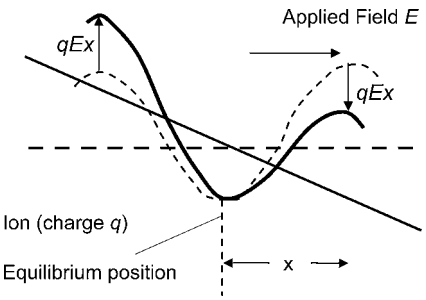


Figure 16.1 Change in the height of a potential barrier between shallow wells due to the application of an electric field. The effective barrier height is reduced by qEx, where E is the electric field and x is the barrier width.

Considering the above dependence of ion hopping on ion energy and barrier height, it should be no surprise that the expression for ion conductivity in the electrolyte is

$$\sigma = \sigma_0 \exp(-E_a/kT) \tag{16.1}$$

where k is Boltzmann's constant, T is absolute temperature, E_a is the activation energy for conduction, and the pre-exponential term σ_0 depends on several factors, including the mobile ion concentration (e.g., σ_0 is in the order of $10^4\,\Omega^{-1}\,cm^{-1}$ for >10 atom% Ag in Ge-S electrolytes [5]). As is evident from Equation 16.1, the activation energy is a major factor in determining ion conductivity. It is directly related to the structure of the host and to the existence of the conduction pathways, both of which govern the effective barrier height. Obviously, the smaller E_a is, the higher the conductivity and the better the electrolyte is. Since E_a is around 0.2–0.3 eV in Ag-saturated Ge-Se and Ge-S electrolytes, the above values lead to ion conductivities in the range 10^{-2}–$10^{-4}\,S\,cm^{-1}$. Just as with electron and hole conduction, we may also define ion conductivity as

$$\sigma = n_i q \mu \tag{16.2}$$

where n_i is the number of mobile ions per unit volume, q is the ionic charge (1.6×10^{-19} C for singly charged ions), and μ is the ion mobility. Interestingly, it is thought that n_i in high ion concentration solid electrolytes such as the Ag-Ge-S ternaries is fairly constant – around 10^{19} ions cm^{-3} [6]. This means that the ion mobilities in the solid electrolytes that are of interest to us are in the order of 10^{-2}–$10^{-4}\,cm^2\,V^{-1}\,s^{-1}$.

16.2.2 Major Inorganic Solid Electrolytes

As mentioned above, a variety of materials can act as solid electrolytes. Anion (oxide) conductors exist, such as ZrO_2, layered La_2NiO_4/La_2CuO_4 [7], or $Bi_{10}V_4MeO_{26}$, where Me is a divalent metal such as Co, Ni, Cu, or Zn [8]. However, for mass-transport devices there is a greater interest in electrolytes that conduct metallic cations as these can be used to form solid metal electrodeposits. In general, the smaller an ion is, the more mobile it should be as it will be able to slip more easily through the pathways in the solid electrolyte. This should be especially true for small-ionic radius elements such as the alkali metals (Li, Na, K). For example, Na^+ has been successfully used in beta-alumina and, to a lesser extent, in non-stoichiometric zirconophosphosilicate [9] to produce good solid ion conductors. The high conductivity in the beta-alumina compounds is a consequence of the structure which has open conduction pathways and a large number of partially occupied sites where cations can reside. Of course, Li^+ conductors in general are of great interest because of their use in high-voltage/high-power density lithium ion batteries, but highly stable Li^+ electrolytes are not easy to produce and there are not many examples of lithium solid electrolyte batteries ($Li/LiI/I_2$ is one of the few commercially available cells). Of course, the high chemical reactivity of these mobile elements makes them unsuitable for most mass transport/

solid electrolyte device applications, and so more stable alternatives must be considered.

The most widely studied solid ion conductors are those which contain silver. These tend to be less difficult to make than alkali metal ion electrolytes, and they have many desirable characteristics, including high ion mobility. Silver is especially appropriate for mass transport applications due to its nobility and ease of both reduction and oxidation. The crystalline Ag halides, principally AgI, and silver *chalcogenides* (e.g., Ag_2S, Ag_2Se, and Ag_2Te) are of particular interest as solid electrolytes. The phases of these materials that are stable at low temperature are semiconductors with moderate to low ion conductivity, but the high-temperature polymorphs (e.g., the cubic phase of Ag_2Se that is stable above 133 °C) are extremely good ion conductors [10]. The effect of this phase transition on the conductivity of AgI is shown in Figure 16.2. The Ag halides and chalcogenides possess a *bcc* structure formed by the covalently bonded halide or chalcogen atoms. An octahedral sublattice which can be occupied by Ag^+ in a multitude of ways is shown in Figure 16.3. The number of the octahedral states is typically much higher than the number of the available Ag ions, and this ensures that there are always non-occupied sites for ions to move into. This abundance of empty sites, in conjunction with the low potential barrier, results in the *superionic* nature of the materials. One other factor contributing to the high conductivity of these electrolytes is the low coordination that Ag^+ has to the immobile chalcogenide/halide sublattice (typically 2–3). This low coordination most likely plays a key role in reducing the activation energy for conduction.

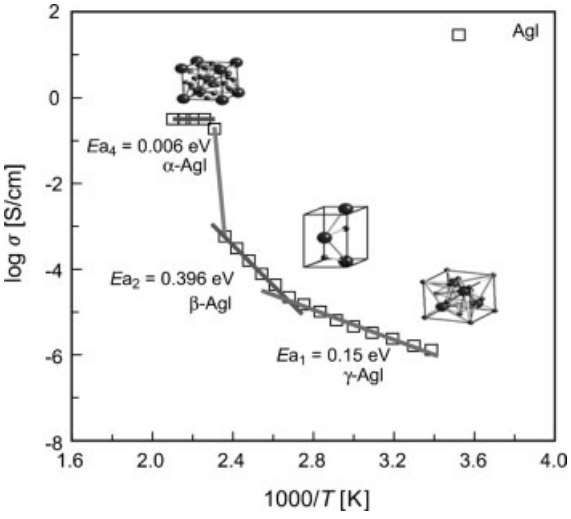


Figure 16.2 Temperature dependence of the conductivity of AgI related to its polymorph structure (Ref. [10]). The polymorph that is stable at high-temperature (denoted by α-AgI on the plot) has the highest conductivity and lowest activation energy compared with the lower temperature (β- and γ-AgI) polymorphs.

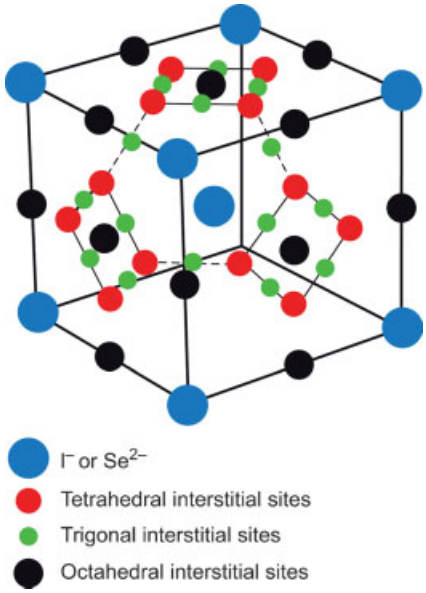


Figure 16.3 Schematic of bcc structure of Ag_2Se (or AgI) showing all possible Ag sublattices (from Ref. [10]). The large number of sites that can be occupied by cations such as Ag^+ lead to high ion mobility.

In practice, even deviations from stoichiometry, δ, as low as 1 part in 10 000 in some cases, lead to the existence of both mobile ions and conduction electrons. To illustrate this, in $\alpha Ag_{2+\delta}Se$, silver atoms are converted into silver ions and free electrons by

$$Ag_{2+\delta}Se \rightarrow (2+\delta)Ag^+ + \delta e^- + Se^{2-} \tag{16.3}$$

Both charge carriers (Ag^+ and e^-) contribute to the total conductivity, so that the material may be regarded as a *mixed conductor* [11]. Analogous effects are expected to occur in Cu-containing electrolytes (e.g., Cu_2S), although the situation will be more complex as participation of the electrons from the Cu d-orbital will result in a variety of bonding configurations. Extensive information on transport in superionic conductors is provided in a review [12]. The main issue with these binary materials is that, whereas they have been widely studied as superionic conductors, it is only their high-temperature phases that are of use in this respect, and this leads to severe practical limitations for electronic device applications.

16.2.3
Chalcogenide Glasses as Electrolytes

A chalcogenide glass is one that contains a large number of group VI or "chalcogen" atoms (S, Se, Te, and O, although oxide glasses are often treated separately from the others in the literature) [13]. These glasses have an astonishing range of physical

characteristics and as such have found a multitude of uses. They lend themselves to a variety of processing techniques, including physical vapor deposition (evaporation and sputtering), chemical vapor deposition, spin casting, as well as melt-quenching. Stable binary glasses typically involve a Group IVB or Group VB atom, such as Ge−Se or As−S, with a wide range of atomic ratios possible. The bandgap of the Group VIB glasses rises from around 1–3 eV for the tellurides, selenides and sulfides, to 5–10 eV for the oxides. The tellurides exhibit the most metallic character in their bonding, and are the "weakest" glasses as they can crystallize very readily (hence their use in so-called phase change technologies as used in re-writable CDs and DVDs), and the others exhibit an increasing glass transition temperature on moving further up the Periodic Table column, with oxides having the highest thermal stability. The non-oxide glasses usually are more rigid than organic polymers but more flexible than a typical oxide glass, and other physical properties follow the same trend. This structural flexibility of these materials offers the possibility of the formation of voids through which the ions can readily move from one equilibrium position to another and, as will be seen later, allows the formation of electrodeposits *within* the electrolyte.

The addition of Group IB elements such as Ag or Cu transforms the chalcogenide glass into an electrolyte as these metals form mobile cations within the material. The ions are associated with the non-bridging chalcogen atoms, but the bonds formed are relatively long −0.27 nm in Ag-Ge-Se and 0.25 nm in Ag-Ge-S ternaries [14]. As with any coulombic attraction, the coulombic energy is proportional to the inverse of the cation–anion distance, so long bonds lead to reduced attractive forces between the charged species. The Ge-chalcogenide glasses are therefore among the electrolytes with the lowest coulombic energies [14]. The slightly shorter $Ag^{+}-S^{-}$ bond length leads to a higher coulombic attraction, which is a factor contributing to the observed lower mobility of Ag in germanium sulfides versus selenides of the same stoichiometry. Thermal vibrations will allow partial dissociation, which results in a two-step process of defect formation followed by ion migration. The activation energy for this process depends heavily on the distance between the hopping cation and the anion located at the next nearest neighbor, as well as the height of the intervening barrier. (A discussion of the relationship between coulombic and activation energies is provided in Ref. [14] but, in addition to having low coulombic energies, the Ge-chalcogenides also have relatively low activation energies for ion transport.) In this respect, the existence of channels due to the structure of the electrolyte is critical in the ion transport process. As an example of this effect, the Ag^{+} conductivity in glassy $AgAsS_2$ is 100-fold larger than that in the crystalline counterpart due to the more "open" structure of the non-crystalline material [15].

The conductivity and activation energy for ion conduction of the ternary glasses is a strong function of the mobile ion concentration. For example, in the Ag concentration range between 0.01 and 3 atom%, the room temperature conductivity of Ag-Ge-S glass changes from 10^{-14} to about $5\times10^{-10}\,\Omega^{-1}\,cm^{-1}$, accompanied by a decline in activation energy from 0.9 to 0.65 eV. However, above a small atomic percent, both conductivity and activation energy change more rapidly as a function of Ag concentration (see Figures 16.4a and b, respectively [16]). This change in the slopes of the conductivity and activation energy curves with Ag content in both Ag-Ge-S and

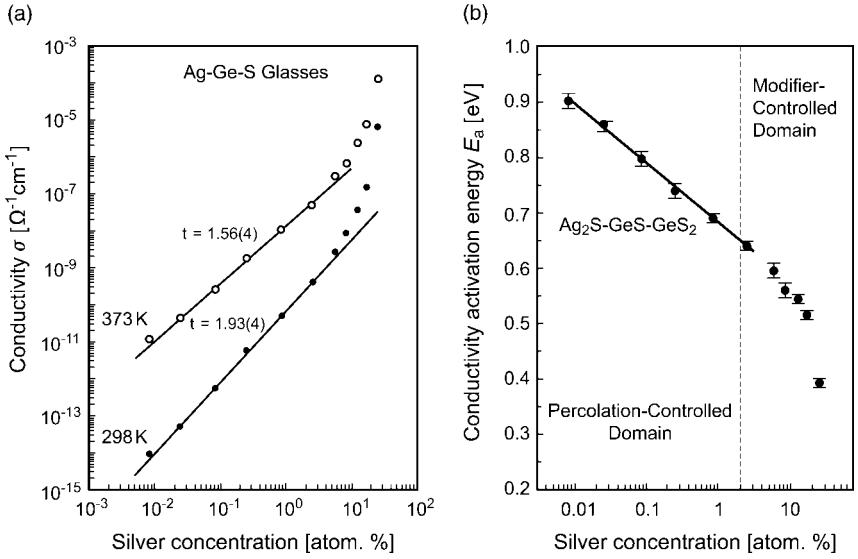


Figure 16.4 (a) Conductivity and (b) activation energy as a function of Ag concentration in Ag-Ge-S ternaries. (From Ref. [16].).

Ag-Ge-Se is a result of a transformation of the ternary material itself caused by the presence of so much silver.

16.2.4
The Nanostructure of Ternary Electrolytes

The transformation that occurs in ternary electrolytes at over a small atomic percent of metal is, by no means, subtle. Indeed, the material undergoes considerable changes in its nanostructure that have a profound effect on its macroscopic characteristics. These changes are a result of *phase separation* caused by the reaction of silver with the available chalcogen in the host to form Ag_2Se in Ag-Ge-Se and Ag_2S in Ag-Ge-S ternaries. For example, if it is assumed that the Ag has a mean coordination of 3, the composition of ternary Ag-Ge-Se glasses may be represented as

$$(Ge_xSe_{1-x})_{1-y}Ag_y = (3y/2)(Ag_2Se) + (1-3y/2)(Ge_tSe_{1-t}) \tag{16.4}$$

where t is the amount of Ge in the Ge-Se backbone $= x(1-y)/(1-3y/2)$ [17]. For a Se-rich glass such as $Ge_{0.30}Se_{0.70}$, $x = 0.30$, and $y = 0.333$ at saturation in bulk glass; hence, $t = 0.40$. This means that the material consists of Ag_2Se and $Ge_{0.40}Se_{0.60}$ (Ge_2Se_3) in the combination

$$16.7\ Ag_2Se + 10\ Ge_2Se_3 = Ag_{0.33}Ge_{0.20}Se_{0.47} \tag{16.5}$$

This electrolyte has a Ag_2Se molar fraction of 0.63 (16.7/26.7) and a Ag concentration of 33 atom% [18]. It has been determined that the dissolution of Ag into a

Se-rich base glass produces a ternary that is a combination of *separate* dispersed crystalline Ag_2Se and continuous glassy Ge-rich phases [19]. The spacing, s, between the Ag_2Se phase regions (and therefore the thickness of Ge_2Se_3 material between them) can be estimated by assuming that the crystalline regions are spherical and uniform in size and dispersion, so that

$$s = d_c(F_v^{-1/3} - 1) \tag{16.6}$$

where d_c is the average measured diameter of the crystalline Ag-rich phase and F_v is the volume fraction of this phase [20]. The volume fraction in the case of Ag_2Se in $Ag_{0.33}Ge_{0.20}Se_{0.47}$ is 0.57 (for a molar fraction of 0.63), so the average spacing between the Ag-rich regions is approximately 0.2 times their diameter. The average diameter of the Ag_2Se crystallites in Ag-diffused $Ge_{0.30}Se_{0.70}$ thin films was determined, using X-ray diffraction (XRD) techniques, to be 7.5 nm [20], which means that by Equation 16.6, they should be separated by approximately 1.5 nm of glassy Ge-rich material. This general structure has been confirmed using high-resolution transmission electron microscopy (TEM). XRD analysis was also performed on a sulfide-based ternary thin film with similar stoichiometry, $Ag_{0.31}Ge_{0.21}S_{0.48}$, with much the same results. In this case, the Ag_2S crystallites are in the order of 6.0 nm in diameter [21]. Even though the detected Ag-rich phases mainly correspond to the room-temperature polymorphs, which are not particularly good ion conductors at room temperature, the ternary is superionic at room temperature. This is not surprising as defects, interfaces, and surfaces play a considerable role in ion transport and the large surface-to-volume ratio of the crystallites within the ternary is likely to greatly enhance ion transport. In addition, it has been noted that the Ag_2Se phases that form following the solid-state diffusion of Ag into Ge-Se may be "distorted" by the effective pressure of the medium to produce high ion mobility phases [19]. The nano-inhomogeneous ternary is ideal for devices such as resistance change memory cells, as the relatively high resistivity leads to a high off-resistance in small diameter devices, although the availability of mobile ions via the dispersed Ag-rich phases means that the effective ion mobility is high.

The addition of Ag (or Cu) to the chalcogenide base glass can be achieved by diffusing the mobile metal from a thin surface film via *photodissolution*. This process utilizes light energy greater than the optical gap of the chalcogenide glass to create charged defects near the interface between the reacted and unreacted chalcogenide layers [22]. The holes created are trapped by the metal, while the electrons move into the chalcogenide film. The electric field formed by the negatively charged chalcogen atoms and positively charged metal ions is sufficient to allow the ions to overcome the energy barrier at the interface, and so the metal moves into the chalcogenide [23]. Prior to introduction of the metal, the glass consists of GeS_4 ($GeSe_4$) tetrahedra and, in the case of chalcogen-rich material, S (Se) chains. The introduced metal will readily react with the chain chalcogen and some of the tetrahedral material to form the ternary. This Ag–chalcogen reaction, which essentially nucleates on the chalcogen-rich regions within the base glass, results in the nanoscale phase-separated ternary described above [24, 25].

16.3
Electrochemistry and Mass Transport

16.3.1
Electrochemical Cells for Mass Transport

In order to move mass, it is clear that an ion current must be generated. Regardless of ion mobility in the electrolyte, a sustainable ion current will only flow if there is a source of ions at one point and a sink of ions at another; otherwise, the movement of ions away from their oppositely charged fixed partners would create an internal field (polarization) which would prevent current flow. The process of *electrodeposition*, in which cations in the electrolyte are reduced by electrons from a negative electrode (cathode), is essentially an ion sink as ions are removed from the electrolyte to become atoms. However, in the absence of an ion source, the reduction of ions at the cathode will occur at the expense of the electrolyte. The concentration of ions in the solid electrolyte will therefore decrease during electrodeposition until the electrode potential equals the applied potential and reduction will cease. Further reduction requires greater applied voltage (governed by the Nernst equation), so that the deposition process is effectively self-limiting for a moderate applied potential. It should be noted also that a depleted electrolyte could allow the subsequent thermal dissolution of an electrodeposit, which would not occur if the glass was maintained at the chemical saturation point. This has important consequences for the stability of any electrodeposit formed. It is therefore necessary to have an *oxidizable* positive electrode (anode) – one which can supply ions into the electrolyte to maintain ion concentration and overall charge neutrality. In the case of a silver ion-containing electrolyte, this oxidizable anode is merely silver or a compound or alloy containing free silver. So, the most basic mass-transport device consists of a solid electrolyte between an electron-supplying cathode and an oxidizable anode (see Figure 16.5). These devices can have both electrodes in a coplanar configuration (as in Figure 16.5a), or on opposite faces of the electrolyte (Figure 16.5b).

In such a device, the anode will oxidize when a bias is applied if the oxidation potential of the metal is greater than that of the solution. Under steady-state conditions, as current flows in the cell, the metal ions will be reduced at the cathode.

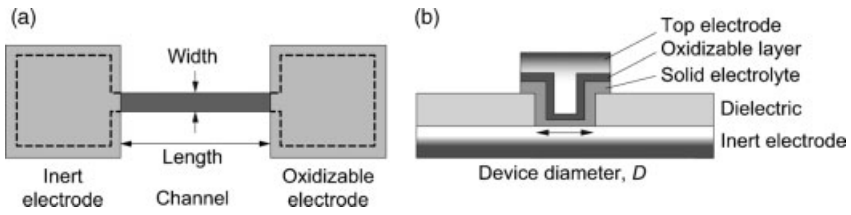


Figure 16.5 Schematic descriptions of two mass transport devices. (a) Coplanar structure that has the two electrodes on the same surface of a solid electrolyte layer. (b) Vertical structure with the solid electrolyte sandwiched between an inert electrode and an oxidizable layer which is covered by the top electrode.

For the case of silver, the reactions are:

$$\text{Anode}: \ \text{Ag} \rightarrow \text{Ag}^{+} + e^{-} \tag{16.7}$$

$$\text{Cathode}: \ \text{Ag}^{+} + e^{-} \rightarrow \text{Ag} \tag{16.8}$$

with the electrons being supplied by the external power source. The deposition of Ag metal at the cathode and partial dissolution of the Ag at the anode indicates that device operation is analogous to the reduction–oxidation electrolysis of metal from an aqueous solution and much the same rules apply, except that in this case the anions are fixed. When a bias is applied across the electrodes, silver ions migrate by the coordinated hopping mechanism (as described above) towards the cathode, under the driving force of the applied field and the concentration gradient. At the boundary layer between the electrolyte and the electrodes, a potential difference exists due to the transfer of charge and change of state associated with the electrode reactions. This potential difference leads to polarization in the region close to the phase boundary, known as the *double layer* [26]. The inner part of the double layer, consisting of ions absorbed on the electrode, is referred to as a *Helmholtz layer*, while the outer part, which extends into the electrolyte and is known to have a steep concentration gradient (over a few tens of nanometers in these systems), is called the *diffuse layer*. Electrically, the double layer very much resembles a charged capacitor, with a capacitance in the order of 10^{-14} F μm^{-2} and resistance around $10^{10}\ \Omega\ \mu m^{2}$ for a typical solid electrolyte under small applied bias [27]. An important consequence of the electric double layer is that, for the reduction–oxidation reaction to proceed, the applied potential must overcome the potential associated with the double layer. This means that no ion current will flow and no sustained electrodeposition will occur until the *concentration overpotential* is exceeded. Below this threshold voltage, the small observed steady-state current is essentially electron leakage by tunneling through the narrow double layer. Above the threshold, the ion current flows and the ions are reduced and join the cathode, effectively becoming part of its structure, both mechanically and electrically. The nature of the electrodeposits will be discussed in greater detail later in the chapter.

The intrinsic threshold is typically in the order of a few hundred millivolts. As the overpotential is governed by the ease of transfer of electrons from the cathode to the ions in the electrolyte, its precise value depends on factors such as the barrier height between the cathode material (including surface/interface states) and the electrolyte, and the bandgap/dielectric constant of the electrolyte. For example, the threshold voltage of a Ni/Ag-Ge-Se/Ag structure is in the order of 0.18 V, whereas the threshold of W/Ag-Ge-Se/Ag is around 0.25 V. Switching to a larger bandgap material, the threshold of a W/Ag-Ge-S/Ag device becomes closer to 0.45 V. The threshold has an Arrhenius dependence on temperature, with an activation energy that is in the order of a few tenths of an electron volt or less, which means that for the W/Ag-Ge-S/Ag structure, the threshold is still 0.25 V even at an operating temperature of 135 °C. Once a silver electrodeposit has formed on the cathode, the Ag metal becomes the new cathode and the threshold for further deposition of Ag is much less, typically less than half the original threshold. This reduced threshold for electrodeposition

following initiation has a profound effect on device operation, and is discussed in more detail below. Of course, long structures that have a high series resistance will require higher voltages to initiate electrodeposition, as most of the applied voltage will be dropped across the electrolyte. For example, the polarization-resistance of a 10 μm^2 electrode will be $10^9\,\Omega$, but if a 50 nm-thick 100 Ω cm Ag-Ge-Se electrolyte between anode and cathode is 10 μm wide and 100 μm long, then the series resistance will be twice this value and so at least 0.75 V will be needed to drop 0.25 V at the cathode and cause electrodeposition.

Just as in any "plating" operation, the ions nearest the electron-supplying cathode will be reduced first. However, in devices with closely spaced interfaces in which the nanoscale roughness of the electrodes is significant and the fields are relatively high, statistical non-uniformities in the ion concentration and in the electrode topography will tend to promote localized deposition or nano-nucleation rather than blanket plating. Even if multiple nuclei are formed, the one with the highest field and best ion supply will be favored for subsequent growth, extending out from the cathode as a single metallic feature. The nature of this somewhat one-dimensional growth will be discussed later in the chapter. The electrodeposition of metal on the cathode does not mean that ions entering from the oxidizable anode must travel the entire length of the structure to replace those that are reduced. As noted earlier, the ions move through the electrolyte by a *coordinated motion*; the ion closest to the reduced ion will move to the vacated negative site on the hosting material, and those upstream will do likewise, each filling the vacated site of the one downstream, until the last vacated space closest to the anode is filled by the incoming ion. So, in the initial stages, the electrodeposit is actually made up of reduced ions from the electrolyte itself; however, as each ion deposited on the growing electrodeposit corresponds to one that has been removed from the metal source, the net effect is a shift of mass from the anode toward the cathode. It should be noted that the growth process in these structures is more complex than a simple plating operation as the deposition interface is moving toward the source of the ions. As the electrodeposit is physically connected to the cathode, it can supply electrons for subsequent ion reduction; hence, the growing electrodeposit will harvest ions from the electrolyte, plating them onto its surface to extend itself outwards from the cathode. This has two consequences: (i) the growth interface continually moves out to meet the ions; and (ii) the growth closes the gap between the electrodes, thereby increasing the field. Both of these outcomes help to speed the overall growth rate of the deposit. The growth rate and electrodeposit morphology are discussed in the following section.

It should be noted that if the cathode is electrochemically inert (not oxidizable), then the electrodeposition process is reversible by switching the polarity of the applied bias. When the electrodeposit is made positive with respect to the original oxidizable electrode, it becomes the new anode and dissolves via oxidation. During dissolution of the electrodeposit, the balance is maintained by deposition of metal back onto the place where the excess metal for the electrodeposition originated. Once the electrodeposit has been completely dissolved, the process self-terminates. It is important to note that it is the *asymmetry* of the device structure that allows the deposition/dissolution process to be cycled repeatedly.

16.3.2
Electrodeposit Morphology

It is clear that the reduction of the ions results in the formation of neutral metal atoms. However, what is not so obvious is the form that the electrodeposits take, as the process depends on a number of factors and involves not only the basic principles of electrochemistry but also transport phenomena, surface science, and metallurgy [28–30]. In this section, some of the more important issues of electrodeposition and deposit morphology with the ternary solid electrolytes will be considered. This process is best illustrated in structures which support electrodeposit growth between coplanar electrodes on an electrolyte layer [31], although growth through an electrolyte film will also be considered.

In the most general case, the process of deposit formation starts with the nucleation of the new metal atom phase on the cathode, and the deposits develop with a structure that generally follows a Volmer–Weber 3-D island growth mechanism [32]. The addition of new atoms to the growing deposit occurs due to a *diffusion-limited aggregation* (DLA) mechanism [33, 34]. In this growth process, an immobile "seed" is fixed on a plane in which particles are randomly moving around. Those particles that move close enough to the seed in order to be attracted to it attach and form the aggregate. When the aggregate consists of multiple particles, growth proceeds outwards and with greater speed as the new deposits extend to capture more moving particles. Thus, the branches of the core clusters grow faster than the interior regions. The precise morphology of these elongated features depends on parameters such as the potential difference and the concentration of ions in the electrolyte [35]. At low ion concentrations and low fields, the deposition process is determined by the (non-directional) diffusion of metal ions in the electrolyte and the resulting pattern is fractal in nature; that is, it exhibits the same structure at all magnifications. For high ion concentrations and high fields, conditions common in the solid electrolyte devices, the moving ions have a stronger directional component, and *dendrite* formation occurs. Dendrites have a branched nature but tend to be more ordered than fractal aggregates and grow in a preferred axis that is largely defined by the electric field. An example of dendritic growth is shown in Figure 16.6 for a Ag electrodeposit on a Ag-saturated $Ge_{0.30}Se_{0.70}$ electrolyte between Ag electrodes. Figure 16.6a is an optical micrograph of the electrodeposit, showing its dendritic character; while Figure 16.6b is an electron micrograph of a similar deposit, showing the extreme roughness of its surface at the nanoscale.

The above model for electrodeposit evolution assumes a homogeneous electrolyte. However, since electrodeposit growth is obviously related to the presence of available Ag ions in the electrolyte surface, the content and consistency of the electrolyte will have a profound effect on electrodeposit morphology. In the case of electrolyte based on a $Ge_{0.30}Se_{0.70}$ glass, the growth of low (about 20 nm high) continuous dendritic deposits are observed on surface of the films (see Figure 16.7a). In the case of the Ge-rich glasses ($Ge_{0.40}Se_{0.60}$), the growth of isolated, tall (>100 nm) electrodeposits can be seen (Figure 16.7b) [36]. The $Ge_{0.30}Se_{0.70}$ material has the higher chalcogen content of the two, and therefore will possess greater and more uniform quantities of

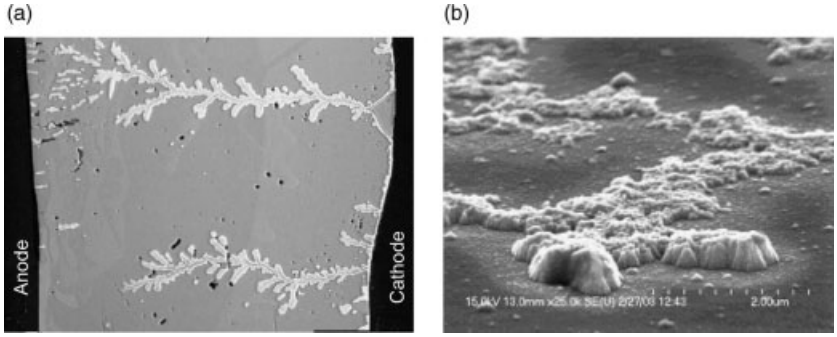
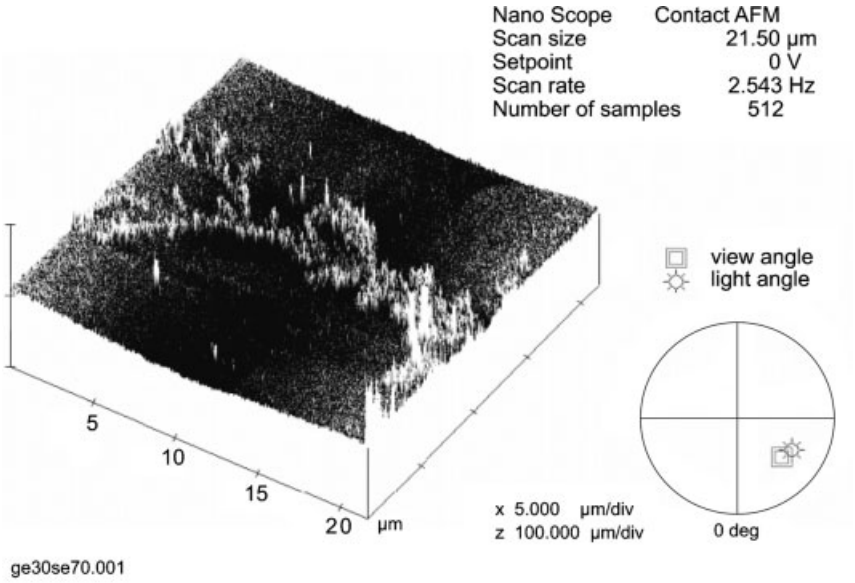


Figure 16.6 Surface electrodeposit of Ag on a Ge-Se-Ag solid electrolyte. (a) Optical micrograph of Ag dendrites. (b) Scanning electron microscopy image of the 3-D structure of the dendrites, showing the nano-roughness and large surface area. (From Ref. [31].).

ion-supplying Ag_2Se following the addition of Ag. This leads to dendritic growth that is closer to that expected with a homogeneous material. The isolated growth on the $Ge_{0.40}Se_{0.60}$ electrolyte is a direct consequence of the greater degree of separation of the Ag-containing phases.

The alternative device configuration has the electrodes on opposite sides of a thin electrolyte film, so that the growth of the electrodeposit is forced to occur *through* rather than *on* the electrolyte. Even though the capture and reduction of ions will essentially be by the same mechanism, it is unlikely that growth inside an electrolyte film will follow the same type of evolution as surface electrodeposition. At this point in time, although our understanding of the exact mechanism of growth within these electrolytes is incomplete, it is clear that the role of the nanoscale morphology should be considered. The confining nature of the medium, with its somewhat flexible channels and voids, will distort the shape of the electrodeposit, and its nano-inhomogeneity (as discussed above) will have a profound effect on local potential and ion supply. The net result is that the electrodeposit will not necessarily appear to be fractal or dendritic in nature, instead taking a form that is governed by the shape of the glassy voids and crystalline regions in the electrolyte. An example of such distorted morphology is shown in Figure 16.8; this is an electron micrograph of an electrodeposit within a 60 nm-thick $Ag_{0.33}Ge_{0.20}Se_{0.47}$ electrolyte between a tungsten bottom electrode and a silver top electrode. This was captured by overwriting a large (5×5 μm) device to produce multiple internal electrodeposits and then using a focused ion beam (FIB) system to ion mill a hole through the electrolyte [37]. The filament appears to be around 20 nm across, but this is misleading as the feature continues to grow through ion reduction by the electron beam.

As a final comment on morphology, reversing the bias dissolves the electrodeposit as it becomes the oxidizable element in the electrochemical cell. Macroscopically, this appears to be the reverse of the growth process, with the electrodeposit dissolving backwards from its tip (or tips in the case of a more two-dimensional dendrite). On closer inspection, the deposit actually dissolves near the tip region into a string of



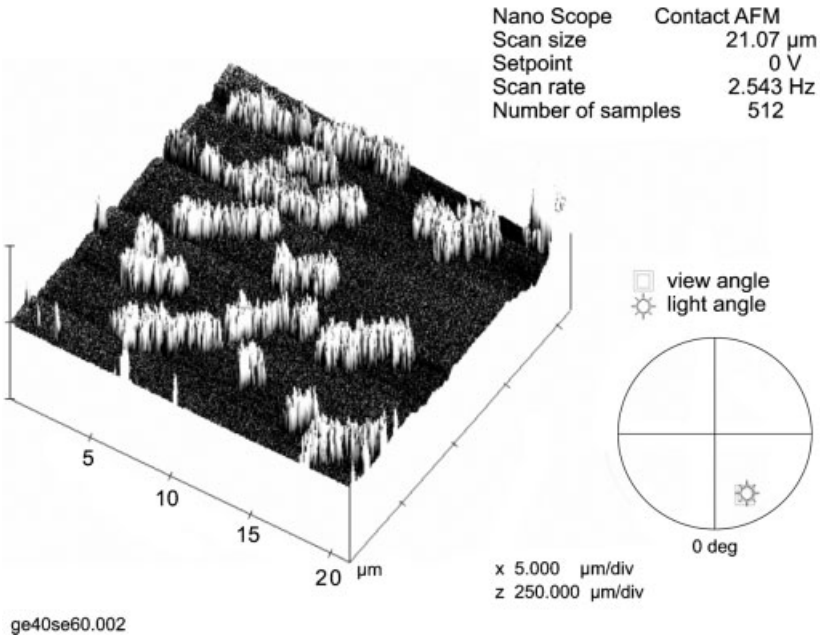


Figure 16.7 Atomic force microscopy image (3-D topographical scan) of (a) Ag electrodeposit grown on Ag-saturated $Ge_{0.30}Se_{0.70}$; the growth is continuous and the maximum electrodeposit height is a few tens of nanometers. (b) Ag electrodeposit grown on Ag-saturated $Ge_{0.40}Se_{0.60}$; the growth appears discontinuous and the maximum electrodeposit height is in the order of 100 nm. (From Ref. [36].).

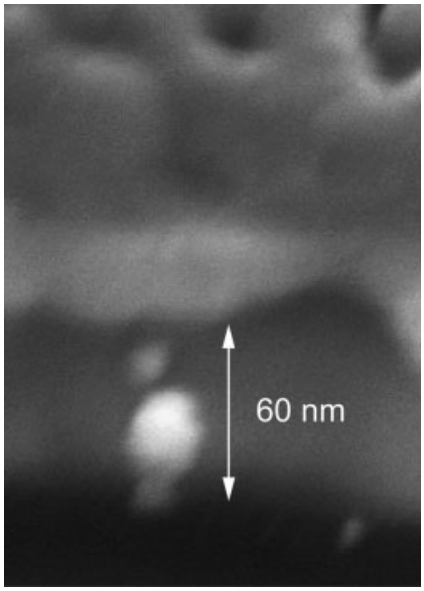


Figure 16.8 Electron micrograph of electrodeposit formation within a 60 nm-thick Ag-Ge-Se solid electrolyte observed by scanning electron microscopy following device sectioning by focused ion beam. (From Ref. [37].).

metal islands which then disappear into the electrolyte. This is a consequence of the uneven nature of the electrodeposit, created in part by the nano-morphology of the electrolyte, which allows some regions (perhaps associated with grain boundaries in the metal) to dissolve slightly faster than others.

16.3.3 Growth Rate

As in any deposition process, the growth rate of the electrodeposit, V, will depend on the ion flux per unit area, F, which corresponds to the current density, J, and the atomic density of the material being deposited, N, by

$$V = F/N = J/qN \tag{16.9}$$

The current density is given as

$$J = \sigma E \tag{16.10}$$

where E is the electric field. In large devices, the field will be relatively low, as will the current density. For example, in a 100 μm-long lateral (coplanar electrode) structure with 10 V applied, the field is $10^3\,\mathrm{V\,cm^{-1}}$. Ag-saturated ternary electrolytes such as Ag-Ge-Se have a conductivity around $10^{-2}\,\mathrm{S\,cm^{-1}}$, and so the ion current density for this field will be $10\,\mathrm{A\,cm^{-2}}$. Dividing current density by the charge on each ion (1.60×10^{-19} for Ag^+) gives the ion flux density, in this case 6.25×10^{19} ions $\mathrm{cm^{-2}\,s^{-1}}$.

Using Equation 16.9 above with the atomic density of Ag (5.86×10^{22} atoms cm^{-3}) gives a growth rate of approximately 10^{-3} cm s^{-1}, or 10 μm s^{-1}. This is a gross simplification, as the complex morphology of the electrodeposits and the moving boundary condition of the advancing electrodeposit will complicate the deposition process. However, this is the approximate average velocity that is measured in a real device for the above conditions.

The electrodeposit growth rates are much more difficult to model in devices that have a thin electrolyte layer sandwiched between two electrodes, and at this point any knowledge of the nano-morphology of the material must be invoked. The average fields in this case range from 10^5 to 10^6 V cm^{-1}, for applied voltages of a few hundred millivolts to a few volts across an electrolyte that is a few tens of nanometers thick. Local fields may be higher still, as most of the applied bias will be dropped across the high-resistance glassy areas between the lower resistivity, metal-rich nanoclusters. Taking a field of 10^6 V cm^{-1} with the conductivity given above suggests that the growth rate will be in the order of 1 cm s^{-1} or 10 nm μs^{-1}. This is much slower than the rate suggested by measured switching speeds observed in actual devices (as will be shown in the next section), so the simple approach that was appropriate for macroscale devices apparently fails at the nanoscale. This may be due to a number of factors. For example, at fields of 10^6 V cm^{-1} or more, the linear conduction equation no longer holds [38] and the mobility will be higher than in the macroscopic case. It is also likely that n_i is larger than the previously assumed 10^{19} cm^{-3}, as the overall silver concentration can be as high as 10^{22} cm^{-3} in Ag-saturated chalcogenide glass electrolytes, and more of this is likely to be mobile due to barrier lowering in materials subjected to such high fields. The overall effect is that the current densities could be sufficiently high to make the electrodeposit growth rates several orders of magnitude higher in nanoscale devices.

16.3.4
Charge, Mass, Volume, and Resistance

The atomic weight of Ag is 107.9 g mol^{-1}, and the metal has a bulk density of 10.5 g cm^{-3}. This means that each atom weighs approximately 1.79×10^{-22} g (107.9 divided by Avogadro's number, 6.022×10^{23} atoms per mol) and this, of course, is the smallest amount of mass that can be transferred using silver as the mobile ion. Each cm^3 of Ag contains 5.86×10^{22} atoms, but a more useful unit in these nanoscale systems in the nm^3; such a volume contains 58.6 atoms on average, and will weigh approximately 10^{-20} g. If this is the electrodeposited mass, then each Ag^+ ion requires one electron from the external circuit to become reduced to form the deposited atom. So, each nm^3 of Ag will require 58.6 times the charge on each electron (1.60×10^{-19} C) which is 9.37×10^{-18} C of Faradaic charge (we can also use Faraday's constant, 9.65×10^4 C mol^{-1}, to perform this calculation). This charge is merely the integral of the current over time, and so a constant current of 1 μA would supply sufficient charge in 1 μs to deposit 10^5 nm^3 or around 1 fg (10^{-15} g) of Ag. So, in this mass-transfer scheme, current and time are the control parameters for the amount of mass deposited.

The amount of charge supplied will also determine the volume of the electrodeposit. The increase in metal volume at a point on the surface of the electrolyte (or decrease in volume at the anode) could be useful in a variety of microelectromechanical applications but it is the electrical resistance of this volume that is perhaps of most interest. The resistance, R, of an electrodeposit is given by

$$R = L/\sigma_m A \tag{16.11}$$

where σ_m is the conductivity of the metal, and L and A are its length and cross-sectional area, respectively (volume is L × A). If the electrodeposited material is silver, σ_m will range from a value close to 5×10^5 S cm^{-1} (slightly higher than the bulk resistance) for features with thickness and width that are greater than a few tens of nanometers to much larger values for sub-10 nm features where surface scattering will play a considerable role. For a silver electrodeposit that is 20 μm long, 2 μm wide, and 20 nm thick (not unlike the example shown in Figure 16.7a), the resistance is about 10 Ω. This volume would take a total of 7.50×10^{-9} C to form. Note that if the underlying Ag-Ge-Se electrolyte (with conductivity 10^{-2} S cm^{-1}) was 20 μm long, 20 μm wide, and 50 nm thick, its resistance would be 2×10^7 Ω (or closer to 4×10^8 Ω for Ag-Ge-S), which is many orders of magnitude higher than that of the electrodeposit. Hence, the overall resistance of the newly formed structure is dominated by the electrodeposit. Figure 16.9 shows, graphically, the measured "on" state data from 50 nm-thick silver-doped arsenic disulfide electrolyte on a thick oxide layer on silicon substrates, patterned into channels with large silver contacts (100 × 100 μm) at the ends [39]. The "off" resistance, R_{off}, is a geometric function of the channel dimensions, following $R_{off} = L/\sigma dW + R_c$, where σ is the conductivity

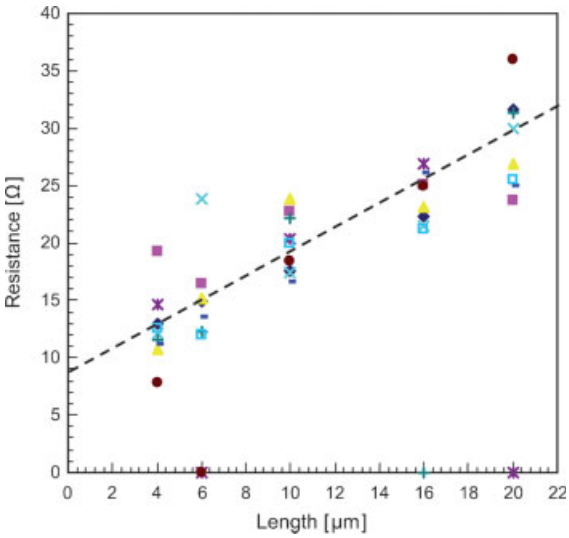


Figure 16.9 Resistance versus length for 10 μm-wide coplanar structures for a 25 mA current limit. The different symbols indicate results from different devices. (From Ref. [39].).

of the electrolyte layer (in the 10^{-3} S cm^{-1} range), d is its thickness, and W and L are width and length, respectively. R_c is the contact resistance (at zero channel length), mainly due to electrode polarization and tunneling at the measurement voltage through the polarization barrier. R_c is in the range of 10^8 to low 10^9 Ω for the electrode configuration used. A 10×10 μm ($W \times L$) device therefore exhibits an R_{off} around 1.5 GΩ. The figure shows the results from a number of 10 μm- wide devices for programming using a 5-s voltage sweep from 0.5 to 1.8 V with a 25 mA current limit. This produces a substantial surface electrodeposit with a resistance of around 1 Ω μm^{-1} of device length. The average electrodeposit contact resistance in this case is around 9 Ω.

The resistance of the electrodeposit that forms within the nanostructured electrolyte is also determined by its volume, but in this case the influence of the different phases present at the nanoscale must also be considered. As discussed above, electrodeposition in its early stages is likely to occur on the metal-rich clusters, through the glassy high resistance regions between them. This means that the initial connection through the electrolyte will essentially consist of metallic bridges between the relatively low-resistivity clusters. In the case of a link that is dominated by the conductivity of the clusters rather than that of the metal, an on-resistance in the order of 20 kΩ in a 50 nm-thick Ag-Ge-Se electrolyte would require a conducting region less than 10 nm in diameter (assuming that the conductivity of the Ag_2Se material is close to the bulk value of 10^3 S cm^{-1} [40]). In the case where the electrodeposit dominates the pathway – that is, when the electrodeposited metal volume is greater than that of the superionic crystallites in the pathway – the electrodeposit resistivity will determine the on-resistance. In this case, a 10 nm-diameter pathway will have a resistance in the order of 100 Ω. This means that the diameter of the conducting pathway will not exceed 10 nm for typical programming conditions which require on-state resistances in the order of a few kΩ to a few tens of kΩ. The small size of the conducting pathway in comparison to the device area explains why on-resistance has been observed to be independent of device diameter, whereas off-resistance increases with decreasing area [41]. An electrodeposit this small means that the entire device can be shrunk to nanoscale dimensions without compromising its operating characteristics. This has been demonstrated by the fabrication of nanoscale devices as small as 20 nm that behave much like their larger counterparts [20, 42]. The other benefit of forming a small-volume electrodeposit is that it takes little charge to do so; in an extreme case, if half the volume of a sub-10 nm-diameter, 50 nm-long conducting region was pure Ag, only a few fC of Faradaic charge would be required to form a low-resistance pathway.

As discussed above, reversing the applied bias reverses the electrodeposition process so that the electrodeposit itself becomes the oxidizable anode and is thereby dissolved. The amount of charge necessary to do this is essentially the same as that required to grow the link in the first place. However, it is not just the oxidation of the electrodeposited metal that is responsible for breaking of the link, especially in the case of a connection between the electrodes that are mostly metal rather than a chain of metallic and Ag-rich electrolytic clusters. The very narrow (and uneven) link is

susceptible to being weakened by electromigration by the current flowing through the metal. This means that the time to failure (initial breaking) of a metallic link depends partly on the current density; it can take years to break the link at a reverse current several orders of magnitude below the critical current density (around 10^7–10^8 A cm^{-2}), but the same link can be broken in less than 1 μs for a reverse current density in excess of this. In practice, for vertical structures, the critical current is typically less than the current limit used to form the link, but not less than 20% of this current. It should be noted that a forward current cannot easily be used to break the link by electromigration if the applied bias is above the threshold for electrodeposition, as any weakness (high-resistance region) in the electrodeposit will be "healed" by electrodeposition in the area due to the elevated voltage drop there. As the resistivity of the electrolyte is many orders of magnitude higher than that of the electrodeposited metal, the overall resistance of the structure rises dramatically as soon as the link is broken. The remainder of the electrodeposited metal in the now-incomplete link is dissolved electrochemically.

16.4 Memory Devices

16.4.1 Device Layout and Operation

As shown in Figure 16.5, the basic elements of a resistance change device – the solid electrolyte, the oxidizable electrode, and the inert electrode – may be configured either laterally or vertically. Whereas lateral devices may have utility in a variety of applications (e.g., microelectromechanical systems; MEMS), it is the vertical configuration that is of most interest in the context of memory devices. Vertical structures occupy the smallest possible area, which is critical for high-density memory arrays. In addition, the distance that the electrodeposit must bridge in order to switch a vertical device to its low-resistance state – a key factor in determining switching speed – is defined by the electrolyte thickness rather than by a lithographically defined gap. As the film thickness can typically be made much smaller than a lateral gap using conventional manufacturing technology, vertical structures switch faster than their lateral counterparts.

A schematic representation of how vertical solid electrolyte memory devices may be integrated in a complementary metal oxide-semiconductor (CMOS) circuit is shown in Figure 16.10. In this case, the inert electrodes are the tungsten plugs that are normally used to connect one layer of interconnect metal to another. The solid electrolyte layer is placed on top of these individual tungsten electrodes, and a common oxidizable electrode (or a bilayer of oxidizable metal and another electrode material) caps the device structures. The individual devices are defined by each tungsten plug. It should be noted that the storage elements are built in the interconnect layers above the silicon devices in a "back-end-of-line" (BEOL) process, which means that the CMOS fabrication scheme need not be changed. A further

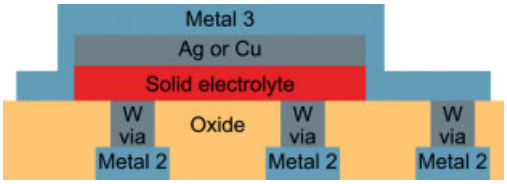


Figure 16.10 Schematic illustration of solid electrolyte device integration between two levels of metal (in this case "metal 2" and "metal 3" in a standard CMOS process. The individual devices are defined by each *W* via plug under the continuous electrolyte layer. One extra mask step is used to define which tungsten via plugs will be covered with the solid electrolyte and oxidizable metal, and which will be through connections.

advantage is that only one extra mask is required to define which tungsten plugs are covered with the device stack, and which are through-connections to the upper layers of the interconnect. This helps to reduce the cost of integration and also facilitates embedding the memory with standard logic. In order to obtain the maximum performance from the devices, each storage cell is connected through the underlying interconnect to a "select" transistor in a "one transistor-one resistor" (1T1R) cell array (Figure 16.11a). In this scheme, the transistor is used to select the cell, and an appropriate programming voltage is then applied across the device. Passive arrays, in which sneak current paths through the cells are avoided using diode elements in the array itself rather than transistors, are also possible using row and column electrodes with device structures at their intersections (Figure 16.11b). This latter approach does not allow high-speed operation but does lead to the densest array possible as there are no transistors to enlarge the total cell area.

The programming of the solid electrolyte memory devices is relatively straightforward. A forward bias (oxidizable electrode positive with respect to the inert electrode) in excess of the threshold required to initiate electrodeposition is used to write the

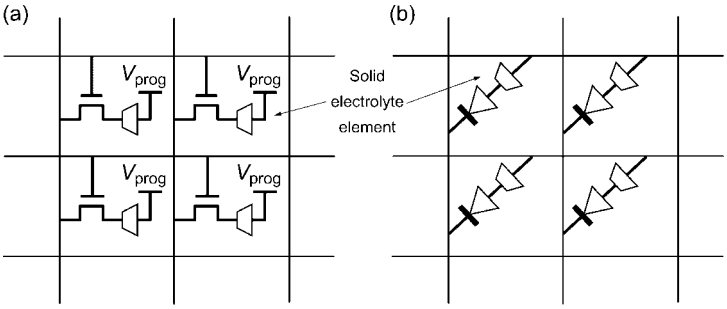


Figure 16.11 (a) Active "1T1R" array configuration. Each memory location consists of a solid electrolyte device and a transistor. V_{prog} is used to set the appropriate voltage across the device for programming. (b) Passive array configuration. Each cell has a diode to prevent "sneak paths" between rows and columns through the cells.

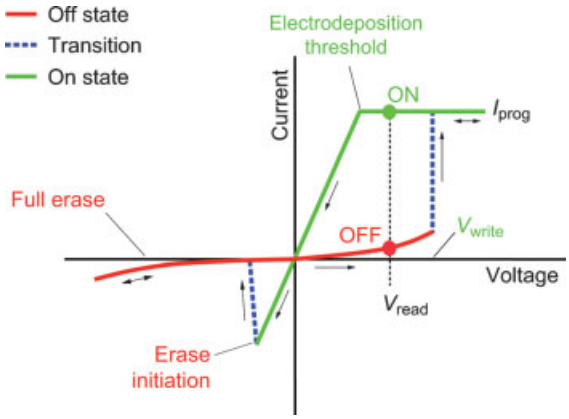


Figure 16.12 Schematic of a current–voltage plot of a solid electrolyte memory device programmed with current limit I_{prog}. The conducting pathway forms at the write voltage (V_{write}) and breaks at the erase initiation voltage/current. Further negative bias is required to fully remove the electrodeposited material and return the device to its original off state. The state of the device is read using a positive voltage below V_{write} to avoid disturbing the state.

device. A negative bias is used to erase the device. Reading the state of the device involves the application of a bias that will not "disturb" or destroy the current state. This typically means that the devices are read using a forward bias that is below the minimum required to write under normal operating conditions. This is shown schematically in Figure 16.12, which shows a current–voltage plot of a solid electrolyte memory device. Only leakage current flows in the off state, but when the conducting pathway forms at the write voltage (V_{write}), the current quickly rises to the programming current limit (I_{prog}). It should be noted that the electrodeposition continues after switching, albeit more slowly than the initial transition, until the voltage across the device reaches the minimum threshold for electrodeposition. The lower on-state resistance is preserved until the erase initiation voltage is reached, at which point the conducting pathway breaks and the device resistance goes high. Further negative bias is required to fully remove the electrodeposited material and return the device to its original off state. The device is read using a positive voltage below V_{write} and the current measured to determine the state. Note that in Figure 16.12, V_{read} has been chosen to be between the minimum voltage for electrodeposition and V_{write} (the consequences of this are discussed in Section 16.4.3). The following section provides results from a variety of fabricated devices.

16.4.2 Device Examples

To date, electrochemical switches have been fabricated using thin Cu_2S [43] and Ag_2S [44] binary chalcogenide electrolytes. The Cu_2S devices have been demonstrated in small memory arrays [45] and reconfigurable logic [46], and although the applications show promise, there is room for improvement in device performance

factors such as retention and endurance with this particular electrolyte. The studies on Ag_2S devices have concentrated on switching by the deposition and removal of small numbers of silver atoms in a nanoscale gap between electrodes. This is of major significance as it demonstrates that the electrochemical switching technique has the potential to be scaled to single atom dimensions. Various oxide-based devices have also been demonstrated [47, 48], and these show great promise as easily integrated elements. However, the lower ion mobility in these materials tends to make the devices slower than their chalcogenide counterparts. Devices based on ternary chalcogenide electrolytes, including Ag-Ge-Se, Ag-Ge-S, and Cu-Ge-S have been the most successful to date, with the silver-doped variants having been applied in sophisticated high-density memory arrays [49] and post-CMOS logic devices [50]. Ag-Ge-Te devices have also been explored [51], but these materials have a tendency to crystallize at low temperatures and so may not be the best choice for devices that must be integrated with CMOS using elevated processing temperatures.

To illustrate the operation of devices based on ternary electrolytes, the discussion will be confined here to those utilizing Ag-Ge-S and Ag-Ge-Se materials. Of these, the Ag-Ge-S electrolyte is the most compatible with BEOL processing in CMOS fabrication as it can withstand thermal steps in excess of 400 °C without any degradation of device characteristics. The sulfides possess better thermal stability as there is less change in the nanostructure at elevated temperature than in the case of selenide electrolytes [21, 52]. A typical device operation is shown in Figure 16.13a and b, which provide current–voltage and resistance–voltage plots respectively for a 240 nm-diameter W/Ag-Ge-S/Ag device with a 60 nm-thick electrolyte [53]. The voltage sweep runs from −1.0 to +1.0 to −1.0 V, and the current limit is 10 μA. As mentioned above, the write threshold for this material combination is 0.45 V, at which voltage the device

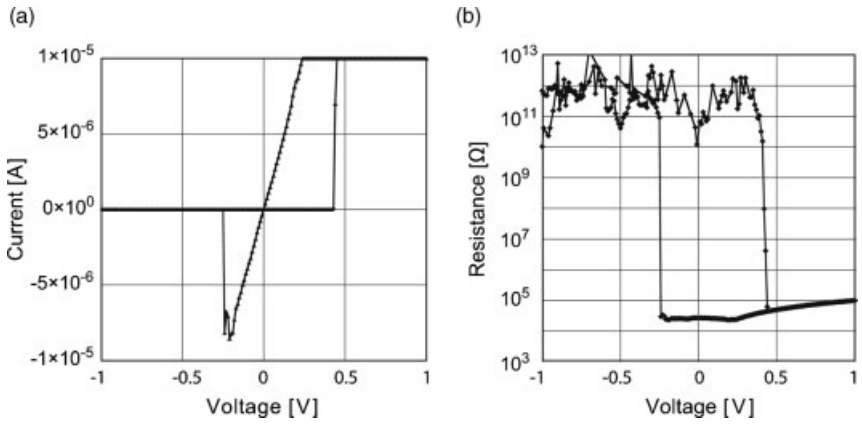


Figure 16.13 (a) Current–voltage plot of a 240 nm-diameter device with a 60 nm-thick Ag-Ge-S electrolyte using a 10 μA current limit. The device has been annealed at 300 °C. The voltage sweep is −1.0 to +1.0 to −1.0 V. (b) Resistance–voltage plot of the same device. The voltage sweep is −1.0 to +1.0 to −1.0 V. (From Ref. [52].).

switches from an off-state resistance, R_{off}, above $10^{11}\,\Omega$ to an on-state resistance, $R_{on} = 22\,k\Omega$, more than six orders of magnitude lower for the 10 μA programming current. This apparent rise in resistance following switching is caused by the current limit control in the measurement instrument. Once electrodeposition is initiated, the threshold for further electrodeposition is decreased, as indicated by the presence of a lower voltage (0.22 V) at which the current drops below the current limit on the negative-going sweep. R_{on} is determined by this voltage divided by the current limit (see below). The device transitions to a high-resistance state at −0.25 V, this being due to an initial breaking of the electrodeposited pathway by the reverse current flow. Continuing the negative sweep, the off-resistance remains above $10^{11}\,\Omega$ as the voltage is swept out to −1.0 V with a leakage current of less than 10 pA at maximum reverse bias. When considering the above characteristics, the device may be written using a voltage in excess of 0.45 V, read by applying a positive voltage that is less than 0.45 V, and erased by a bias greater than −0.25 V. These voltages are compatible with devices at the 22-nm node of the International Technology Roadmap for Semiconductors (ITRS).

Figure 16.14 illustrates the dependence of R_{on} on the programming current limit, I_{prog}, in the range 1–10 μA for a W/Ag-Ge-Se/Ag device with a 50 nm-thick electrolyte [54]. For this electrolyte/electrode combination, the write threshold is 240 mV and the electrodeposition threshold 140 mV. R_{on} is related to I_{prog} by $R_{on} = 0.14/I_{prog}$ (the solid line in the figure). This relationship between electrodeposition voltage, programming current, and on-resistance is common to all material combinations. It is explained by the fact that as long as sufficient potential difference is maintained for the situation where electrodeposition is already underway (in this case 140 mV), the reduction of silver ions will continue and the decrease in resistance of the conducting bridge will be maintained, even after it has formed. If the external current source is limited, when the resistance falls to a point where the voltage drop is no longer sufficient to support electrodeposition, the resistance lowering process ceases. The resulting resistance in ohms is therefore given approximately by the minimum potential to sustain electrodeposition in volts divided by the current limit of the external supply in amperes.

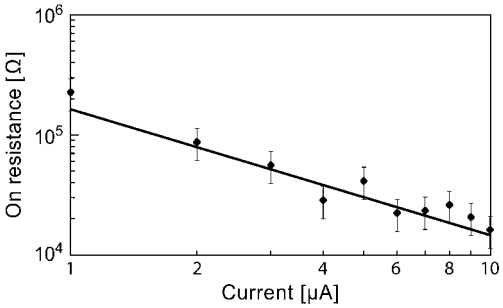


Figure 16.14 On-state resistance versus programming current for a 1 μm-diameter W/Ag-Ge-Se/Ag device. The resistance value (in Ω) is approximately 0.14 divided by the current in A (solid line). (From Ref. [54].).

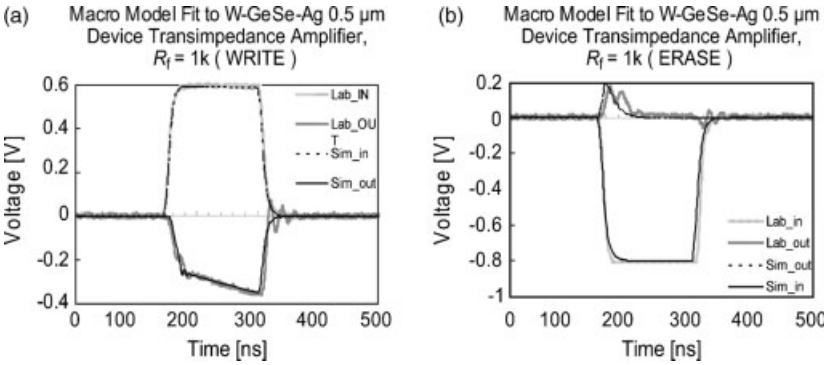


Figure 16.15 (a) Result of a 150-ns write pulse of 600 mV applied to a 500 nm W/Ag-Ge-Se/Ag device showing the output of the transimpedance measurement amplifier. The final on-resistance is 1.7 kΩ. (b) Result for a 150-ns erase pulse of −800 mV on the same device, showing the output of the transimpedance measurement amplifier. Lab_in and Lab_out are the measured input and output signals; Sim_in and Sim_out are the corresponding model generated curves. (From Ref. [55].).

The switching speed of a 500 nm-diameter W/Ag-Ge-Se/Ag device with a 50 nm-thick electrolyte is illustrated in Figure 16.15, which shows both measured and simulated device results for write (Figure 16.15a) and erase (Figure 16.15b) operations [55]. For the write, a 150-ns pulse of 600 mV was applied to the device, and the output of the transimpedance measurement amplifier shows that the device initially switches in less than 20 ns, while the resistance continues to fall more slowly, ultimately reaching an on-resistance of 1.7 kΩ at the end of the write pulse. For the erase, a 150-ns pulse of −800 mV was applied, whereupon the output of the transimpedance measurement amplifier shows that the device transitions to a high-resistance state (the start of the voltage decay in the output signal) in around 20 ns. The electrodeposit is essentially metal that has been added to a chemically saturated electrolyte, and this local supersaturation leads to high stability of the electrodeposit and excellent device retention characteristics. The results of a retention assessment test on a 2.5 µm-diameter W/Ag-Ge-S/Ag device with a 60 nm-thick electrolyte annealed at 300 °C, and programmed using a 0 to +1.0 V sweep is shown in Figure 16.16. The plot shows the off and on resistances measured using a 200 mV read voltage. The off state was in excess of 10^{11} Ω (above the limit of the measurement instrument) and remained undisturbed by the read voltage at this level for the duration of the test. The on resistance remained below 30 kΩ during the test. Following this, the device was erased using a 0 to −1.0 V sweep, and the off-state resistance measured using a 200 mV sensing voltage as before. The device remained above 10^{10} Ω beyond 10^5 s, demonstrating that the erased state was also stable. Other studies have show that both on and off states are also stable at elevated temperature, with a margin of several orders of magnitude being maintained even after 10 years at 70 °C [42]. Figure 16.17 provides an example of cycling for a 75-nm Ag-Ge-Se electrolyte device [20]. Trains of positive (write) pulses of 1.2 V in magnitude and 1.6 µs duration followed by −1.3 V negative (erase) pulses of 8.7 µs duration were

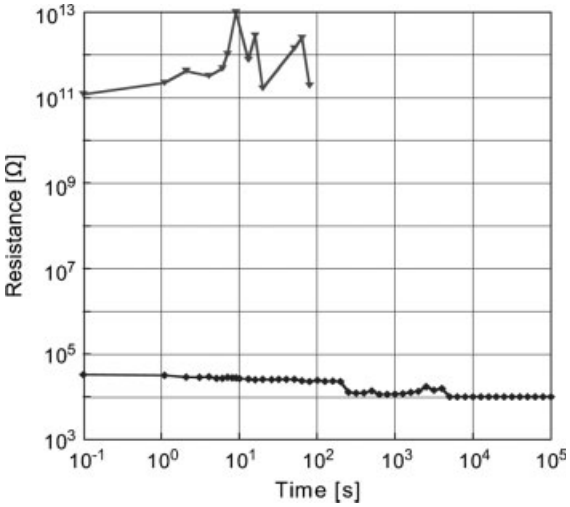


Figure 16.16 Off- (upper plot) and on-state (lower plot) resistance versus time measured using a 200 mV read voltage for a 2.5 μm W/Ag-Ge-S/Ag device programmed using a 10 μA current limit. The off-state resistance remained above $10^{11}\,\Omega$ for the duration of the test. (From Ref. [53].).

used to cycle the devices. A 10 kΩ series resistor was used to limit current flow in the on state. The results are shown in 10^9 and 10^{11} cycle ranges. The data in Figure 16.17 show that there might be a slight decrease in on current, but this is gradual enough to allow the devices to be taken well beyond 10^{11} write–erase cycles (if this decrease is maintained, there will only be a 20% decrease in on current at 10^{16} cycles).

As mentioned above in this section, oxide-based electrolytes may also be used in memory devices. Of these, Cu-WO_3 [47] and Cu-SiO_2 [48] are of particular interest as they utilize materials that are already in common use in the semiconductor industry,

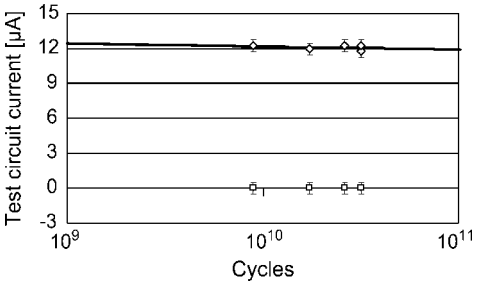


Figure 16.17 Current in the on (upper plot) and off (lower plot) state at various numbers of cycles for a 75-nm Ag-Ge-Se device. The device was cycled using trains of positive (write) pulses of 1.2 V in magnitude and 1.6 μs duration, followed by −1.3 V negative (erase) pulses of 8.7 μs duration. The solid line is a logarithmic fit to the on current data. (From Ref. [20].).

namely Cu and W for metallization and SiO_2 as a dielectric, and this will help to reduce the costs of integration. In general, the switching characteristics for both systems are very similar to those observed in metal-doped chalcogenide glasses, and that is why the same switching mechanism is assumed for the oxide-based cells, even though the material nanostructure is quite different from that found in the ternary chalcogenide electrolytes. For example, in the case of Cu-WO_3, the Cu must exist within oxide in unbound form for successful device operation [47]. For Cu-SiO_2, the best results are attained via the use of porous oxide, formed by physical vapor deposition, into which the metallic copper is introduced by thermal diffusion so that it exists in "free" form in the nano-voids in the base glass. In the case of W-(Cu/SiO_2)-Cu devices with a 12 nm-thick electrolyte, both unipolar (positive voltage for both write and erase) as well as bipolar switching has been observed [48]. Unipolar switching requires high programming currents (several hundred μA to several mA) to thermally break the electrodeposited copper connection in forward bias. Bipolar switching with a resistance ratio of 10^3 is achieved with switching voltages below 1 V and currents down to the sub-μA range. Highly stable retention characteristics beyond 10^5 s and switching speeds in the microsecond regime have been demonstrated, and the possibility of multi-bit storage exists due to the relationship between on-state resistance and programming current. These results, combined with the initial endurance testing which showed that more than 10^7 cycles were possible with these structures, indicate that this technology shows promise as a low-cost, low-energy Flash memory replacement technology.

16.4.3
Technological Challenges and Future Directions

The above results indicate that memory devices based on electrodeposition in solid electrolytes show great promise. However, although several substantial development efforts are under way, many questions remain unanswered with regards to the physics and long-term operation of this technology. The most pressing issues relate to the reliability of such devices. In any memory technology, the storage array is only as good as its weakest cell. Reduced endurance (cycling between written and erased states), poor retention, and "stuck" bits plague even the most mature memory technologies. It may be many years before the issues concerning the solid electrolyte approach are fully understood, but considerable optimism exists regarding reliability which may set this technology apart from others. For example, many technologies suffer from reduced endurance due to changes in the material system with time. In this respect, solid electrolyte devices can exhibit diminishing off-to-on-resistance ratio with cycle number if incorrect programming (overwriting and/or incomplete erase) leads to a build up of electrodeposited metal within the device structure. The convergence of the off and on states eventually leads to an inability to discriminate between them. However, it is possible electrically to "reset" the solid electrolyte using an extended or "hard" erase; this will then plate the excess material back on to the oxidizable electrode and return the electrolyte to its original composition. This ability to change material properties using electrical signals allows such corrections to be

performed in the field, and this may have a profound effect on device reliability. Another issue that can occur in written devices is the upward drift in programmed on-state resistance with time at elevated temperature. This is thought to be due to thermal diffusion of the electrodeposited metal, but it may also be a consequence of electromigration during repeated read operations. However, a read voltage that lies between the write voltage and the minimum voltage for electrodeposition will essentially repair or "refresh" a high-resistance/open on state. To illustrate this, the device characteristics shown in Figure 16.13 are revisited. If a read voltage between 0.22 and 0.45 V is used, an off/erased device will not be written, but a device that has been previously programmed will actually have its on state strengthened. This "auto-refresh" above the minimum threshold for electrodeposition is unique to electrochemical devices. It should be noted that although this effect is extremely useful, it can also lead to problems in incorrectly erased devices (those which are open circuit but still have electrodeposited material on the cathode), as these can also be written at read voltages. Clearly, under-erasing must be avoided in order to maintain high device reliability.

Attention is now turned to the scaling of solid electrolyte memory devices. This involves two points of consideration: physical scaling; and electrical scaling. Physical scaling of the types of device described in the previous section has already been demonstrated to below 22 nm, with good operational characteristics [42]. In addition, studies on the bridging of nanometer-sized gaps between a solid electrolyte and a top electrode seem to suggest that *atomic-scale* electrodeposits could be used to change the resistance of the device, and this may represent the ultimate scaling of the technology [44]. What is not known is how the "high-performance" phase-separated chalcogenide electrolytes will scale, as these contain crystallites that approach 10 nm in diameter. Clearly, further investigations are required in this area, although some are already under way. The other aspect of scaling is electrical scalability. For example, the supply voltage for highly scaled systems around the 22 nm node of the ITRS is on the order of 0.4–0.6 V. This means that, in order to avoid the use of area-, speed-, and energy-sapping charge pumps, the memory cells must be able to operate at the very low voltages at which solid electrolyte devices can function. In addition, the critical current density for 22 nm interconnect is only a few tens of μA, and the devices must also be able to operate at these current levels which, once again, is achievable by solid electrolyte devices.

The final consideration for the future relates to memory density in the Tb (10^{12} bits)/chip regime. Such high storage densities will eventually be required for high-end consumer and business electronics to replace mechanical hard drives in small-form factor, portable systems. If it is assumed that a $20 \times 20\,\mathrm{mm}^2$ chip has an extremely compact periphery such that most of the area is storage array, and a compact cell at $4\,F^2$ (where F is the half-pitch), then Tb storage would require F to be 10 nm at most. Such small wires cannot be produced using standard semiconductor fabrication technologies without significant variations, and their current carrying capacity is very small. Backing-off to $F = 22\,\mathrm{nm}$ means that multi-level cell (MLC) storage – more than one bit per physical storage cell – will be necessary to achieve Tb storage. The ability in solid electrolyte devices to control the on-resistance using the programming current allows multiple resistance levels to be stored in each cell. For

example, four discrete resistance levels leads to 2 bits of information in each cell (00, 01, 10, 11). Such MLC storage has already been demonstrated in a solid electrolyte memory array that was integrated with CMOS circuitry [56]. A combination of the above characteristics demonstrated physical scalability with low-voltage/-current/-power and MLC operation, and it would appear that solid electrolyte memory devices are a strong contender for future solid-state memory and storage.

16.5 Conclusions

Considering the characteristics described in the previous sections, devices based on mass transport in solid electrolytes appear to be appropriate for use as scalable, low-power memory elements. A reduction in resistance of several orders of magnitude is attainable in vertical devices for a write power below 1 μW, and since the on-resistance is a function of programming current, then MLC operation with simple sensing schemes is possible. The elements are non-volatile, with extrapolated retention results suggesting that the reduced resistance, with a large off-to-on ratio, will persist for well over 10 years. Even sub-100 nm devices show excellent endurance with no significant degradation to over 10^{10} cycles, and with stable operation indicated well beyond this. Both, simulated and measured data show that the devices write and erase within 20 ns, and further scaling – especially in the vertical dimension – is likely to result in even greater programming speeds. It should be noted that write times in the order of a few tens of nanoseconds mean that the write energy is less than 100 fJ, which makes solid electrolyte cells memory one of the lowest energy non-volatile technologies. The low resistivity and small size of the electrodeposits mean that the entire device can be shrunk to nanoscale dimensions, without compromising operating characteristics. This physical scalability, combined with low-voltage and low-current operation, suggests that extremely high storage densities will be possible. The other benefit of forming a small-volume electrodeposit is that it takes little charge to do so – in the order of a few fC to create an extremely stable low-resistance link. The charge required to switch a solid electrolyte element to a non-volatile state is therefore comparable to that required to program a typical dynamic random access memory cell, with the potential to further reduce this charge in future devices.

References

1 Maier, J. (2005) *Nature Materials*, **4**, 805.
2 Faraday, M. (1838) *Philosophical transactions of the Royal Society of London*,
3 Kummer, J.T. and Weber, N. (1966) US Patent 3,458,356.
4 Kirby, P.L. (1950) *British Journal of Applied Physics*, **1**, 193.
5 Bychkov, E., Tsegelnik, V., Vlasov, Yu., Pradel, A. and Ribes, M. (1996) *Journal of Non-Crystalline Solids*, **208**, 1.
6 Miyamoto, Y., Itoh, M. and Tanaka, K. (1994) *Solid State Communications*, **92**, 895.
7 Goodenough, J.B. (2003) *Annual Review of Materials Research*, **33**, 91.

8 Lazure, S., Vernochet, Ch., Vannier, R.N., Nowogrocki, G. and Mairesse, G. (1996) *Solid State Ionics*, **90**, 117.

9 Kreuer, K.-D., Kohler, H. and Maier, J. (1989) *High Conductivity Solid Ionic Conductors* (ed. T. Takahashi), World Scientific, Singapore, p. 242.

10 Kartini, E., Kennedy, S.J., Sakuma, T., Itoh, K., Fukunaga, T., Collins, M.F., Kamiyama, T., Suminta, S., Sugiharto, A., Musyafaah, E. and Bawono, P. (2002) *Journal of Non-Crystalline Solids*, **312–314**, 628.

11 Ogawa, H. and Kobayashi, M. (2002) *Solid State Ionics*, **148**, 211.

12 Shahi, K. (1977) *Physica Status SolidI A-Applied Research*, **41**, 11.

13 Tanaka, K. (2000) Chalcogenide glasses, in *Encyclopedia of Materials: Science and Technology*, Elsevier.

14 Tanaka, K., Miamoto, Y., Itoh, M. and Bychkov, E. (1999) *Physica Status SolidI A-Applied Research*, **173**, 317.

15 Elliott, S.R. (1991) Chalcogenide glasses, in *Materials Science and Technology* (ed. J. Zarzycki), VCH, New York.

16 Bychkov, E. (2000) *Solid State Ionics*, **1111**, 136–137.

17 Mitkova, M., Wang, Y. and Boolchand, P. (1999) *Physical Review Letters*, **83**, 3848.

18 Mitkova, M. and Kozicki, M.N. (2002) *Journal of Non-Crystalline Solids*, **1023**, 299–302.

19 Kozicki, M.N., Mitkova, M., Zhu, J. and Park, M. (2002) *Microelectronic Engineering*, **63**, 155.

20 Kozicki, M.N., Park, M. and Mitkova, M. (2005) *IEEE Transactions on Nanotechnology*, **4**, 331.

21 Balakrishnan, M., Kozicki, M.N., Poweleit, C.D., Bhagat, S., Alford, T.L. and Mitkova, M. (2007) *Journal of Non-Crystalline Solids*, **353**, 1454–1459.

22 Rennie, J.H.S. and Elliott, S.R. (1987) *Journal of Non-Crystalline Solids*, **1239**, 97–98.

23 Kolobov, A.V., Elliott, S.R. and Taguirdzhanov, M.A. (1990) *Philosophical Magazine B*, **61**, 857.

24 Mitkova, M., Kozicki, M.N., Kim, H. and Alford, T. (2004) *Journal of Non-Crystalline Solids*, 338–340. 552.

25 Mitkova, M., Kozicki, M.N., Kim, H.C. and Alford, T.L. (2004) *Thin Solid Films*, **449**, 248.

26 Kotzeniewski, C. in (1997) *The Electrochemical Double Layer* (ed. B.E. Conway), The Electrochemical Society Inc.

27 West, W.C., Sieradzki, K., Kardynal, B. and Kozicki, M.N. (1998) *Journal of the Electrochemical Society*, **145**, 2971.

28 Dini, J.W. (ed.) (1992) *Electrodeposition: the Materials Science of Coatings and Substrates*, Noyes Publications NJ, USA.

29 Budevski, E., Staikov, G. and Lorenz, W.J. (1996) *Electrochemical Phase Formation and Growth*, VCH Publishers, NY, USA.

30 Watanabe, T. (2004) *Nano-Plating Microstructure Control Theory of Plated Film and Data Base of Plated Film Microstructure*, Elsevier.

31 Kozicki, M.N., Maroufkhani, P. and Mitkova, M. (2004) *Superlattices and Microstructures*, **34**, 467.

32 See for example: Henrich, V.E. and Cox, P.A. (1994) *The Surface Science of Metal Oxides*, Cambridge University Press, Cambridge. Chapter 5.

33 Witten, T.A. and Sander, L.M. (1981) *Physical Review Letters*, **47**, 1400.

34 Meakin, P. (1983) *Physical Review A*, **27**, 1495.

35 Sawada, Y., Dougherty, A. and Gollub, J.P. (1986) *Physical Review Letters*, **56**, 1260.

36 Mitkova, M., Kozicki, M.N. and Aberouette, J.P. (2003) *Journal of Non-Crystalline Solids*, **425**, 326–327.

37 Ratnakumar, C., Mitkova, M. and Kozicki, M.N. (November 2006) Proceedings of the 2006 Non-Volatile Memory Technology Symposium, San Mateo, California p. 111.

38 Dignam, M.J. (1968) *Journal of Physics and Chemistry of SolidS*, **29**, 249.

39 Kozicki, M.N., Yun, M., Hilt, L. and Singh, A. (1999) *Electrochemical Society Proceedings*, **13**, 298.

40 Miyatani, S.-Y. (1960) *Journal of the Physical Society of Japan*, **15**, 1586.

41 Symanczyk, R., Balakrishnan, M., Gopalan, C., Grüning, U., Happ, T., Kozicki, M., Kund, M., Mikolajick, T., Mitkova, M., Park, M., Pinnow, C., Robertson, J. and Ufert, K. (November 2003) Proceedings of the 2003 Non-Volatile Memory Technology Symposium, San Diego, California, p. 17-1.

42 Kund, M., Beitel, G., Pinnow, C., Röhr, T., Schumann, J., Symanczyk, R., Ufert, K. and Müller, G. (2005) *IEDM Technical Digest*, 31.5.

43 Sakamoto, T., Sunamura, H., Kawaura, H., Hasegawa, T., Nakayama, T. and Aono, M. (2003) *Applied Physics Letters*, **82**, 3032.

44 Terabe, K., Hasegawa, T., Nakayama, T. and Aono, M. (2005) *Nature*, **433**, 47.

45 Kaeriyama, S., Sakamoto, T., Sunamura, H., Mizuno, M., Kawaura, H., Hasegawa, T., Terabe, K., Nakayama, T. and Aono, M. (2005) *IEEE Journal of Solid State Circuits*, **40**, 168.

46 Sakamoto, T., Banno, N., Iguchi, N., Kawaura, H., Kaeriyama, S., Mizuno, M., Terabe, K., Hasegawa, T. and Aono, M. (2005) *IEDM Technical Digest*, 19.5.

47 Kozicki, M.N., Gopalan, C., Balakrishnan, M. and Mitkova, M. (2006) *IEEE Transactions on Nanotechnology*, **5**, 535.

48 Schindler, C., Thermadam, S.C.P., Waser, R. and Kozicki, M.N.(2007) *IEEE Trans. Electron Devices*, **54**, 2762.

49 Hönigschmid, H., Angerbauer, M., Dietrich, S., Dimitrova, M., Gogl, D., Liaw, C., Markert, M., Symanczyk, R., Altimime, L., Bournat, S. and Müller, G. (June 2006) IEEE VLSI Circuits Symposium, Honolulu, Hawaii, 132.

50 Fujita, S., Fujita, S., Abe, K. and Lee, T.H. (May 2005) *NSTI-Nanotech*, Anaheim, California. 31.04.

51 Kim, C.-J., Yoon, S.-G., Choi, K.-J., Ryu, S.-O., Yoon, S.-M., Lee, N.-Y. and Yu, B.-G. (2006) *Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures*, **24**, 721.

52 Mitkova, M., Kozicki, M.N., Kim, H.C. and Alford, T.L. (2006) *Journal of Non-Crystalline Solids*, **352**, 1986.

53 Kozicki, M.N., Balakrishnan, M., Gopalan, C., Ratnakumar, C. and Mitkova, M. (2005) IEEE Non-Volatile Memory Technology Symposium, D5, p. 1.

54 Kozicki, M.N., Gopalan, C., Balakrishnan, M., Park, M., Mitkova, M. (November, 2004) Proceedings of the 2004 Non-Volatile Memory Technology Symposium, Orlando, Florida, USA, p. 10.

55 Gilbert, N.E., Gopalan, C. and Kozicki, M.N. (2005) *Solid-State Electronics*, **49**, 1813.

56 Gilbert, N.E. and Kozicki, M.N. (2007) *IEEE Journal of Solid-State Circuits*, **42**,1383.

I
Logic Devices and Concepts

Nanotechnology. Volume 4: Information Technology II. Edited by Rainer Waser

ISBN: 978-3-527-31737-0

1
Non-Conventional Complementary Metal-Oxide-Semiconductor (CMOS) Devices

Lothar Risch

1.1
Nano-Size CMOS and Challenges

The scaling of complementary metal-oxide-semiconductor (CMOS) is key to following Moore's law for higher integration densities, faster switching times, and reduced power consumption at reduced costs. In today's research laboratories MOSFETs with minimum gate lengths below 15 nm have already been demonstrated. An example of such a small transistor is shown in Figure 1.1a, where the transmission electron microscopy (TEM) cross-section shows a functional, fully depleted silicon-on-insulator (SOI) transistor with 14 nm gate length, 20 nm spacers using a 17 nm thin silicon layer and a 1.5-nm gate dielectric. The gate has been defined with electron-beam (e-beam) lithography. For the contacts, elevated source drain regions were grown with selective Si epitaxy to lower the parasitic resistance, and a high dose of dopants was implanted into the epi layer for source and drain. In Figure 1.1b, a TEM cross-section through the fin of a SONOS memory FinFET is shown with a diameter of 8 nm, surrounded by the ONO charge-trapping dielectric. As can be seen, many critical features in Si-MOSFETs are already in the range in the range of 1 to 20 nm.

However, achieving the desired performance gain in electrical parameters from scaling will in time become very challenging, as indicated in the International Technology Roadmap for Semiconductors (ITRS) by many red brick walls [1] (see Figure 1.2).

The three main limiting factors for a performance increase are related to physical laws. Gate leakage stops SiO_2 scaling (see Figure 1.3), while source drain leakage reduction needs higher channel doping and shallower junctions. However, this increases junction capacitance, junction leakage, gate-induced drain currents, reduces carrier mobility and increases parasitic resistance. Because of this, transistors with astoundingly small gate lengths down to 5 nm have been realized [2]; although these are the smallest MOSFETs produced until now, their performance is worse than that of a 20-nm device.

When considering memories, the situation is not much different, and for mainstream DRAM and Floating Gate Flash several constraints can be foreseen. For DRAM, the storage capacitance at small cell size and a low leakage cell transistor

Nanotechnology. Volume 4: Information Technology II. Edited by Rainer Waser

ISBN: 978-3-527-31737-0

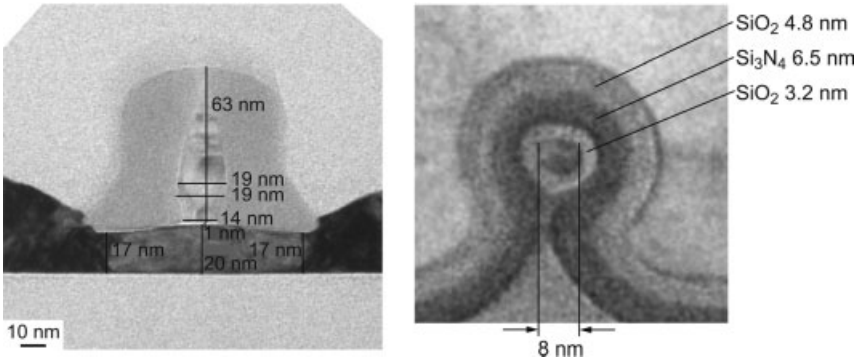


Figure 1.1 (a) A TEM cross-section of a 14-nm gate SOI transistor with raised source/drain (S/D) on 17 nm Si, $t_{ox} = 1.5$ nm. (b) TEM cross-section of a SONOS SOI FinFET across a 8-nm wire-type fin.

become a critical issue. For Floating Gate, the high drain voltages and scaling of the gate dielectric, as well as coupling to neighboring cells, are critical.

Therefore, on the way to better devices, two strategies are proposed by ITRS. The first strategy is to implement new materials as performance boosters. Among these are high-k dielectrics and metal gates, high-mobility channels and low-resistivity or

Year		04	07	10	13	16
Node [nm]		90	65	45	32	22
L_G [nm]	hp	37	25	18	13	9
	lop	53	32	22	16	11
	lstp	65	37	25	18	13
V_{dd} [V]	hp,lstp	1.2	1.1	1.0	0.9	0.8
	lop	0.9	0.8	0.7	0.6	0.5
I_{on} [mA/μm]	hp	1.1	1.5	1.9	2.05	2.4
	lop	0.53	0.57	0.77	0.78	0.92
	lstp	0.44	0.51	0.76	0.88	0.86
I_{off} [nA/μm]	hp	50	70	100	300	500
	lop	3	5	7	10	30
	lstp	0.01	0.025	0.06	0.08	0.1

Figure 1.2 ITRS 04 roadmap: gate lengths and currents for high performance, low operation power, low standby power.

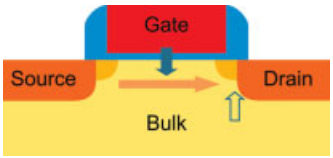


Figure 1.3 Scaling limits of scaled MOSFETs: source to drain, gate dielectric tunneling and junction leakage.

metal source drain junctions. This will lead to a remarkable improvement in the performance of transistors. The second strategy is to develop new device structures with better electrostatic control, such as fully depleted SOI and multi-gate devices. These can also be utilized in DRAMs as low leakage cell transistors, as well as in nanoscale non-volatile Flash memories.

1.2 Mobility Enhancement: SiGe, Strained Layers, Crystal Orientation

Carrier mobility enhancement of electron and holes provide the key to increase the on-currents without higher gate capacitance and without degrading the off-currents. Several methods have been developed, including SiGe heterostructures [3] with a higher hole mobility for the p-channel transistor. This is achieved by growing a thin epitaxial $Si_{1-x}Ge_x$ layer, where x is the Ge concentration, with a thickness of 5–10 nm for the channel region directly on Si (see Figure 1.4). On top of the SiGe layer a thin Si cap layer is deposited with a thickness of 3–5 nm, which is also used for the growth of the gate oxide. This forms a quantum well for the holes due to a step in the valence band of the Si/SiGe/Si heterostructure, with a depth of about 150 mV for a Ge content of 20%. The SiGe layer is under bi-axial compressive strain due to the smaller lattice constant of Si compared to SiGe (see Figure 1.4a). The mobility is enhanced because of the lower effective mass of the holes in SiGe and a split of the degenerated three-valence bands, thus reducing intervalley scattering. Compared to pure Si with a peak hole mobility of about 110 $cm^2 Vs^{-1}$, with 0.25 Ge 210 $m^2 Vs^{-1}$ have been achieved [4],

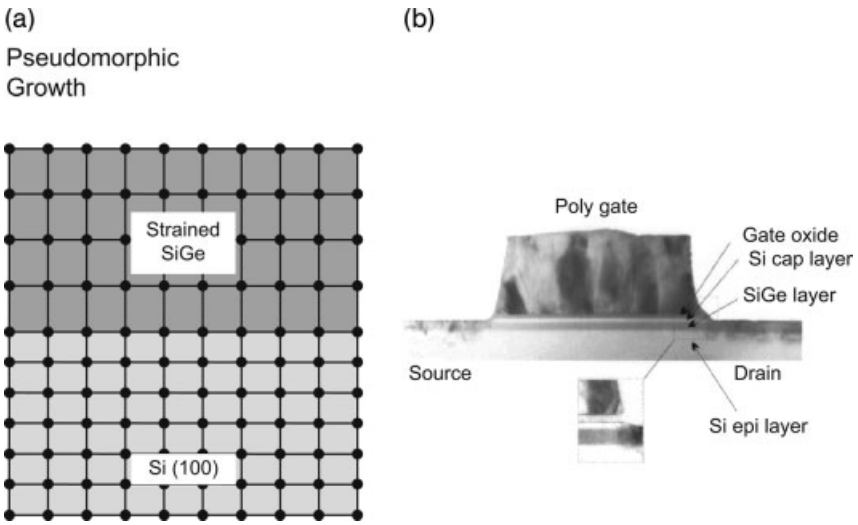


Figure 1.4 (a) Crystal lattice of a Si/SiGe heterostructure. (b) TEM cross-section of a p-channel MOSFET with a Si/SiGe/S quantum well.

extracted from MOSFET measurements. Whereas, the SiGe channel on Si is beneficial for the hole mobility, strained silicon offers both an improved electron and hole mobility, together with a surface channel [5]. The strain is created by a relaxed graded SiGe buffer layer, typically with a thickness of about 3 μm and a Ge concentration of 20–30%. A thin Si layer is grown on top of the relaxed SiGe layer in the range of 10 to 20 nm, which is now under biaxial tensile strain due to the larger lattice constant of the SiGe buffer layer.

Both techniques provide global bi-axial strain on the wafer and are based on Si/SiGe epitaxy. A critical issue here is the increased process complexity, the density of defects and wafer cost. Moreover, the implementation of tensile strain for the n-channel and compressive strain for the p-channel would be desirable, and is difficult to achieve with global strain. Therefore, local uni-axial strain techniques have now become mainstream for mobility enhancement, and these can provide tensile and compressive strain by depositing dedicated layers around the transistor. This method was introduced [6] for the 90-nm CMOS generation. In the n-channel transistor a nitride capping layer with tensile strain is used to improve the drive current by 10–15%. For the p-channel transistor, an embedded SiGe source drain region provides compressive strain and increases the drive current by 25%. TEM cross-sections of the n- and p-channel devices are shown in Figure 1.5 [6].

Another mobility-enhancement technique is based on the crystal orientation dependence of the mobility. Until now, the (1 0 0) surface of silicon wafers has been used with a channel orientation of the transistors in the <0 1 1> direction (see Figure 1.6). This is optimal for the electron mobility but will decrease the hole mobility, which is twice that at the (1 1 0) plane in the <1 0 0> direction. If (1 1 0) wafers are used or rotated (1 0 0) wafers by 45° with the channel in the <1 0 0> direction, the hole mobility is improved remarkably while electron mobility is reduced only moderately (see Figure 1.7) [7].

Therefore, another channel orientation is an effective means to increase p-channel performance, and an improvement of up to 15% has been reported [8]. Unfortunately, the embedded SiGe source drain regions with compressive strain have no remarkable influence in this crystal direction.

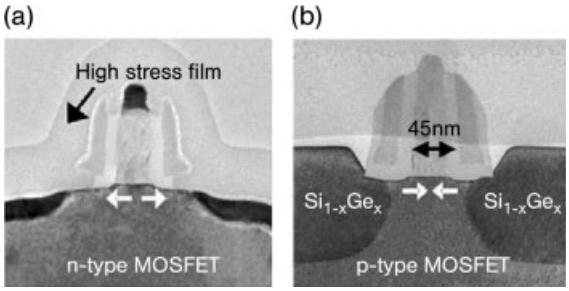


Figure 1.5 (a) 90-nm technology NMOS transistor with tensile stress nitride layer; (b) PMOS, showing heteroepitaxial SiGe source/drain inducing compressive strain [6] .

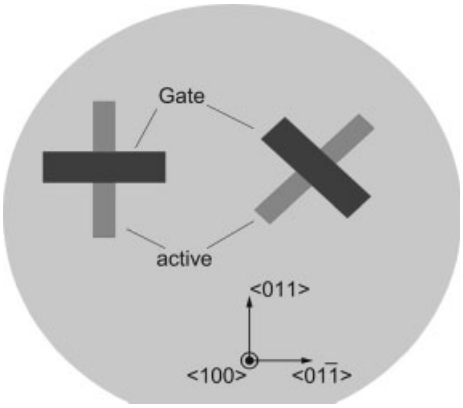


Figure 1.6 Crystal orientation and channel direction on (100) Si wafers.

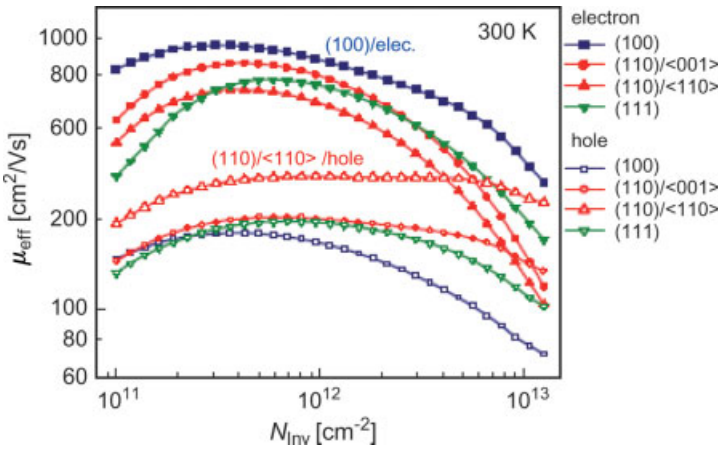


Figure 1.7 Mobility dependence for electrons and holes on crystal orientation and channel direction [7].

1.3
High-*k* Gate Dielectrics and Metal Gate

As indicated in the ITRS roadmap, scaling of the classical SiO_2 gate dielectric and increasing the gate capacitance in order to achieve higher drive currents reached completion at about 2 nm for low standby power, due to Fowler–Nordheim tunneling currents through the gate dielectric. By using nitrided oxides, the minimum thickness could be extended to about 1 nm for high-performance applications with a gate leakage current of about 10^3 A cm^{-2} [9]. The introduction of high-*k* dielectrics allows the use of thicker dielectric layers in order to reduce the tunneling currents at the same equivalent oxide thickness, or to provide thinner dielectrics for continuous scaling. Unfortunately,

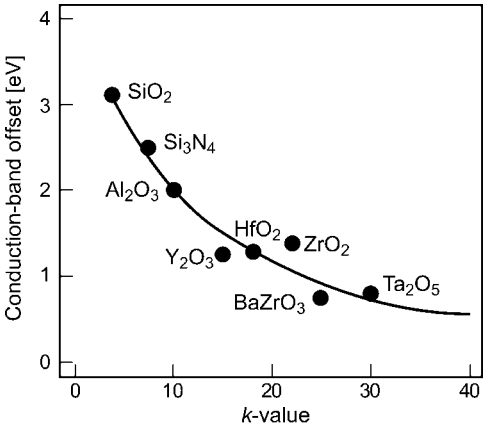


Figure 1.8 Conduction band offset versus k-value for different high-*k* materials [10].

all known high-*k* materials have a smaller bandgap than SiO_2. In Figure 1.8 the conduction band offset as a function of the dielectric constant is shown for different materials [10]. For the highest *k* materials such as Ta_2O_5 ($k = 30$) or TiO_2 ($k = 90$), the bandgap becomes too small and leads to increased gate leakage. Other critical issues are the growth of an interfacial layer during processing. Today, the most mature high-*k* dielectrics are based on Hf. Among these, HfO_2 ($k = 17$–25), HfSiO ($k = 11$) and HfSiON ($k = 9$–11), the latter are the more temperature-stable. An equivalent oxide thickness of below 1 nm has been demonstrated for these high-*k* materials [10]. Other candidates are ZrO_2 and La_2O_3 with dielectric constants between 20 and 30; however, the former is incompatible with a poly silicon gate and requires a metal gate.

For most high-*k* dielectrics a degradation of mobility is observed due to an increased scattering by phonons or a high fixed charge density at the interface. Especially for Al_2O_3, the hole mobility reduction is not acceptable. For the best Hf-based high-*k* dielectrics a 20% lower mobility has been achieved until now, compared to SiO_2.

Closely related to the high-*k* dielectric is a new gate material which avoids the depletion layer of poly silicon gates and the reaction of the high-*k* material with silicon at higher process temperatures. Moreover, metal gates offer the possibility of adjusting the threshold voltage with the workfunction of the gate material instead of doping in the channel, and this decreases the mobility at higher doping concentrations. The desired workfunctions for bulk with n+ poly and p+ poly silicon gates for low-power/high-performance applications with low doped transistor channels are shown in Figure 1.9.

Midgap-like materials such as TiN, TiSiN and W are suitable for n- and p-channel transistors with threshold voltages in the range of 300 to 400 mV, especially for fully depleted SOI or multi-gate transistors with lower channel doping concentrations. For optimized logic processes with low V_t transistors for high performance, in the range of 100 to 200 mV, dual metals with n+ and p+ poly-silicon-like workfunctions must be integrated. For n-channel transistors Ru is a candidate, and for p-channel Ta or RuTa alloys.

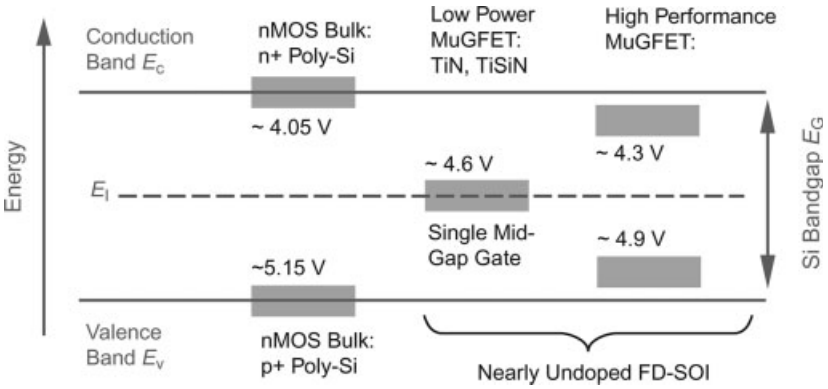


Figure 1.9 Desired workfunction for bulk and FD MOSFETS [24], Pacha ISSCC 2006.

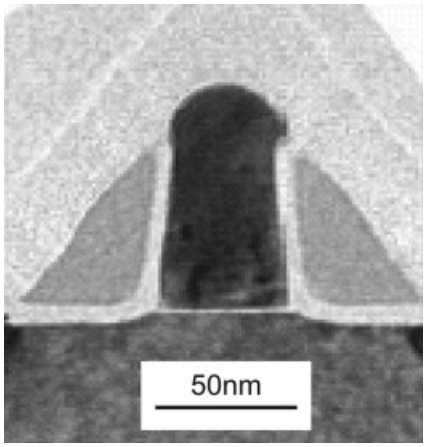


Figure 1.10 A fully silicided NiSi gate transistor [10].

Another gate material option is a tunable workfunction, such as fully silicided NiSi implanted with As and B, or Mo implanted with N. Until now, a shift of the workfunction in a conduction band direction of 200 to 300 mV has been reported [11]. A cross-section of a 50-nm transistor with a fully silicided NiSi gate is shown in Figure 1.10. Here, two approaches have been pursued: the first approach, with Thin Poly, allows the simultaneous silicidation of the source/drain (S/D) and gate, while the second approach uses CMP, offers the independent silicidation of the S/D and gate, and also avoids the formation of thick silicides in the S/D [10].

1.4 Ultra-Thin SOI

Many of the device problems due to short channel effects are related to the silicon bulk. The SOI [12] uses only a thin silicon layer for the channel, which is isolated from

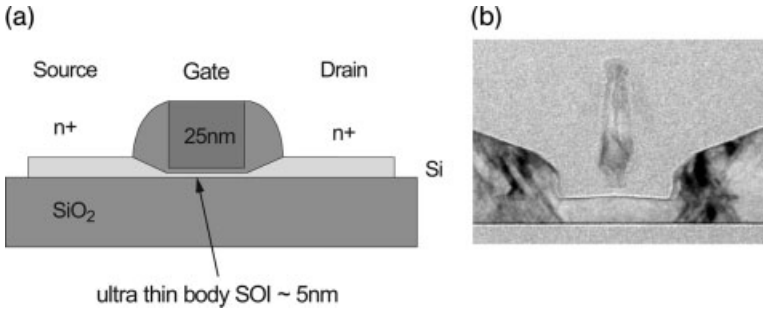


Figure 1.11 (a) A schematic cross-section of a fully depleted SOI transistor with a raised source drain. (b) TEM cross-section of a 12-nm gate fully depleted SOI transistor on 16-nm silicon.

the bulk by a buried oxide. Several companies producing semiconductors have already switched to SOI for high-performance microprocessors or low-power applications. Typically, the thickness of the Si layer is in the range of 50 to 100 nm, and the doping concentrations are comparable to those of bulk devices. This situation, which is referred to as *partially depleted SOI*, has several advantages, most notably a 10–20% higher switching speed. However, further down-scaling faces similar issues as the bulk, and here thinner Si layers [13], which lead to fully depleted channels, are of interest.

A schematic representation and a TEM cross-section of a thin-body SOI transistor with 12-nm gate length and 16-nm Si thickness on 100 nm buried oxide is shown in Figure 1.11. The gate has been defined with e-beam lithography while, for the contacts, raised source drain regions were grown with selective Si epitaxy and a high dose of dopands was implanted into the epi layer.

The experimental current–voltage (I–V) characteristics of n-channel SOI transistors with gate lengths down to 12 nm are shown in Figure 1.12. For gate lengths

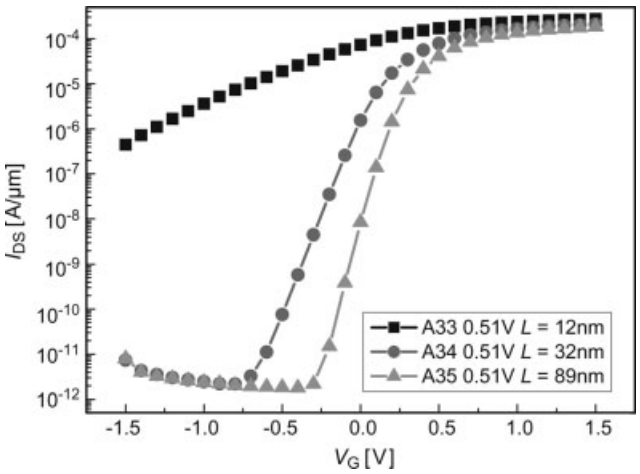


Figure 1.12 Experimental I–V characteristics of 89 to 12-nm SOI transistors on 16-nm silicon with undoped channel, n+ poly gate, $t_{ox} = 1.5$ nm.

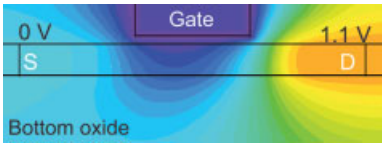


Figure 1.13 Potential distribution in a 30-nm single gate SOI transistor ($t_{Si} = 10$ nm, $t_{ox} = 2$ nm, $V_g = 0$ V, $V_d = 1.1$ V, midgap gate material).

>32 nm, subthreshold slopes of 65 mV dec^{-1} have been reached but, due to the still relatively thick Si body of 16 nm, short channel effects begin to increase below 30 nm gate length, and the transistors with 12 nm gate length cannot easily be turned off.

A two-dimensional (2-D) device simulation of the electrostatic potential of an SOI transistor with undoped channel and a thinner silicon body of 10 nm is shown in Figure 1.13 at a drain voltage of 1.1 V and a gate voltage of 0 V. For a gate length of 30 nm the gate potential controls the channel quite well. However, even with 10 nm Si thickness the potential barrier is slightly lowered at the bottom of the channel.

>This gives rise to an increase in the subthreshold slope as function of gate length, even for 5 nm Si thickness and 1 nm gate oxide (see the device simulation in Figure 1.16). A single-gate SOI exhibits the ideal subthreshold slope of 60 mV dec^{-1} down to about 50-nm gate lengths. In the gate length range of 50 to 20 nm, the turn-off characteristics are still good, and therefore ultra-thin SOI can provide a device architecture which is superior to that of bulk and suitable for the 32-nm node. A simple scaling rule for fully depleted SOI devices proposes a Si thickness of about one-fourth of the gate length in order to achieve good turn-off characteristics.

Whilst in these devices the channel was either low or undoped, this is not feasible in bulk devices because of the punch through from source to drain. The mobility of the charge carriers and the on-current is higher due to lower electric fields; this is shown graphically in Figure 1.14 for different channel doping concentrations. At a

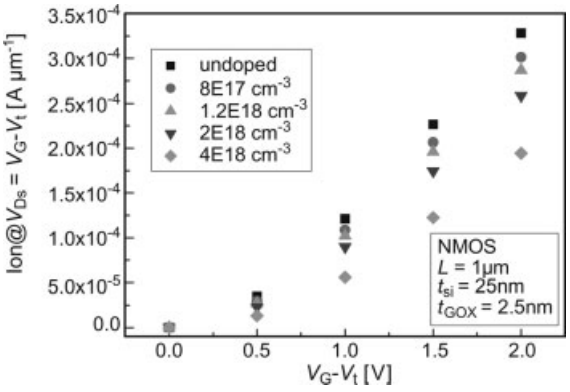


Figure 1.14 Measured on-currents at doped and undoped fully depleted SOI transistors at $V_g - V_t = 1$V.

gate voltage overdrive of 1 V the saturation current of the undoped transistor is twice that of the doped channel, at 4E18 cm^{-3} [14].

Moreover, without channel doping the Zener tunneling currents are reduced as well as electrical parameter variations, due to statistical fluctuations of the doping atoms.

1.5 Multi-Gate Devices

Further reduction of the gate length will require two or more gates for control of the channel, together with thin Si layers. The advantage of a multi-gate is to suppress the drain field much more effectively.

This is illustrated in Figure 1.15, by using the same simulation conditions as in Figure 1.13 and adding a bottom gate to the 30-nm SOI transistor. As shown in Figure 1.15, the electrostatic potential barrier is much higher than in the single-gate device. The better electrostatic control results in a steeper subthreshold slope; this can be seen in Figure 1.16, with a drift diffusion simulation of single- and double-gate transistors. A very thin Si thickness of 5 nm and a equivalent gate oxide thickness of 1 nm has been assumed, with a drain voltage of 1 V. Compared to the single gate, a

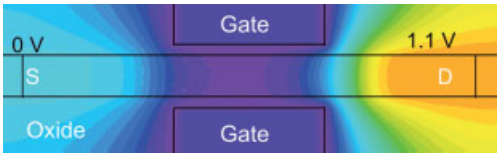


Figure 1.15 Electrostatic potential in a double-gate transistor with 30-nm gate length and 10-nm Si thickness; $V_g = 0\,V$; $V_d = 1.1\,V$; midgap gate material.

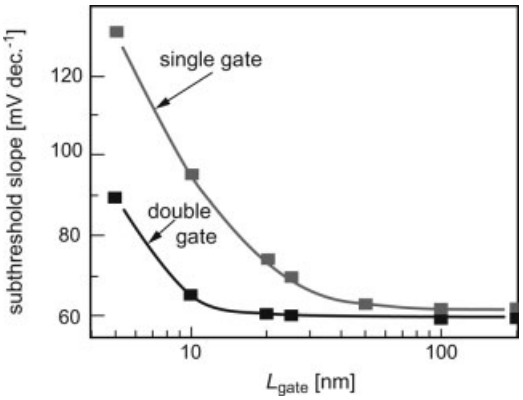


Figure 1.16 Simulated subthreshold slopes of single- and double-gate SOI transistors.

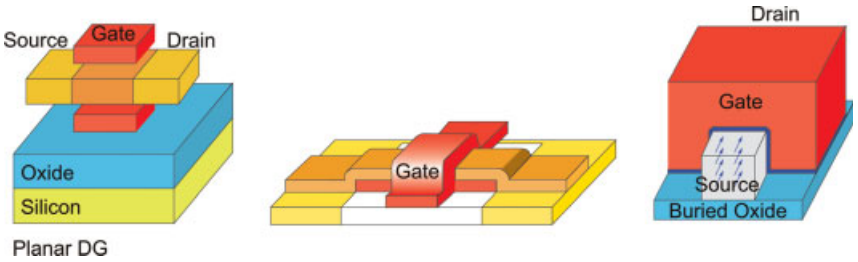


Figure 1.17 Three architectures for multi-gate devices. Left: Planar double-gate wafer-bonded [16]; Center: Gate all-around device [17]; Right: FinFET [18].

10-nm gate length and a subthreshold slope of 65 mV dec^{-1} are predicted for a double gate, and even 5 nm seems feasible with a reasonable subthreshold slope.

The challenge for multi-gate transistors will be to develop a manufacturable process with self-aligned gates to S/D regions. Three promising concepts have been investigated within the EC project NESTOR [15]: the first was a planar double-gate SOI transistor, which uses wafer bonding [16]; the second was a gate all-around device, based on silicon-on-nothing (SON) [17]; and the third was a FinFET type [18] (see Figure 1.17).

1.5.1 Wafer-Bonded Planar Double Gate

The planar double-gate transistor is an evolution of the ultra-thin SOI transistor, with a top and a bottom gate being used for better control of the channel. Processing starts with the bottom gate, spacers and elevated S/D regions using a SOI wafer with a thin silicon layer (see Figure 1.18). The gate is then encapsulated with dielectric layers and planarized with chemical mechanical polishing (CMP). Next, a handle wafer with an oxide layer is bonded onto the wafer with the bottom gate. The bulk Si of the top wafer is then completely removed down to the buried oxide, which acts as an etch stop. After removal of the buried oxide, a gate dielectric is deposited on the thin Si layer. Finally, the top gate and metallization are processed as in a conventional transistor.

An atomic force microscopy (AFM) image of a double-gate transistor test-structure with two separated contacts for the bottom and top gate is depicted in Figure 1.19a, using e-beam litho and an alignment mark for the top gate. A TEM cross-section of the first devices with a p+ poly-Si top and a n+ poly-Si bottom gate for V_t adjustment is shown in Figure 1.19b [19].

Recently, functional double-gate transistors with a TiN metal gate and lengths down to 12 nm and 8 nm for the top and bottom gates have been demonstrated [16] (see Figure 1.20). The 20-nm devices show good short-channel characteristics with $S = 102$ mV dec^{-1}, an off-current in the range of 1 μA μm^{-1}, and an on-current of 1250 μA μm^{-1}.

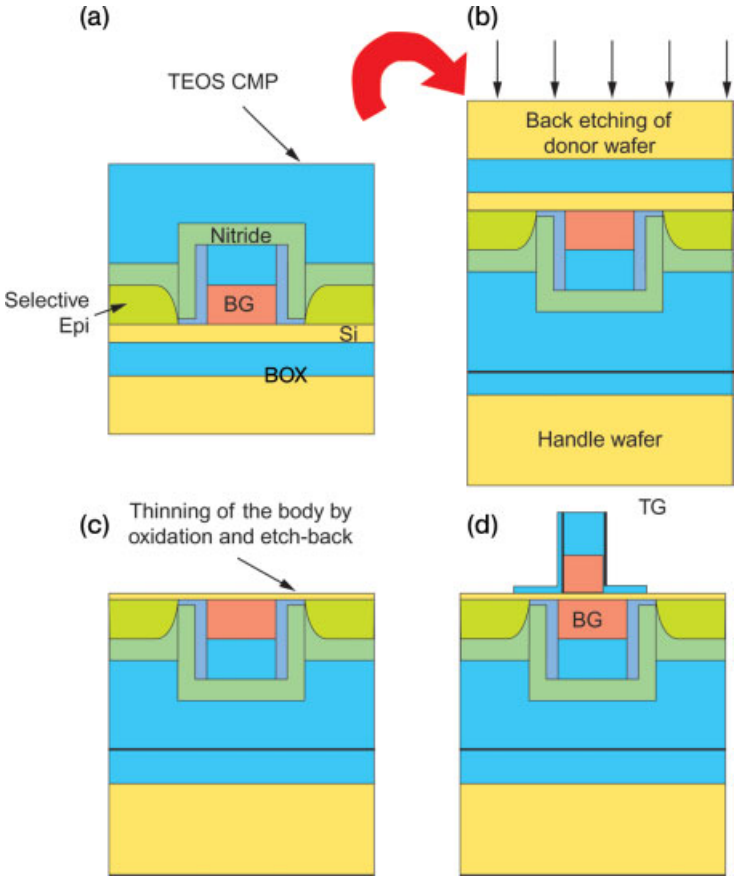


Figure 1.18 Process flow for a wafer-bonded double-gate transistor: Bottom gate, raised source drain and planarization, wafer bonding and back etch of Si bulk wafer, back etch Si channel, gate dielectric and top gate. BOX = Buried Oxide; BG = Buried Gate; TEOS = tetraethyl orthosilicate; CMP = chemical mechanical polishing.

1.5.2
Silicon-On-Nothing Gate All Around

The second approach for multi-gate architectures is based on silicon-on-nothing, as proposed by [20], which uses bulk Si wafers instead of SOI. A SiGe layer is grown with selective chemical vapor deposition (CVD) epitaxy and on top, non-selectively, a thin Si layer for the channel (see Figure 1.21). Next, the SiGe layer is removed by an isotropic etching process. The gate dielectric is then deposited around the silicon bridge, followed by the gate material, which is either poly-Si or a TiN metal gate. A 40-nm gate length and very thin Si channels down to 15 nm have been successfully

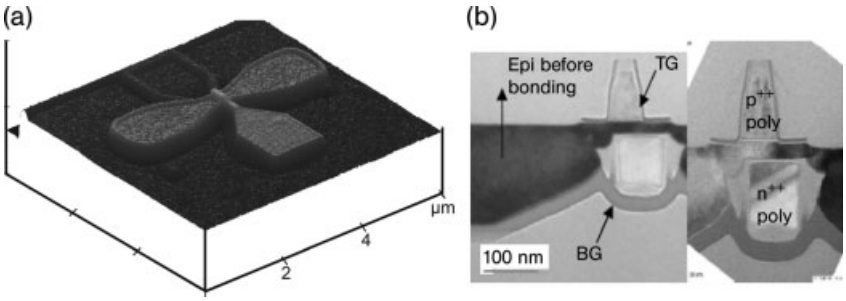


Figure 1.19 (a) AFM image of planar double-gate transistor with top and bottom gate. (b) TEM cross-section of planar double-gate transistor with n+/p+ poly gates.

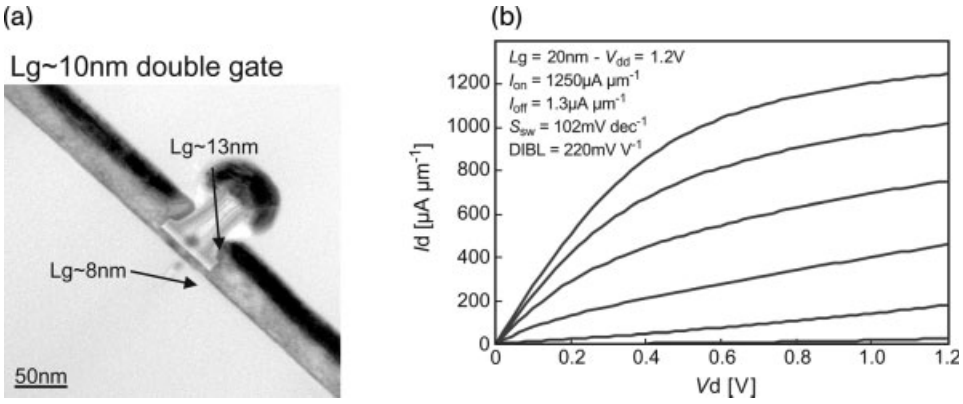


Figure 1.20 (a) TEM cross-section of a 10-nm double-gate transistor realized with wafer bonding [16]. (b) *I–V* characteristics of a 29-nm wafer-bonded double-gate device with a TiN metal gate [16].

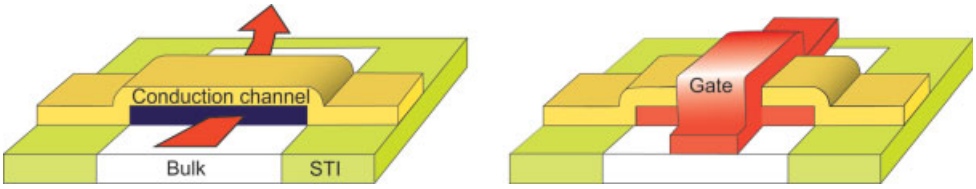


Figure 1.21 Gate all-around transistor processing based on silicon-on-nothing (SON) with a SiGe layer, which is removed for the gate [17].

fabricated [17]. Within the EC project NESTOR, devices with gate lengths of 25 nm have been achieved (see Figure 1.22a). These exhibit excellent short-channel characteristics, with $S = 70\,\mathrm{mV\,dec^{-1}}$, $\mathrm{DIBL} = 11.8\,\mathrm{mV}$, and high on currents of $1540\,\mu\mathrm{A\,\mu m^{-1}}$ ($I_{\mathrm{off}} = 2\,\mu\mathrm{A\,\mu m^{-1}}$, $t_{\mathrm{ox}} = 2\,\mathrm{nm}$) at 1.2 V (see Figure 1.22b). As shown in

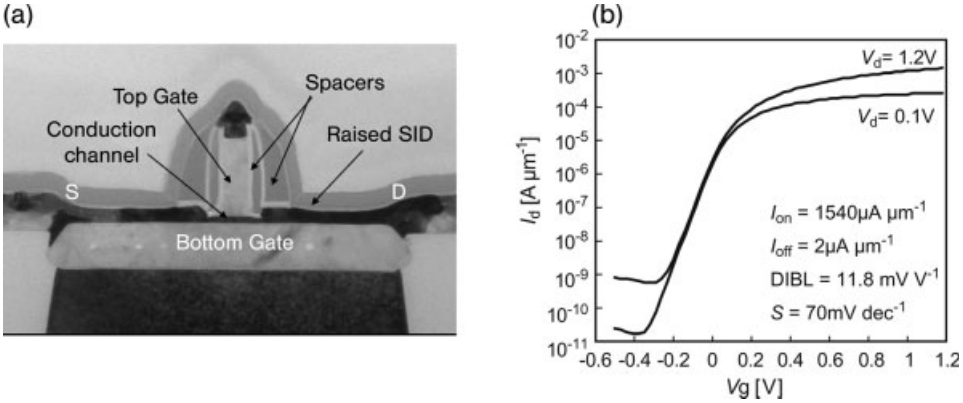


Figure 1.22 (a) TEM cross-section of 25-nm gate all-around SON transistor ($t_{ox} = 2$ nm; $t_{Si} = 10$ nm [15]. (b) Electrical characteristics of 25-nm gate all-around transistor [15].

Figure 1.22a, the bottom gate is still larger than the top gate. Ongoing studies have focused on a reduced bottom gate capacitance and a self-aligned approach.

Recently, multi-bridge transistors [21] have been reported using a similar type of SiGe layer etch technique for the fabrication of two or more channels stacked above each other and with an on-current of up to 4.2 mA μm^{-1} at 1.2 V.

1.5.3 FinFET

The FinFET [18, 22] can provide a double- or triple-gate structure with relatively simple processing (see Figure 1.23). First, the fin on SOI is structured with a tetraethylorthosilicate (TEOS) hardmask (Figure 1.23, left). A Si_3N_4 capping layer shields the top of the fin for a double-gate FinFET, and the same process flow can be used for triple-gate devices, without the capping layer. Next, a gate dielectric and the poly-Si gate are deposited and structured with litho and etching (Figure 1.23, center). The buried oxide provides an etch stop for the definition of the fin height. After this, a gate spacer is formed, raised source/drain regions are grown with epitaxy, and highly doped n+ or p+ regions implanted (Figure 1.23, right). The source/drain regions are enhanced using selective Si epitaxy to lower the sheet resistance. The facet of the Silicon epitaxy has been optimized to reduce the capacitance of drain to gate.

A TEM cross-section of a 20-nm tri-gate FinFET [23] is shown in Figure 1.24. Here, the top of the Si fin is also used for the channel, and no corner effects are observed at low fin doping concentrations. The fin and the gate layer have been processed with e-beam lithography. The smallest fin widths are in the range of 10 nm (see also Figure 1.30).

TEM cross-sections of a tri-gate device with larger fins of about 36 nm are also shown in Figure 1.24. The fin height is in the range of 35 nm, the corners are rounded by sacrificial oxidation, the gate dielectric is 2–3 nm SiO_2, and the poly gate surrounds the fin with a slight under-etch of the buried oxide.

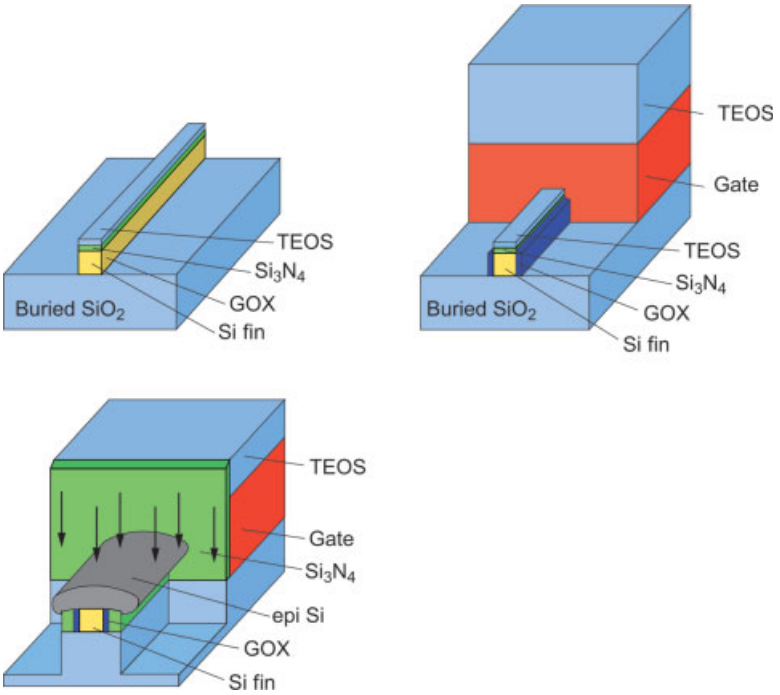


Figure 1.23 Process flow for a FinFET on buried oxide with a capping layer on top of the fin, a poly-Si gate, and raised source/drain regions with implantation. For details, see the text.

The measured I–V_g characteristics of n- and p-channel FinFETs with 20-nm and 30-nm gate length, respectively, are depicted in Figure 1.25. For the n-channel transistor a saturation current of 1.3 mA μm^{-1} (normalized by fin height) at an off-current of 100 nA μm^{-1} has been achieved at a gate voltage of 1.2 V, despite a relaxed

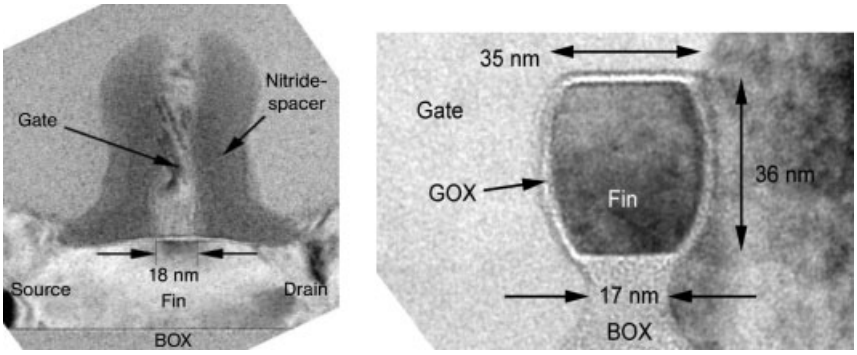


Figure 1.24 TEM cross-sections of a tri-gate FinFET on 100-nm buried oxide along and across the fin. Left: cross section along the fin; only the gate length is visible (18 nm). Right: cross section of the fin; on top it is 35 nm, height 36 nm, bottom ~17 nm.

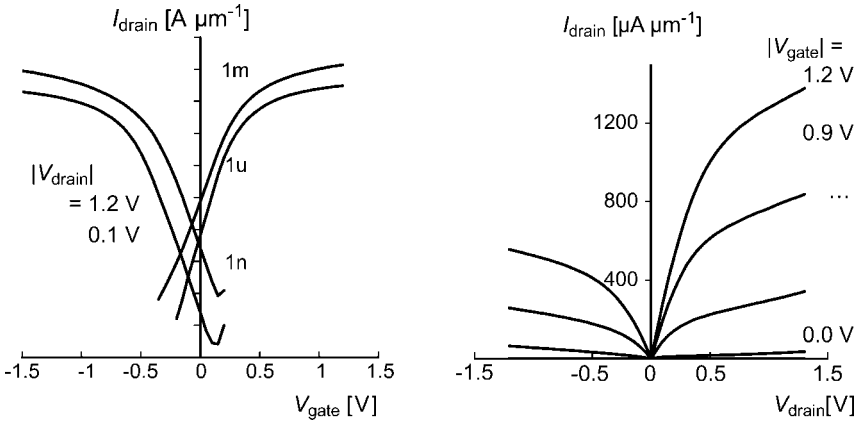


Figure 1.25 Measured *I–V* characteristics of 20-nm n-FinFETs (left) and 30-nm p-FinFETs (right) with $t_{ox} = 3$ nm (n) and 2 nm (p).

gate oxide thickness of 3 nm. For the p-channel, a high on current of 500 µA/µm and an off current in the range of 5 nA µm^{-1} is measured at 30-nm gate length. The FinFET has the advantage of self-aligned source and drain regions.

In Figure 1.25 the current was normalized on the height of a single fin. The electrical width of the device would be 2.2 times larger. For circuit applications, multi-fins are often needed in order to achieve higher drive currents (in Figure 1.26 the device has four fins) [24]. For a comparison with planar transistors it is important how many fins with height, width and pitch can be integrated on the same area as for the conventional device.

With respect to the switching time of multi-gate devices, the drive current together with the gate capacitance must be considered. Here, it was shown by simulation, that multi-gate devices can achieve 10–20% faster delay times compared to single-gate devices, mainly due to the better I_{off}/I_{on} ratio [25]. This was confirmed experimentally in Ref. [24] for inverter FO2 ring oscillators, where tri-gate FinFETs with TiSiN gate,

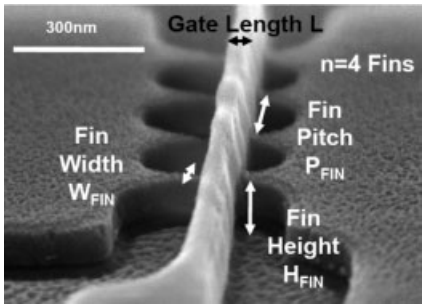


Figure 1.26 Scanning electron microscopy image of a multi-channel FinFET [24] with four fins on SOI. The gate length is 60 nm, fin width 30 nm, and pitch 200 nm.

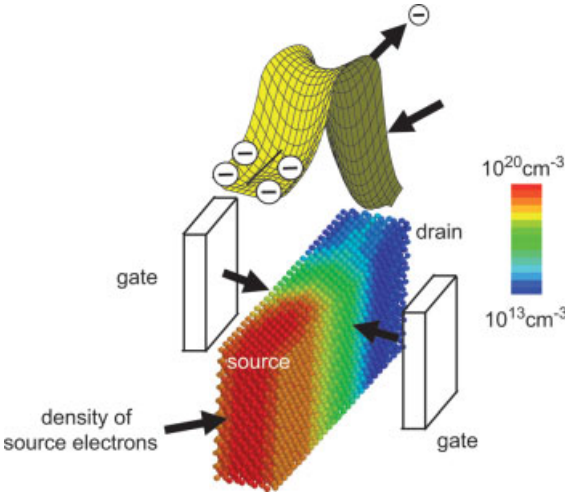


Figure 1.27 Atomistic simulation of a double-gate FinFET using the tight binding method.

55 nm length, and a low-doped channel achieved, with 21 ps, a much better speed performance than comparable planar MOSFETs in a 65 nm low-power CMOS technology, especially for sub-1 V power supply voltages.

1.5.4
Limits of Multi-Gate MOSFETs

The physical limit for the minimum channel length of multi-gate transistors has been investigated with 3-D quantum mechanical simulations using the tight binding method [26]. The device is composed of atoms in the silicon crystal lattice; the current can flow either by thermionic emission across the potential barrier of the channel, or directly via tunneling through the barrier from source to drain (see Figure 1.27).

In Figure 1.28, the simulated source drain current as a function of gate voltage is given with and without band to band tunneling for different gate lengths. An aggressive Si thickness of 2 nm and equivalent oxide thickness of 1 nm has been assumed. For gate lengths of 8 nm the tunneling contribution is on the order of the current over the potential barrier. At 4 nm the current is increased by two orders of magnitude by tunneling, but even 2-nm gates seem possible with off currents in the range of $\mu A\,\mu m^{-1}$, corresponding to ITRS hp specifications. Gate control is still effective and would achieve a subthreshold slope of about $140\,mV\,dec^{-1}$.

1.6
Multi-Gate Flash Cell

Multi-gate transistors are also very suitable for highly integrated memories with small gate lengths. Flash memory cells require thicker gate dielectrics than in logic

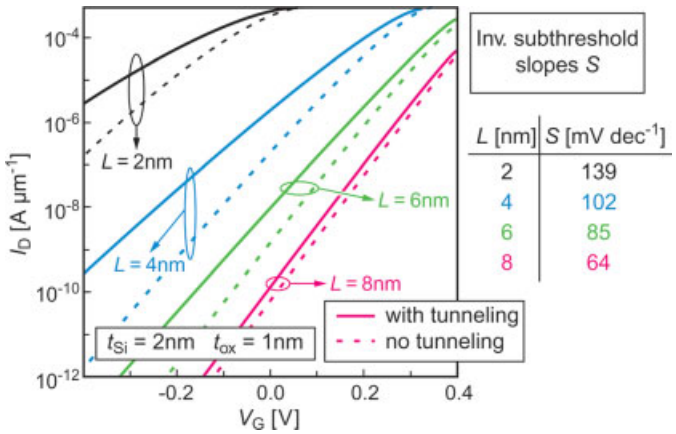


Figure 1.28 Thermionic and total current (+tunneling) of double-gate FinFETs simulated with the tight binding method [26].

applications, and therefore exhibit enhanced short channel effects. Currently, the most widely used Flash cell consists of a transistor with a floating gate [27] or a charge-trapping dielectric [28] sandwiched between the gate electrode and the channel region. A small amount of charge is transferred into the storage region either by tunneling or hot electron injection. This can be stored persistently and read out by a shift in the I–V_g characteristics. A schematic cross-section of a tri-gate FinFET memory transistor with improved electrostatic channel control compared to a planar device is shown in Figure 1.29, where a multilayer ONO gate dielectric around the fin serves as the storage element.

An experimentally realized memory structure [29] with a very small Si fin of 12 nm width and height of 38 nm is shown in Figure 1.30. The multilayer dielectric consisted of 3 nm SiO_2, 4 nm Si_3N_4 and 6.5 nm SiO_2.

The charge is uniformly injected into the nitride trapping layer by Fowler–Nordheim tunneling. The electrical function has been verified experimentally down to

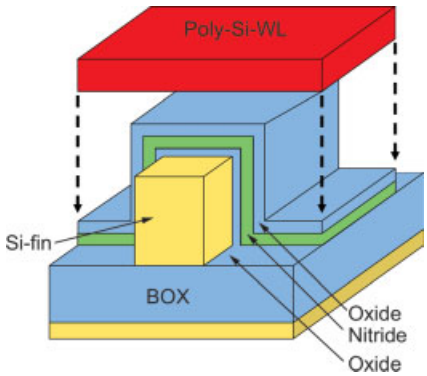


Figure 1.29 Schematic cross-section of a tri-gate charge-trapping memory field-effect transistor (FET).

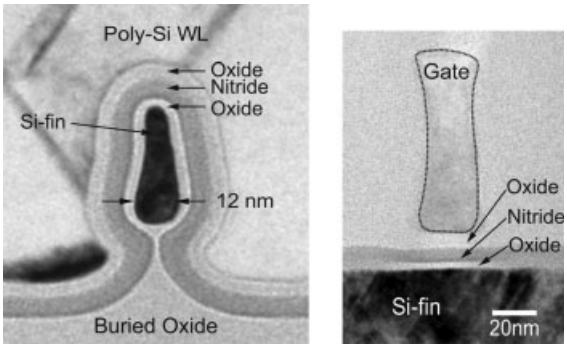


Figure 1.30 TEM cross-section of a tri-gate memory cell with 12 nm fin width and ONO dielectric.

20 nm gate length [29]. During an applied gate voltage of +12.5 V, 2 ms, electrons are injected and shift the I–V_g curves to positive voltages (write) (see Figure 1.31).

A V_t shift of about 4 V (write) was obtained using a fin width of 12 nm at gate lengths of 80 to 20 nm. The application of a negative gate voltage (erase) of 11 V, 2 ms, injects holes into the nitride layer or detraps electrons and shifts the I-Vg curves back to low V_t.

Due to the large V_t shift, multi-level storage becomes also feasible. Four levels with about 1 V separation have been programmed in the 40-nm memory transistor. The charge of one level corresponds to about 100 electrons.

The retention time for the charge-trapping dielectric has been tested, and a programming window of 3.6 V for single level was extrapolated after 10 years. Excellent retention properties between all levels are observed (see Figure 1.32).

If operated in a 4–5 F^2 high-density array such as NAND, the tri-gate cell would enable memory densities up to 32 Gbit at a die size of 130 mm^2 for the 25-nm node. A schematic NAND layout is shown in Figure 1.33. Finally, scaling is limited by the thickness of the two oxide–nitride–oxide layers, plus the minimal gate electrode thickness between the fins.

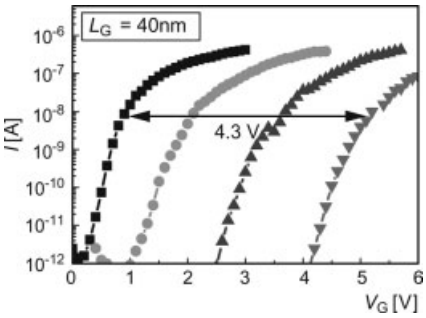


Figure 1.31 Write characteristics of a tri-gate memory cell with 40 nm gate length and multi-level operation. The different symbols represent write voltages between 0 V and 4.3 V.

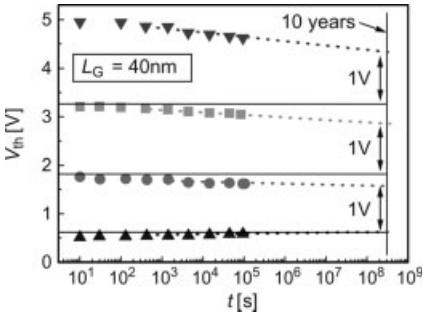


Figure 1.32 Retention time for the 40-nm tri-gate memory cell with oxide–nitride–oxide (ONO) dielectric at room temperature. The different symbols indicate the different write voltages of Figure 1.31.

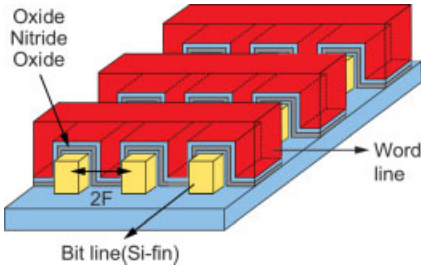


Figure 1.33 4 F^2 NAND array with tri-gate memory cells.

1.7 3d-DRAM Array Devices: RCAT, FinFET

For DRAMs, extremely low leakage current array devices below 1 fA per cell are required in order to avoid too-high charge losses during the refresh time interval. A contribution to the leakage current originates from the sub-V_t current of the cell transistor, while others are junction leakage and tunneling currents through the dielectric of the storage capacitor. With respect to the cell transistor, the channel doping cannot be increased in order to improve the turn-off characteristics, because of the electric field, which will initiate trap-assisted tunneling leakage currents [30] at $E > 0.5\,MV\,cm^{-1}$. Therefore, the planar DRAM cell transistors can be scaled down only to about 70 nm [30]. A schematic and a SEM cross-section of the 70-nm trench DRAM cell are shown in Figure 1.34.

For future applications, new cell transistor structures must be implemented. For stack DRAM cells, the transition to a recessed channel array transistor (RCAT) has already been reported for the 90- to 80-nm generation [31]. Such a device, with a U-shaped groove etched into silicon with a depth of about 200 nm, is shown in

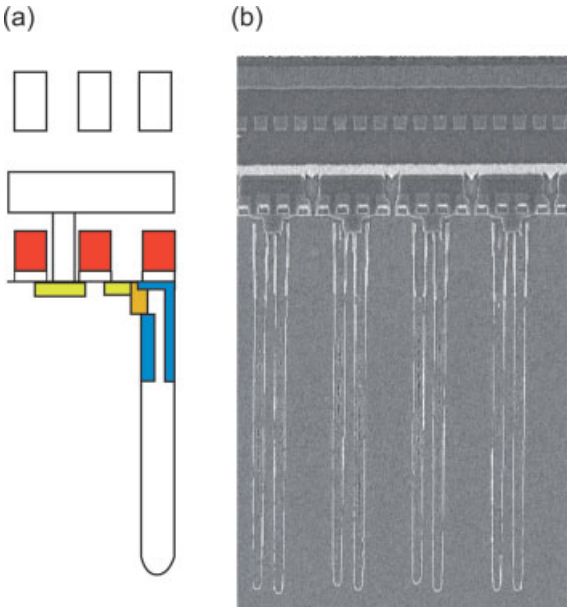


Figure 1.34 (a) Schematic cross-section of a trench DRAM cell with planar cell transistor and buried strap capacitor node contact [30]. The yellow rectangle indicates a n+ doped region in p-well; the red area is the gate (wordline); the blue is an isolation oxide. The other line is the bitline and the second wordline on top. (b) SEM cross-section of the 70-nm trench DRAM cell [30].

Figure 1.35 [32]. After gate dielectric growth, the groove is filled with the poly-silicon gate material. Bitline and storage node contacts are on the planar silicon. Such a structure is suitable for sub-70-nm generations because it provides a longer channel for lower I_{off} currents. In this Extended U-shape Device, a gate wrap-around the Si sidewalls with a depth of 6–10 nm increases the on-current and improves the subthreshold slope. The 3d device has been integrated into a 90-nm DRAM test array [32]. Simulation and measurement are shown in Figure 1.35b, with and without a corner device of about 6 nm. The subthreshold slope is in the range of 95 to 130 mV dec^{-1} at 85 °C, and the side gates enhance the on-current by 30%.

Reducing the width of the cell transistor to sublithographic dimensions and utilizing deeper vertical sidewalls leads to a fully depleted FinFET device with improved electrostatic control and increased on-currents [32]. A schematic cross-section in bitline direction of a trench cell with a FinFET array transistor, together with a SEM cross-section of a realized structure in 90 nm technology, are shown in Figure 1.36. The fin has a width of about 20 nm and a height of 50 nm. The transistor has been implemented using a local Damascene technique for fin and gate. The local gates are connected with a WSi wordline, which is also used for the gate layer of the planar transistors in the periphery circuits. The body is connected to the substrate and isolated to the neighboring fin with Sallow Trench Isolation.

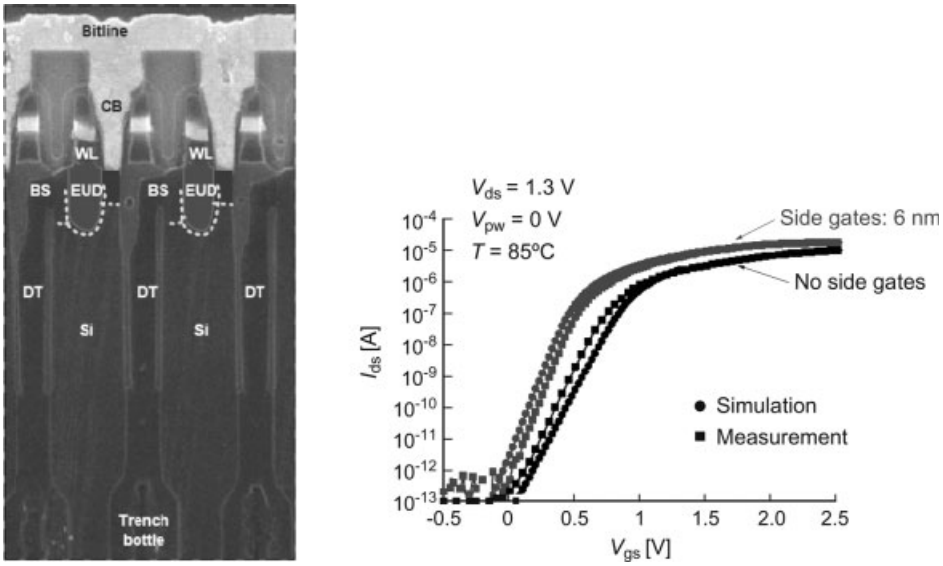


Figure 1.35 (a) SEM cross-section along bitline of a 90-nm trench DRAM cell with Extended U-Shape Device [32]. (b) Measurement and simulation of the *I–V* characteristics of the Extended U-Shape Device with and without 6-nm side gates at the corner [32].

A steep sub-threshold slope of 77 mV dec^{-1} without any drain-induced barrier lowering and back bias effect has been measured [32] in a 90-nm demonstrator (see Figure 1.37). According to simulations, the high I_{off}/I_{on} ratio will be maintained at least down to 40 nm, with a subthreshold slope of 89 mV dec^{-1} and without any remarkable influence of the adjacent trench cells, which can disturb the potential in the array device.

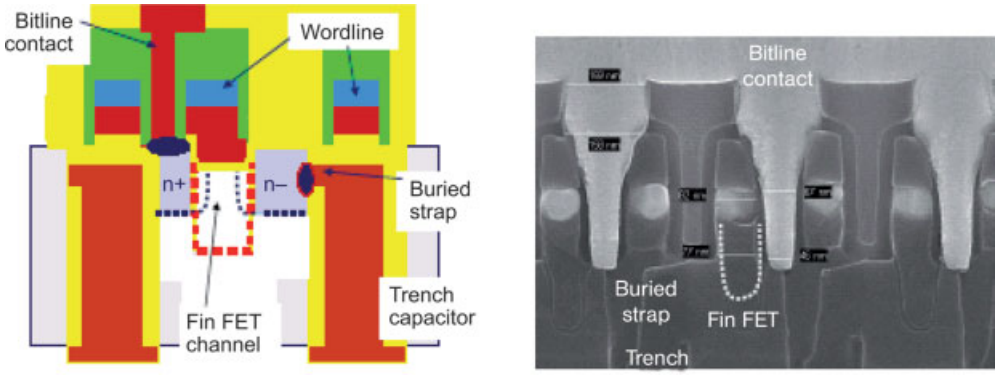


Figure 1.36 Trench DRAM with a FinFET-type cell transistor [32].

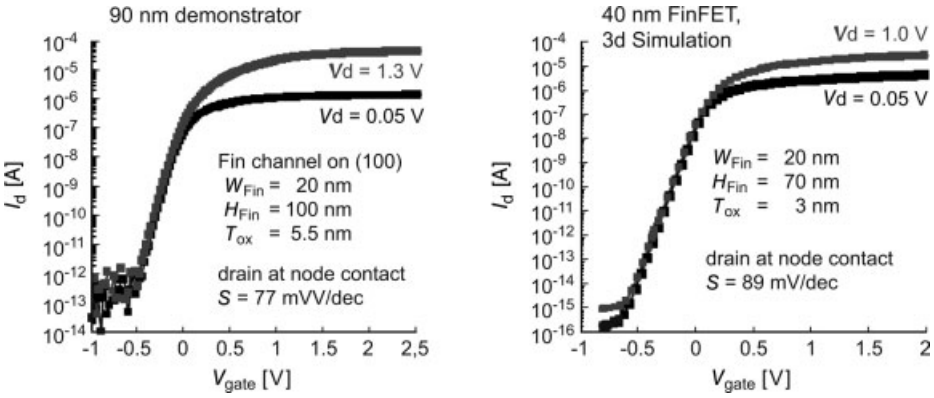


Figure 1.37 Measured FinFET array device *I–V* characteristics in a 90-nm trench cell demonstrator and simulation for 40 nm.

1.8 Prospects

Assuming that lithography tools such as Extreme Ultra-Violet will be available for the sub-45-nm technology nodes, it seems very likely that the scaling of Si CMOS will continue down to the 22-nm node, with the start of production in the year 2016, according to the ITRS roadmap. In this scenario – which is known as *More Moore* – technology costs must be reduced per chip from generation to generation, and performance must be increased. This will be expected especially for memories and microprocessors, and in order to fulfill these requirements more challenging new process modules, such as metal gate, high-*k* dielectrics, and strain will need to be integrated with high yield and in good time. On the other hand, conventional bulk CMOS may run into performance constraints below the 45-nm generation. Multi-gate devices with thin silicon channels and better electrostatic control may take over and will allow further downscaling, but with more complex processing. For DRAMs and Flash, the integration of such 3d transistors with very low leakage currents has already been started. Ultimately, beyond 10 nm the process tolerances and variability of the electrical parameters will become the most limiting factors. In addition, with the consistently good scaling potential of Si MOSFETs, many applications such as low-frequency RF, analogue, and powerFETs, displays and sensors do not require extremely small feature sizes. Therefore, additional functionality on the chip – referred to as *More than Moore* – will be another key trend.

Another important issue is the increasing research into new logic and memory devices. Among these are the 1d wire structures of Si, Ge or carbon with source, drain and gate, such as Si MOSFETs. These devices show similar *I–V* characteristics to Si (or even better), depending on the normalization of the current on the small width of the devices. However, the manufacturability and integration on a large scale has still

to be proven, and the key for success would be the integration capability with Si CMOS.

With regards to memories, many promising new concepts have appeared, based on new materials such as the storage element. Among these are included non-volatile memories, with a large change in resistance, such as Phase-Change or Conductive Bridging. These memories can be combined very well with a Si access transistor and CMOS circuitry. With these evolutionary elements, non-conventional CMOS represents the most realistic approach for high-density logic and memories, and will undoubtedly represent the dominant technology of the nanoelectronics era.

Acknowledgments

The studies on SOI MOSFETs have been partly supported within the BMBF project HSOI, and Multi-Gate Devices within Extended CMOS and the EC Project NESTOR, IST-2001-37114. The author thanks the NESTOR partners for their courtesy, especially S. Deleonibus, T. Poiroux, P. Coronel, S. Harrison, N. Collaert, and Y. Ponomarev. Thanks are also expressed to the author's colleagues at Infineon/Qimonda for their contributions, notably M. Alba, L. Dreeskornfeld, J. Hartwich, F. Hofmann, G. Ilicali, J. Kretz, E. Landgraf, T. Lutz, H. Luyken, W. Rösner, M. Specht, M. Staedele, C. Pacha, and W. Mueller.

References

1 ITRS Roadmap 2004 edition, http://public.itrs.net.

2 H. Wakabayashi, S. Yamagami, N. Ikezawa, A. Ogura, M. Narihiro, K.-I. Arai, Y. Ochiai, K. Takeuchi, T. Yamamoto, T. Mogami, Sub-10 nm planar-bulk-CMOS devices using lateral junction control (5 nm CMOS), *IEDM Technical Digest* 2003, 989.

3 D. K. Nayak, J. C. S. Park, K. Wang, K. P. MacWilliams, Enhancement-Mode Quantum-Well GexSi1-x PMOS, *IEEE-EDL* 1991, **12**, 154.

4 L. Risch, *et al.*, Fabrication and electrical characterization of Si/SiGe p-channel MOSFETs with a delta doped boron layer, Proceedings of ESSDERC, p. 465, 1996.

5 K. Rim, S. Koester, M. Hargrove, J. Chu, P. M. Mooney, J. Ott, T. Kanarsky, P. Ronsheim, M. Ieong, A. Grill, H.-S. P. Wong, Strained Si CMOS (SS CMOS) Technology, Proceedings VLSI Symposium, p. 59, 2001.

6 T. Ghani, M. Armstrong, C. Auth, M. Bost, P. Charvat, G. Glass, T. Hoffmann, K. Johnson, C. Kenyon, J. Klaus, B. McIntyre, K. Mistry, A. Murthy, J. Sandford, M. Silberstein, S. Sivakumar, P. Smith, K. Zawadzki, S. Thompson, M. Bohr, A 90 nm high volume manufacturing logic technology featuring novel 45 nm gate length strained silicon CMOS transistors, *IEDM Technical Digest* 2003, 978.

7 H. Irie, K. Kita, K. Kyuno, A. Toriumi, In-plane mobility anisotropy and universality under uni-axial strains in n- and p-MOS inversion layers on (1 0 0), (1 1 0), and (1 1 1) Si, *IEDM Technical Digest* 2004, 225.

8 H. Sayama, Y. Nishida, H. Oda, T. Oishi, S. Shimizu, T. Kunikiyo, K. Sonoda, Y. Inoue, M. Inishi, Effect of <1 0 0> channel direction for high performance SCE

immune p-MOSFET with less than 0.15 μm gate length, *IEDM Technical Digest* 1999, 657.

9 B. Tavel, M. Bidaud, N. Emonet, D. Barge, N. Planes, H. Brut, D. Roy, J. C. Vildeuil, R. Difrenza, K. Rochereau, M. Denais, V. Huard, P. Llinares, S. Bruyère, C. Parthasarthy, N. Revil, R. Pantel, F. Guyader, L. Vishnubhotla, K. Barla, F. Arnaud, P. Stolk, M. Woo, Thin oxynitride solution for digital and mixed-signal 65 nm CMOS platform, *IEDM Technical Digest* 2003, **27.6**, 643.

10 S. De Gendt, Advanced Gate Stacks: high k and metal gates, 2004 IEDM Short Course 45 nm CMOS Technology.

11 J. Kedzierski, D. Boyd, P. Ronsheim, S. Zafar, J. Newbury, J. Ott, C., Jr. Cabral, M. Ieong, W. Haensch, Threshold voltage control in NiSi-gated MOSFETs through silicidation induced impurity segregation, *IEDM Technical Digest* 2003, **13.3**, 315.

12 J. P. Colinge, *SOI Technology: Materials to VLSI*, 2nd edition, Boston, MA, Kluwer, 1997.

13 B. Doris, M. Ieong, H. Zhu, Y. Zhang, M. Steen, W. Natzle, S. Callegari, V. Narayanan, J. Cai, S. H. Ku, P. Jamison, Y. Li, Z. Ren, V. Ku, D. Boyd, T. Kanarsky, C. D'Emic, M. Newport, D. Dobuzinsky, S. Deshpande, J. Petrus, R. Jammy, W. Haensch, Device design considerations for ultra-thin SOI MOSFETs, *IEDM Technical Digest* 2003, 631.

14 J. Hartwich, L. Dreeskornfeld, F. Hofmann, J. Kretz, E. Landgraf, R. J. Luyken, M. Specht, M. Staedele, T. Schulz, W. Rösner, L. Risch, Off-current adjustments in ultra-thin SOI MOSFETs, *Proceedings of ESSDERC*, p. 305, 2004.

15 EC Project NESTOR. IST-2001-37114.

16 M. Vinet, T. Poiroux, J. Widiez, J. Lolivier, B. Previtali, C. Vizioz, B. Guillaumot, Y. Letiec, P. Besson, B. Biasse, F. Allain, M. Casse, D. Lafond, J.-M. Hartmann, Y. Morand, J. Chiaroni, S. Deleonibus, High performance 10 nm bonded planar double metal gate CMOS transistors,*IEEE-EDL* May 2005, 317.

17 S. Harrison, P. Coronel, F. Leverd, R. Cerutti, R. Palla, D. Delille, S. Borel, S. Jullian, R. Pantel, S. Descombes, D. Dutartre, Y. Morand, M. P. Samson, D. Lenoble, A. Talbot, A. Villaret, S. Monfray, P. Mazoyer, J. Bustos, H. Brut, A. Cros, D. Munteanu, J.-L. Autran, T. Skotnicki, Highly performant double gate MOS-FET realized with SON process, *IEDM Technical Digest* 2003, 449.

18 X. Huang, W.-C. Lee, C. Kuo, D. Hisamoto, L. Chang, J. Kedzierski, E. Anderson, H. Takeuchi, Y.-K. Choi, K. Asano, V. Subramanian, T.-J. King, J. Bokor, C. Hu, Sub 50 nm FinFET, *IEDM Technical Digest* 1999, 67.

19 G. Ilicali, W. Weber, W. Rösner, L. Dreeskornfeld, J. Hartwich, J. Kretz, T. Lutz, J. P. Mazellier, M. Städele, M. Specht, J. R. Luyken, E. Landgraf, F. Hofmann, L. Risch, R. Käsmaier, W. Hansch, Planar double gate transistors with asymmetric independent gates, Proceedings International SOI Conference, 2005.

20 S. Monfray, D. Chanemougame, S. Borel, A. Talbot, F. Leverd, N. Planes, D. Delille, D. Dutartr, R. Palla, Y. Morand, S. Decombes, M.-P. Samsan, N. Vulliet, T. Sparks, A. Vandooren, T. Skotnicki, SON technological CMOS platform: Highly performant devices and SRAM cells, *IEDM Technical Digest* 2004, 635.

21 S.-Y. Lee, E.-J. Yoon, S.-M. Kim, C. W. Oh, M. Li, J.-D. Choi, K.-H. Yeo, M.-S. Kim, H.-J. Cho, S.-H. Kim, D.-W. Kim, D. Park, K. Kim, A novel sub 50 nm multi-bridge-channel MOSFET (MBCFET) with extremely high performance, 2004 Symposium on VLSI Technology, p. 200.

22 B. Yu, L. Chang, S. Ahmed, H. Wang, S. Bell, C.-Y. Yang, C. Tahery, C. Ho, Q. Xiang, T.-J. King, J. Bokor, C. Hu, M.-R. Lin, D. Kyser, FinFET scaling to 10 nm gate length, *IEDM Technical Digest* 2002, 251.

23 W. Roesner, E. Landgraf, J. Kretz, L. Dreeskornfeld, H. Schäfer, M. Städele, L. Risch, Nanoscale FinFETs for low-power applications, *Solid-State Electronics* 2004, **48**, 1819.

24 C. Pacha, K.v. Arnim, T. Schulz, W. Xiong, M. Gostkowski, G. Knoblinger, A. Marshall, T. Nirschl, J. Bertold, C. Russ, H. Gossner, C. Duvvury, P. Patruno, R. Cleavelin, K. Schruefer, Circuit design issues in multi-gate FET CMOS technologies, Proceedings ISSCC, 2006.

25 M. Städele, R. J. Luyken, M. Specht, G. Ilicali, W. Rösner, L. Risch, Speed considerations of fully depleted single and double gate SOI transistors, *Proceedings ULIS*, p. 87, 2005,

26 M. Städele, A. Di Carlo, P. Lugli, F. Sacconi, B. Tuttle, Atomistic tight-binding calculations for the transport in extremely scaled SOI devices, *IEDM Technical Digest* 2003, 229.

27 J.-H. Park, S.-H. Hur, J.-H. Lee, J.-T. Park, J.-S. Sel, J.-W. Kim, S.-B. Song, J.-Y. Lee, S.-J. Son, Y.-S. Kim, M.-C. Park, S.-J. Choi, U.-I. Chung, J.-T. Moon, K.-T. Kim, K. Kim, B.-L. Ryu, 8 Gb MLC NAND flash memory using 63 nm process technology, *IEDM Technical Digest* 2004, 873.

28 J. Willer, C. Ludwig, J. Deppe, C. Kleint, S. Riedel, J.-U. Sachse, M. Krause, R. Mikalo, E. Stein, V. Kamienski, S. Parascondola, T. Mikolajick, J.-M. Fischer, M. Isler, K.-H. Kuesters, I. Bloom, A. Shapir, E. Lusky, B. Eitan, 110 nm NROM technology for code and data flash products, 2004 Symposium on VLSI Technology, p. 76.

29 M. Specht, U. Dorda, L. Dreeskornfeld, J. Kretz, F. Hofmann, M. Staedele, R. J. Luyken, W. Rösner, H. Reisinger, E. Landgraf, T. Schulz, J. Hartwich, R. Kömmling, L. Risch, 20 nm tri-gate SONOS memory cells with multi-level operation, *IEDM Technical Digest* 2004, 1083.

30 J. Amon, A. Kieslich, L. Heineck, T. Schuster, J. Faul, J. Luetzen, C. Fan, C.-C. Huang, B. Fischer, G. Enders, S. Kudelka, U. Schroeder, K.-H. Kuesters, G. Lange, J. Alsmeier, A highly manufacturable deep trench based DRAM cell layout with a planer array device in a 70 nm technology, *IEDM Technical Digest* 2004, 73.

31 H. S. Kim, D. H. Kim, J. M. Park, Y. S. Hwang, M. Huh, H. K. Hwang, N. J. Kang, B. H. Lee, M. H. Cho, S. E. Kim, J. Y. Kim, B. J. Park, J. W. Lee, D. I. Kim, M. Y. Jeong, H. J. Kim, Y. J. Park, Kinam. Kim, An outstanding and highly manufacturable 80 nm DRAM technology, *IEDM Technical Digest* 2003, **17.2**, 411.

32 W. Mueller, G. Aichmayr, W. Bergner, E. Erben, T. Hecht, C. Kapteyn, A. Kersch, S. Kudelka, F. Lau, J. Luetzen, A. Orth, J. Nuetzel, T. Schloesser, A. Scholz, U. Schroeder, A. Sieck, A. Spitzer, M. Strasser, P.-F. Wang, S. Wege, R. Weiset, Challenges for the DRAM cell scaling to 40 nm, *IEDM Technical Digest* 2005, **14.1**.

Further Reading

Short Course on Silicon+: Augmented Silicon Technology, Organizer T.-J. King, IEDM, December 7, 2003.

Emerging Nano-Electronics: Scaling MOSFETs to the Ultimate Limits and Beyond-MOSFET Approaches, Organizers P. Zeitzoff, T. Mogami, VLSI Technology Short Course, June 14, 2004.

Advanced CMOS Devices on bulk and SOI: Physics, modeling and characterization, T. Poiroux, G. Le Carval, Short Courses ESSDERC, September 12, 2005.

Non Classical CMOS: Novel Materials, Novel Device Structures, and Technology Roadmap, H.-S. Philip Wong, Short Courses ESSDERC, September 12, 2005.

2
Indium Arsenide (InAs) Nanowire Wrapped-Insulator-Gate Field-Effect Transistor

Lars-Erik Wernersson, Tomas Bryllert, Linus Fröberg, Erik Lind, Claes Thelander, and Lars Samuelson

2.1
Introduction

Semiconductor nanowires [1–9] offer the possibility to form a new class of semiconductor device. Nanowire technology enables new material combinations and also the possibility to enhance the potential control in down-scaled channels using wrap-around gates. As the lateral dimensions of semiconductor materials are scaled down towards 100 nm and below (which can be easily achieved with the nanowire technology), fewer constraints become apparent in terms of lattice matching between materials. This opens the path to a heterogeneous materials integration that cannot be accomplished with conventional bulk semiconductor technology. For example, it has been shown that segments of InP can be incorporated in indium arsenide (InAs) nanowires [10], and that InAs nanowires can be grown on Si substrates [11], in spite of about 3.5% and 7% lattice mismatch, respectively. These material combinations cannot be synthesized in the bulk, nor with planar epitaxial techniques. The second advantage is related to the challenges that the technology is facing as the planar transistors are scaled down towards the 22 nm node and beyond. At this length scale, the transistors are more sensitive to short-channel effects related to the reduced potential control in the channel and the body of the devices. This is reflected in an increased output conductance and sub-threshold swing of the transistors that degrade the transistor performance. Dual gates, trigates and FinFETs have been demonstrated to reduce these issues. Taking the technology one step further is to completely surround the channel with a wrapped gate, and for this technology vertical nanowires are ideal.

Several groups have reported on the successful fabrication of vertical nanowire transistors [12–22]. Various implementations of Si transistors have been reported and, in particular, it has been shown that the wire geometry may be used to fabricate different advanced transistors with benefits in sub-threshold characteristics and switching behavior. The present authors' effort has been focused on the vertical

Nanotechnology. Volume 4: Information Technology II. Edited by Rainer Waser

ISBN: 978-3-527-31737-0

implementation of III–V transistors, and in the following sections are described the processing and characteristics of both long- and short-channel transistors. The importance is also demonstrated of introducing a high-*k* dielectric, its influence on the device characteristics, and the benefits of heterostructure design.

2.2 Nanowire Materials

In these studies, InAs has been the primary choice of material in the transistors. For various reasons, wrap-gate transistors based on silicon will naturally have a very strong standing, due primarily to the compatibility with silicon technology in general and also to the fact that Si nanowires can be made with diameters even <5 nm and yet still be conducting. InAs, on the other hand, shows very strong lateral confinement effects already for diameters around 30 nm, making very narrow uncapped InAs transistors depleted of charge-carriers and, in that sense, less promising. In contrast, n-type InAs has highly attractive material properties, with a reproducible Fermi-level pinning in the conduction-band, with a very high room-temperature mobility and ideal ohmic-contact properties. The remainder of this chapter focuses on the use of InAs as the active transistor channel material and the use of P-containing $InAs_{1-x}P_x$-alloys for enhancement of the transistor functionality and performance.

2.3 Processing

The nanowires used here are grown with chemical beam epitaxy (CBE), using patterned Au discs to locate the nanowire growth and to set the diameter of the nanowires (Figure 2.1). The ability to form well-defined matrices of nanowires is a key feature both for the post-growth device processing and for the transistor design, in that the number of wires determines the drive current and the transconductance of the nanowire transistor. The uniformity in length provides good starting conditions for uniform top contact formation. Typically, nanowire matrices ranging from 1×1 to 10×10 are used to form the vertical transistors in order to reach drive currents approaching 10 mA, but a nanowire transistor may be defined by anything from a single wire to, say, 10^4 wires. Outside the active transistor region, smaller arrays of nanowires are formed to create alignment markers for optical lithography in the post-growth processing described below. The seed for these wires are formed in the same seed and growth steps as the actual transistor wires.

After the growth, either long-channel or short-channel transistors may be formed by processing the transistors in two different ways, as shown in Figures 2.2 and 2.4, respectively [15–22]. For the long-channel transistors, the vertical nanowire matrix is first covered by a SiN_x gate-dielectric layer, followed by a sputtered Ti/Au gate metal that is covering the nanowires uniformly. The sample is spin-coated by a photoresist, which is back-etched to the desired gate length, after which the gate metal is

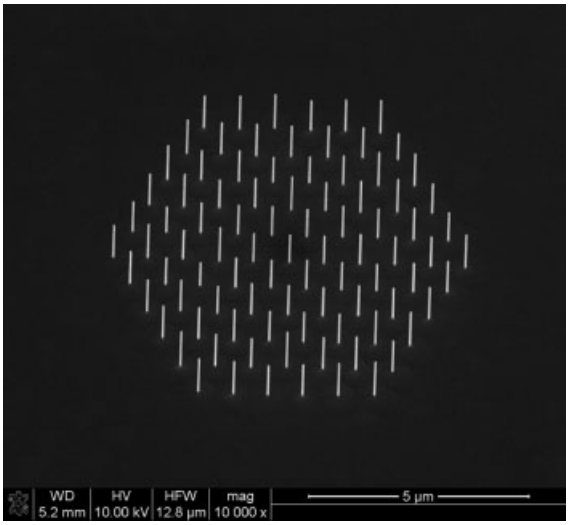


Figure 2.1 Scanning electron microscopy image of a nanowire matrix grown by chemical beam epitaxy.

selectively removed by wet-etching. After removal of the resist, the sample is covered by a second resist layer and the SiN_x is etched to open for formation of the drain Ti/Au top contact by evaporation. Finally, an airbridge is created by electroplating from the drain contact, and the resist is dissolved. With this technology, the transistor structure shown in Figure 2.3 is formed. While this technology provides good long-channel devices, in which fluctuations in the gate-length are less important due to the smaller relative change, is seems difficult to reproducibly scale the definition of the

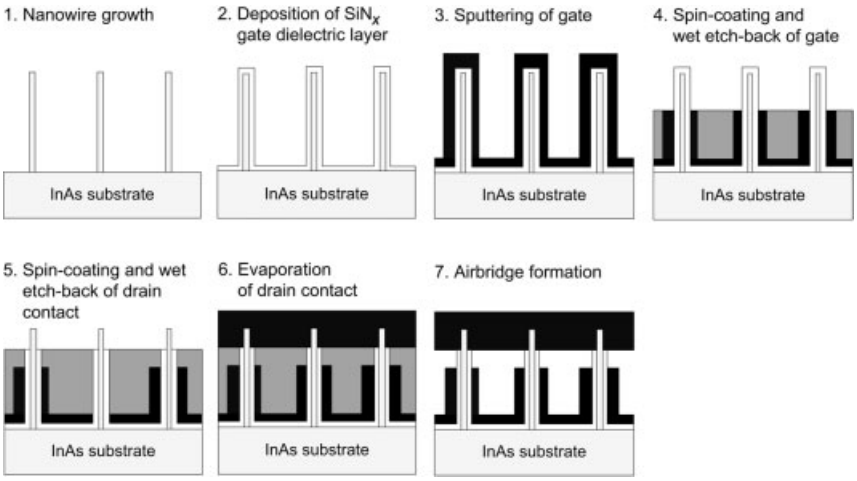


Figure 2.2 Processing scheme for long-channel transistors with sputtering and back-etching.

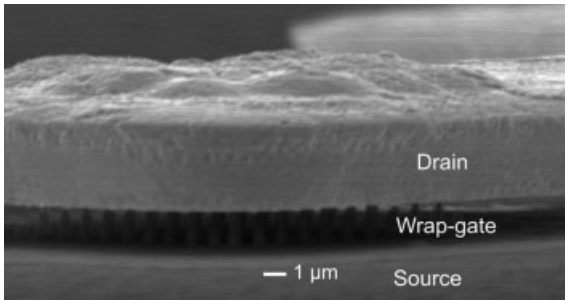


Figure 2.3 Scanning electron microscopy image of wrap-gate transistor with air-bridge drain contact [15].

gate-length below 100 nm when using this back-etch process. Instead, a direct evaporation method is used to form gates with a length below 100 nm, as described next.

For the processing of short-channel transistors, a direct evaporation of the gate metal has been developed. In this process, the metal gate is evaporated onto the SiN_x-covered nanowires, the main benefit of this approach being that the gate-length is determined by the thickness of the evaporated layer. This is in contrast to the previously described long-channel process, where it is set by the thickness of the back-etched polymer film. A scanning electron microscopy (SEM) image of a formed 80 nm-thick gate is shown in Figure 2.5. As can be seen in the image, an intimate contact is formed between the gate layer and the nanowire. Following gate formation, the drain contact is formed by spin-coating the sample with a resist and wet etching of

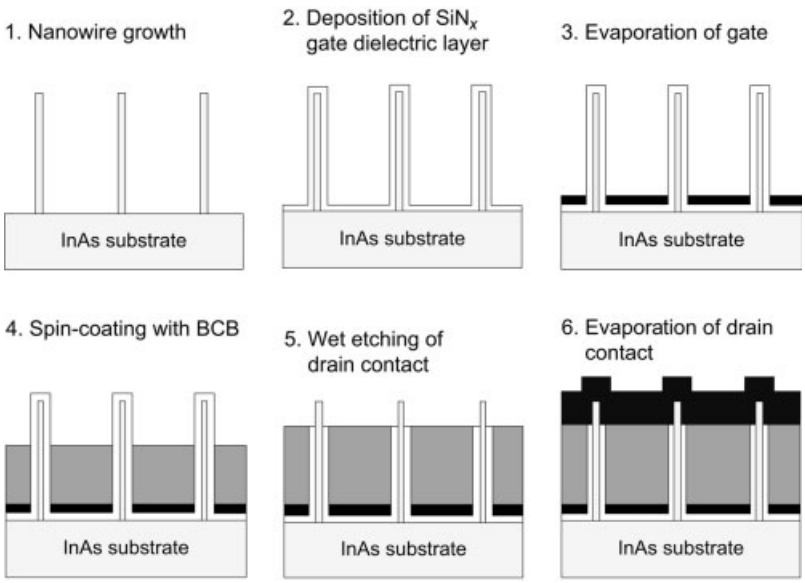


Figure 2.4 Processing scheme for short-channel transistors with evaporation of the gate [21]. BCB = benzocyclobutane.

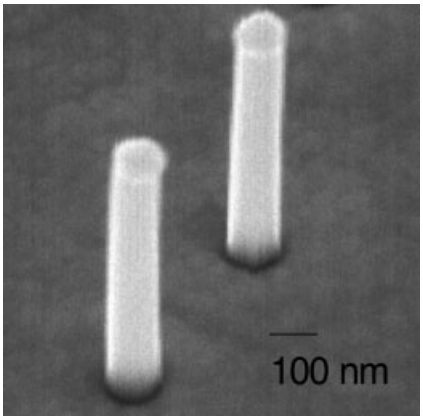


Figure 2.5 InAs nanowires coated with SiNx penetrating an evaporated Ti/Au gate electrode [15].

the tips of the nanowires that penetrate the organic film. Finally, an evaporated drain top contact is formed over the wires.

2.4 Long-Channel Transistors

The long-channel transistors have been characterized using the substrate as a grounded source contact. The data for a 10×10 nanowire matrix with a gate length of 800 nm, a wire diameter of 80 nm, and a thickness of 40 nm in the SiN_x-layer is shown in Figure 2.6. The transistor shows a good current saturation already at low drain biases $V_{sd} = 0.2\,V$ and a transconductance of about 1 mS. At larger drain biases ($V_{sd} > 1\,V$) the transistor shows an increase in the drain current, most likely related to impact ionization processes in the InAs channel. In order to explain the transistor operation, the transistor structure is modeled as a cylindrical version of the planar metal-insulator-semiconductor field-effect transistor (MISFET), as shown in Figure 2.7. In the MISFET, the gate potential is used to deplete carriers in the channel and thus to modulate the conducting area in the cross-sectional core of the nanowire. As the gate length is a factor ten-fold larger than the diameter of the wire and the thickness of the dielectric, a good potential control is obtained in the channel, and this is reflected in the measured low-output conductance of the transistor.

The long-channel transistors show the expected $I_{sd} \sim {V_g}^2$ dependence, as demonstrated for a 40-nanowire transistor in Figure 2.8. From these data the threshold voltage is deduced to be $V_t = -0.16\,V$. From the analytic fitting in Figure 2.8, values are deduced for the carrier concentration and the mobility ($N_d = 3 \times 10^{17}\,cm^{-3}$, $\mu = 9600\,cm^2\,V^{-1}\,s^{-1}$). The sub-threshold characteristics of the transistor are shown in Figure 2.9. At $V_{sd} = 0.2\,V$ an inverse sub-threshold slope of 100 mV decade^{-1} was measured, and a maximum current on-off ratio above 1000. To further verify the mode of operation in the transistor, the transistor characteristics were also simulated

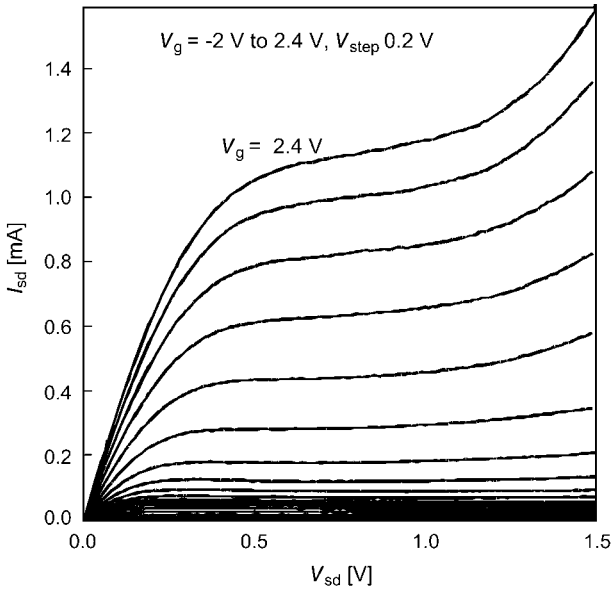


Figure 2.6 Measured *I–V* characteristics for an InAs long-channel NW-transistor [16].

using the Atlas device simulator created by Silvaco [16]. As these devices have a long channel length and a wide diameter, effects related to lateral quantization, doping fluctuations and ballistic transport may be omitted, and the transistors may be modeled within the drift-and-diffusion formalism. The simulated data in Figure 2.9 are obtained for $N_d = 2 \times 10^{17}\,\mathrm{cm}^{-3}$ and $\mu_e = 10\,000\,\mathrm{cm}^2\,\mathrm{V}^{1-1}\,\mathrm{s}^{-1}$, and reproduce the measured data both in the on-state and in the off-state. Thus, the measured data may be reproduced both by analytical modeling and by simulation in a MISFET model.

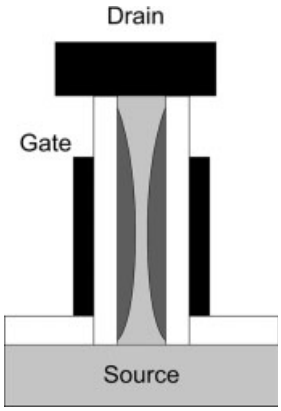


Figure 2.7 Schematic illustration of nanowire MISFET operation.

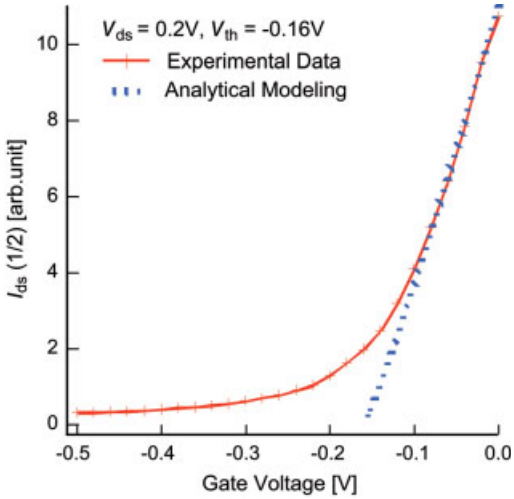


Figure 2.8 Measured transfer characteristics with analytical fitting to deduce the threshold voltage [16].

2.5 Short-Channel Transistors

Scaling is of importance to any FET-technology, and for the nanowire FET the scaling of both the gate length and nanowire diameter must be considered. The processing outlined above has been used the to fabricate transistors with 80 nm nominal gate length [19, 20]. During the growth of these nanowires, matrices with different nanowire diameters (70 and 55 nm) have been included on the same sample. Both types of nanowire transistor were processed in the same batch, and the transistor characteristics compared (see Figure 2.10). In both cases, good transistor characteristics were observed, with both transistors showing a limited current saturation, even

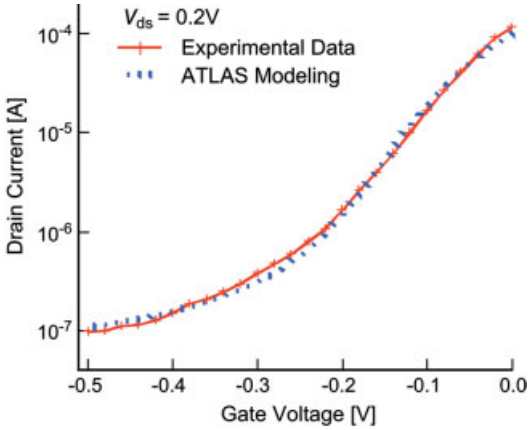


Figure 2.9 Measured and simulated sub-threshold characteristics for a long-channel transistor [15].

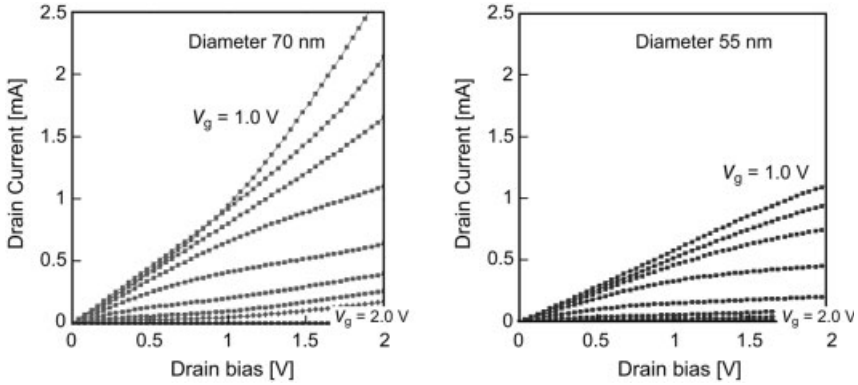


Figure 2.10 Measured transistor characteristics for 80 nm gate-length NW-transistor with 70 nm (left) and 55 nm (right) diameters [20].

at comparably large drain voltages. The increased saturation voltage arises from a series resistance in the 1 μm separation between the gate and the drain. For biases above 1 V, the larger-diameter transistors show a punch-through in the characteristics, a feature that was not observed in the narrower transistors that have a better potential control. When the drain current was normalized with the circumference of the nanowire, only a minor drive current reduction per gate width was observed as the diameter was reduced; this demonstrates good scalability in the technology.

In order to scale the gate length further, the relatively thick SiN_x dielectric was replaced with 10 nm HfO_2, a material with a higher dielectric constant (15) (Figure 2.11) [21, 22]. The HfO_2 was deposited using atomic layer deposition (ALD), which gives a uniform dielectric coverage and a very accurate thickness. A 100-nm layer of silicon oxide was also deposited, to act as a lower-k spacer layer between the InAs substrate and the wrap-gate. Next, a 50-nm Cr gate layer was formed by metal evaporation. Finally, a 100- to 200-nm-thick polymer layer was deposited on top of the gate to provide insulation between the gate and the drain contact. Despite a very short gate length (50 nm), considerably improved dc characteristics were observed compared to previous device designs. Transconductance values up to 0.8 S mm^{-1} were obtained ($V_{sd} = 1$ V), with an inversed sub-threshold slope around 100 mV dec^{-1}. The transconductance values were in this case normalized to the total nanowire circumference for the array. In addition to a gate swing of 0.5 V, an I_{on}/I_{off} ratio >1000 at $V_{sd} = 0.5$ V following the conventional definition [23] was observed, whereas a maximum I_{on}/I_{off} ratio above 10^4 was measured.

2.6 Heterostructure WIGFETs

The wrap-gate transistors show a good on-state characteristics, but even the long gate transistor characteristics suffer from a comparably large inverse sub-threshold slope (100 mV dec^{-1}) and a non-negligible off-state current. This is worse for the devices

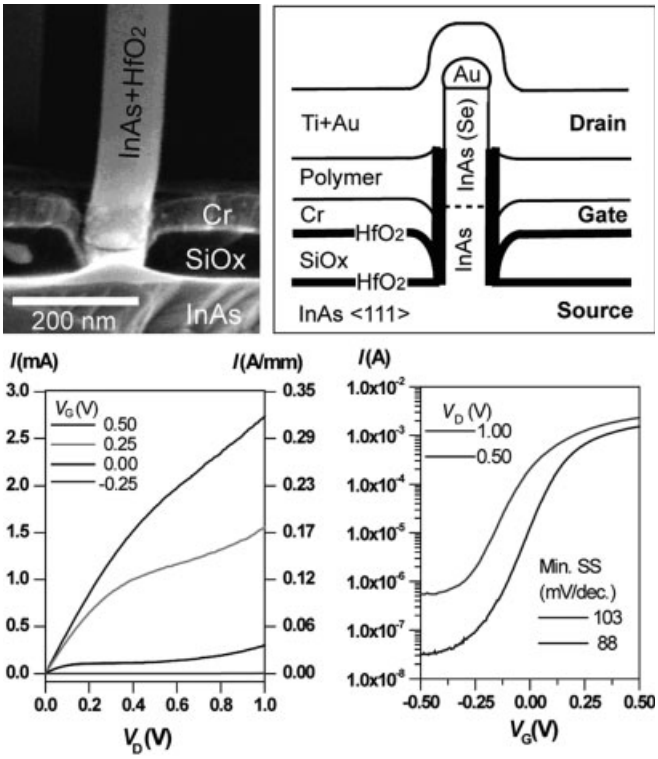


Figure 2.11 (Top left) A cross-section of a test sample showing the SiOx spacer layer, and the Cr wrap gate. (Top right) Illustration of the device design for an individual nanowire element in an array. (Bottom left) Output characteristics for a 61 nanowire wrap-gate array. (Bottom right): Sub-threshold characteristics for the same device for two different drive voltages [21].

with a short gate and a comparably thick (40 nm) gate-dielectrics. The transistors with high-*k* gate oxides also show effects related to the narrow InAs band gap that allows for impact ionization processes and thus creates a limited potential barrier in the off-state. The nanowire technology offers alternative transistor designs in that heterostructure segments may be incorporated into the transistor channel to alter the band gap in critical regions. A segment of InAsP was introduced into the InAs channel of a nanowire transistor and the role of the barrier in transistor performance subsequently investigated [24]. A 150 nm-long segment of InAsP was introduced into a 4 μm-long, 50 nm-diameter InAs nanowire grown by CBE. The nominal P content in the InAsP segment was 30%, and the conduction band barrier 180 meV. The nanowire was placed in a lateral geometry with a Si/SiO_2 back gate, where two drain contacts and one source contact was used in order to fabricate and evaluate transistors with the same geometry differing in only the InAsP barrier (Figure 2.12). Room-temperature data for the two types of transistor are shown graphically in Figure 2.13.

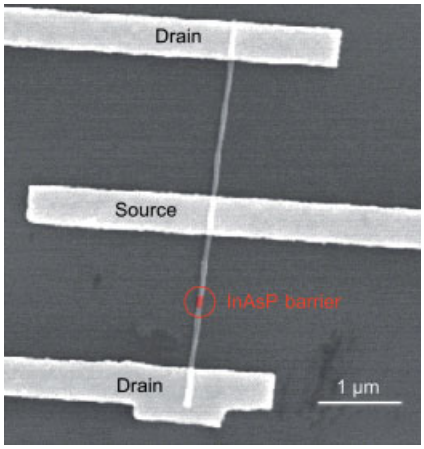


Figure 2.12 Scanning electron microscopy top image of lateral heterostructure nanowire transistor [23].

Both transistors showed good characteristics with current saturation and a decent drive current level. For a given bias condition ($V_{sd} = 0.3$ V and $V_g = 2.0$ V) the InAs transistor had a factor 2:1 higher drive current than the InAsP transistor. This was expected due to the introduction of the barrier. From the transfer characteristics, however, it should be noted that the current reduction was not related to a degradation in the transconductance, but rather to a shift in the threshold voltage. In fact, the measured transconductance remained constant, and for a fixed gate-overdrive the drive current was the same. When turning to the sub-threshold characteristics, major improvements were noted in both the inverse sub-threshold slope and the maximum I_{on}/I_{off} ratio as the barrier was introduced. Finally, temperature-dependent measurement of the current level was used to verify the presence of the barrier and to evaluate its height (Figure 2.14). The role of the barrier in this geometry is not only to block the off-current in the body of the wire, but also (and even more in this lateral geometry) to block the leakage current along the edges of the wires.

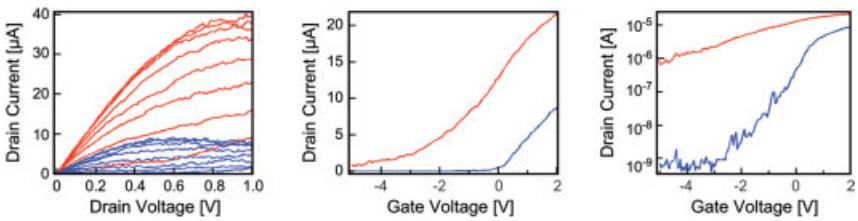


Figure 2.13 Measured transistor characteristics for lateral InAs/InAsP NW transistor (left) with transfer characteristics (middle) and sub-threshold characteristics (right) [23].

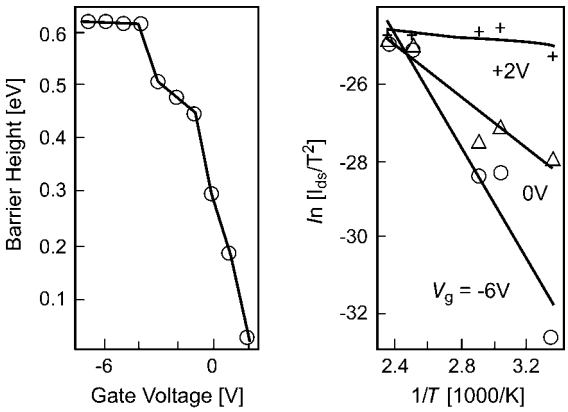


Figure 2.14 Deduced activation energies for varying gate bias (left) and the corresponding Arrhenius plot (right).

2.7 Benchmarking

It is of great value to perform an early evaluation of the potential in this new wrap-gate transistor technology. Hence, the performance of 100 nm gate-length transistors (structure shown in Figure 2.15) has been simulated and the characteristics evaluated according to the metrics of high-performance logic devices, including the gate delay

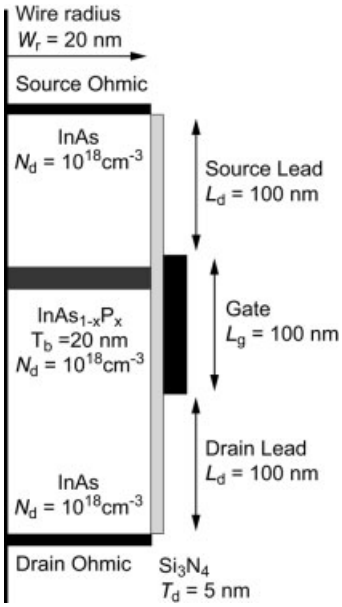


Figure 2.15 Schematic of nanowire geometry used for the bench-marking [25].

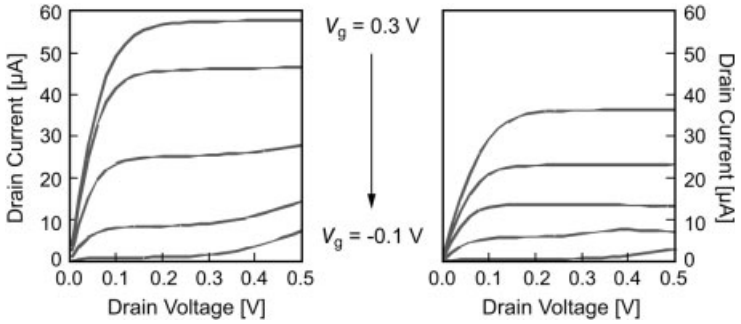


Figure 2.16 Simulated *I–V* characteristics for InAs nanowire wrap-gate transistor (left) and InAs/InAsP nanowire wrap-gate transistor (right) [25].

($\tau = C_{gg} V_{ds}/I_{on}$), the energy-delay-product ($EDP = \tau C_{gg} V_{ds}^2$), the current I_{on}/I_{off} ratio, and the inverse sub-threshold slope [25]. InAsP barriers with different P contents were further introduced into the InAs channel as a way of reducing the off-current, in analogy with the lateral devices previously described. In these simulations, the wire diameter was set to 40 nm and the doping level fixed to $1.0 \times 10^{18}\,cm^{-3}$ in order to obtain usable threshold voltages around $V_t = 0\,V$ and to avoid parasitic access resistance in the source and drain leads. The mobility was set to $10\,000\,cm^2\,V^{-1}\,s^{-1}$ and v_{sat} to $3 \times 10^5\,m\,s^{-1}$, in accordance with the fitting in Figure 2.9. The simulated output characteristics are shown in Figure 2.16 for a pure InAs channel and for a channel with a barrier of 20% P, respectively. Both types of transistor showed a good saturation already at about $V_{sd} = 0.1\,V$, which was a reflection of the low access resistance and the possibility of achieving excellent ohmic contacts to InAs. Obviously, the drive current for fixed gate bias was reduced by a factor 1.5 as the 20% barrier was introduced.

By evaluation of the transfer characteristics (Figure 2.17), this reduction in the drive current was found to be mainly related to a positive shift of the threshold

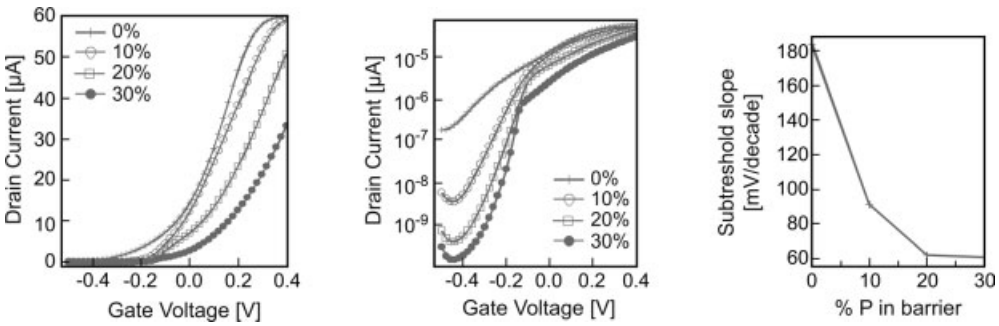


Figure 2.17 Simulated transfer characteristics (left), sub-threshold characteristics, and deduced inverse sub-threshold slope (right) for InAsP nanowire transistors with varying P content [24].

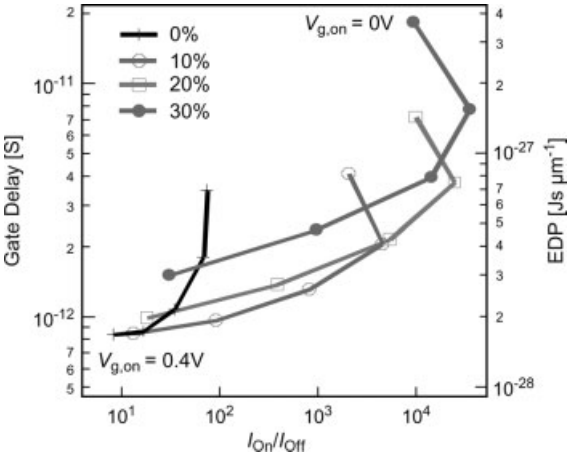


Figure 2.18 Deduced gate delay and energy–delay product as function of current on/off ratio for InAsP nanowire transistors with different P content and at varying bias conditions [25].

voltage, whilst there was only a minor degrading in the transconductance. On a logarithmic scale, it should be noted not only that the barrier provides a way to shift the threshold voltage, but also that the transition from the exponential to the linear characteristics becomes sharper, the current on/off ratio increases, and the inverse sub-threshold slope is reduced. All of these factors reduce the power consumption in circuits. The deduced critical metrics – that is, the gate delay, the energy delay product, the current on/off ratio, and the inverse sub-threshold swing – are shown in Figure 2.17 for various P contents in the barrier. As the threshold voltage shift with the barrier height, the evaluation is performed for various $V_{g,on}$, while the gate swing is kept constant at 0.5 V for all cases. The main point in Figure 2.18 is that the pure InAs channel provides the shortest gate delay due to the largest drive current, but also the lowest current I_{on}/I_{off} ratio related to the narrow band gap. While the introduction of the P barrier into the channel increases the minimum gate delay, it provides a way of increasing the I_{on}/I_{off} ratio to above 10^3 even for a gate swing of 0.5 V, a value required for many circuit applications. As an added benefit, the sharper transition between the on- and off-states further reduces the sensitivity to the choice of gate bias. These simulations show the potential in the technology and also demonstrate the benefit of introducing heterostructures into the nanowires.

2.8 Outlook

Based on the experimental results obtained to date, the question might be asked as to how far the nanowire FET technology may be developed? Critical issues for scaled devices are related to the growth of narrow nanowires with diameters of 10 to 30 nm and the processing of vertical gates on the 20 nm length scale. Based on experimental

results, devices processed on these dimensions seem feasible in the near future. However, in order for these devices to be competitive it will be necessary for the drive current to be increased and the parasitics reduced. Likewise, good control of the carrier concentration in the channel and in the source and drain regions will be needed, as will an understanding and control of the interface properties in capped wires. The main benefit of the wire geometry – the possibility for heterostructure design in the axial and radial directions – may well prove to be the key when addressing these issues.

Acknowledgments

These studies were conducted within the Nanometer Structure Consortium at Lund University, with financial support from the Swedish Research Council (V.R.), the Swedish Foundation for Strategic Research (S.S.F.), the Knut and Alice Wallenberg Foundation (K.A.W.), and from the European Union via the project NODE 015783.

References

1 C. P. Auth, J. D. Plummer, Scaling theory for cylindrical, fully-depleted, surrounding-gate MOSFETs, *IEEE Electron Device Lett.* 1997,**18** (2), 74–76.

2 H. Takato, K. Sunouchi, N. Okabe, A. Nitayama, K. Hieda, F. Horiguchi, F. Masuoka, Impact of Surrounding Gate Transistor (SGT) for ultra-high-density LSIs, *IEEE Trans. Electron. Dev.* 1991, **38** (3), 573.

3 S. D. Suk, S.-Y. Lee, S.-M. Kim, E.-J. Yoon, M.-S. Kim, M. Li, C. W. Oh, K. H. Yeo, S. H. Kim, D.-S. Shin, K.-H. Lee, H. S. Park, J. N. Han, C. J. Park, J.-B. Park, D.-W. Kim, D. Park, B.-I. Ryu, High performance 5 nm radius twin silicon nanowire MOSFET (TSNWFET): fabrication on bulk Si wafer, characteristics, and reliability, *Int. Electron Devices Meeting Tech. Dig.* 2005, 735–738.

4 S. C. Rustagi, N. Singh, W. W. Fang, K. D. Buddharaju, S. R. Omampuliyur, S. H. G. Teo, C. H. Tung, G. Q. Lo, N. Balasubramanian, D. L. Kwong, CMOS inverter based on gate-all-around silicon-nanowire MOSFETs fabricated using top-down approach, *IEEE Electron Device Lett.* 2007, **28** (11), 1021–1024.

5 X. C. Jiang, Q. H. Xiong, S. Nam, *et al.*, InAs/InP radial nanowire heterostructures as high electron mobility devices, *Nano Lett.* 2007, **7** (10), 3214–3218.

6 J. Xiang, W. Lu, Y. J. Hu, Y. Wu, H. Yan, C. M. Lieber, Ge/Si nanowire heterostructures as high-performance field-effect transistors, *Nature* 2006, **441**, 489–493.

7 Y. Li, J. Xiang, F. Qian, *et al.*, Dopant-free GaN/AlN/AlGaN radial nanowire heterostructures as high electron mobility transistors, *Nano Lett.* 2006, **6** (7), 1468–1473.

8 P. Mohan, J. Motohisa, T. Fukui, Fabrication of InP/InAs/InP core-multishell heterostructure nanowires by selective area metal organic vapor phase epitaxy, *Appl. Phys. Lett.* 2006, **88**, 133105.

9 C. Thelander, P. Agarwal, S. Brongersma, *et al.*, Nanowire-based one-dimensional electronics, *Mater. Today* 2006, **9** (10), 28–35.

10 A. I. Persson, M. T. Björk, S. Jeppesen, L. Samuelson, J. B. Wagner, L. R. Wallenberg,

$InAs_{1.5-.5x}P_x$ nanowires for device engineering, *Nano Lett.* 2006, **6**, 403.

11 T. Mårtensson, J. B. Wagner, E. Hilner, A. Mikkelsen, C. Thelander, J. Stangl, B. J. Ohlsson, A. Gustafsson, E. Lundgren, L. Samuelson, W. Seifert, Epitaxial growth of indium arsenide nanowires on silicon using nucleation templates formed by self-assembled organic coatings, *Adv. Mater.* 2007, **19** (14), 1801–1806.

12 M. T. Björk, O. Hayden, H. Schmid, *et al.*, Vertical surround-gated silicon nanowire impact ionization field-effect transistors, *Appl. Physics Lett.* 2007, **90** (14), 142110.

13 O. Hayden, M. T. Björk, H. Schmid, *et al.*, Fully depleted nanowire field-effect transistor in inversion mode, *Small* 2007, **3** (2), 230–234.

14 H. T. Ng, J. Han, T. Yamada, P. Nguyen, Y. P. Chen, M. Meyyappan, Single crystal nanowire vertical surround-gate field-effect transistor, *Nano Lett.* 2004, **4** (7), 1247–1252.

15 T. Bryllert, L. Samuelson, L. E. Jensen, L.-E. Wernersson, Vertical high mobility wrap-gated InAs nanowire transistors, in: Proceedings, 63rd Device Research Conference, Santa Barbara, CA, USA, 2005.

16 L.-E. Wernersson, T. Bryllert, E. Lind, L. Samuelson, Wrap-gated InAs nanowire, in: Proceedings, Field Effect Transistor 2005 International Electron Device Meeting, December 5–7, IEDM Technical Digest, Washington DC, USA, pp. 265–268, 2005.

17 T. Bryllert, L.-E. Wernersson, L. E. Froberg, L. Samuelson, Vertical high-mobility wrap-gated InAs nanowire transistor, *IEEE Electron Device Lett.* 2006, **27** (5), 323–325.

18 T. Bryllert, L.-E. Wernersson, T. Löwgren, L. Samuelson, Vertical wrap-gated nanowire transistors, *Nanotechnology* 2006, **17** (11), 227–230.

19 L.-E. Wernersson, E. Lind, L. Samuelson, T. Löwgren, J. Ohlsson, Nanowire field-effect transistor, *Jap. J. Appl. Phys.* 2007, **46** (4B), 2629–2631.

20 T. Löwgren, J. Ohlsson, L. Samuelson, L.-E. Wernersson, Control of threshold voltage in 80 nm gate length InAs vertical nanowire WIGFETs, *Device Research Conference Tech. Digest* 2007, 165–166.

21 C. Thelander, L. E. Fröberg, C. Rehnstedt, L. Samuelson, L.-E. Wernersson, Vertical enhancement-mode InAs nanowire field-effect transistor with 50 nm wrap-gate, *IEEE Electron Device Lett.* 2008, **29**, 206–208.

22 C. Rehnstedt, C. Thelander, L. E. Fröberg, B. J. Ohlsson, L. Samuelson, L.-E. Wernersson, Drive current and threshold voltage control in vertical InAs wrap-gate transistors, *Electron Lett.* (accepted for publication) 2008, **44**, 236–237.

23 R. Chau, S. Datta, M. Doczy, B. Doyle, B. Jin, J. Kavalieros, A. Majumdar, M. Metz, M. Radosavljevic, Benchmarking nanotechnology for high-performance and low-power logic transistor applications, *IEEE Trans. Nanotechnol.* 2005, **4** (2), 153–158.

24 E. Lind, A. I. Persson, L. Samuelson, L.-E. Wernersson, Improved subthreshold slope in an InAs nanowire heterostructure field-effect transistor, *Nano Lett.* 2006, **6** (9), 1842–1846.

25 E. Lind, L.-E. Wernersson, InAsP/InAs nanowire heterostructure field effect transistors, *Device Res. Conf. Tech. Digest* 2006, 173–174.

3
Single-Electron Transistor and its Logic Application

Yukinori Ono, Hiroshi Inokawa, Yasuo Takahashi, Katsuhiko Nishiguchi, and Akira Fujiwara

3.1
Introduction

Complementary metal-oxide-semiconductor (CMOS) technology will face significant technological limitations shortly after 2010 [1], and intensive studies are currently being conducted in computational architecture, circuit design, and device fabrication to find ways to overcome this impending crisis. The major problem, especially for logic large-scale integrated circuits (LSIs), is that rapidly increasing power dissipation due to ever larger numbers of transistors and higher levels of interconnections is pushing CMOS circuits beyond their cooling limit. This points to the need for some drastic change in how LSIs are built, either at the system architecture or base device level, or both. Roughly speaking, achieving low-power operation of LSIs requires that both the total capacitance of circuits and the operation voltage are reduced, which means in turn that the number of electrons participating in the operation of some unit instruction must also be reduced. Single-electron transistors (SETs) [2–5], the characteristics of which are literally governed by the movement of single electrons, are considered to be the devices that will allow such a change. Their operation is basically guaranteed even when device size is reduced to the molecular level. Their performance, such as the peak-to-valley current ratio, improves as they become smaller. These properties are quite beneficial for large-scale integration. In addition, SETs are able not only to operate as simple switches but also to have high functionality. Many theoretical studies have been conducted to evaluate the possibility of building SET-based LSIs, and fundamental computational capability has already been proved.

In this chapter, after a brief explanation of SET operation principles and fabrication processes, some experimentally tested SET logic circuits will be introduced. In addition, the merits and demerits of the SET as a logic device will also be discussed, and some brief ideas proposed concerning SET logic circuits.

Nanotechnology. Volume 4: Information Technology II. Edited by Rainer Waser

ISBN: 978-3-527-31737-0

3.2 SET Operation Principle

Before considering the SEToperation principle, imagine a small conductive sphere or "island" floating over the ground. If one electron is taken from the ground and placed in the sphere, then there will be an increase in the electrostatic potential of the sphere. This is given by e/C, where C is the capacitance of the sphere to the ground and e is the elementary charge, 1.6×10^{-19} C. When the sphere – and hence C – is extremely small, the potential increase becomes significant. For example, for a nanometer-scale sphere having a capacitance C of say 1 aF (1×10^{-18} F), the increase in the potential e/C reaches 160 mV. This is much larger than the thermal noise voltage at room temperature, 25.9 mV. The potential increase prevents another electron from entering the sphere unless that electron has an energy larger than e^2/C. This phenomenon is called the *Coulomb blockade*. If the potential can be decreased by e/C, by applying an external bias, then a second electron can (but the third one cannot) enter the sphere. If this occurs by quantum-mechanical tunneling, then it is called *single-electron tunneling*. Any single-electron device, including the SET, has at least one small conductive island and its operation relies on the Coulomb blockade and single-electron tunneling.

The cross-sectional view and equivalent circuit of the SET is shown in Figure 3.1. The SET is, like a conventional transistor, a three-terminal device consisting of a source, drain, and gate. There is however an additional component, called the "Coulomb island", between the source and drain. The Coulomb island is also called

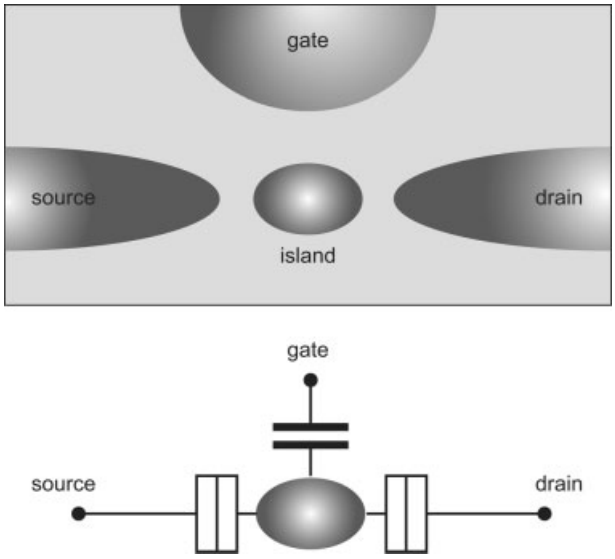


Figure 3.1 Cross-sectional view and equivalent circuit of the SET. In the equivalent circuit, double boxes indicate the tunnel junctions.

simply the island or a quantum dot. The Coulomb island must be conductive so that electrons can travel from the source to the drain via it. The role of this island is to capture/donate one electron from/to the source/drain, and otherwise to hold captured electrons. The region between the island and source (and also drain) must not be a good conductor; it must basically be insulating, so that electrons move to/from the island only by the tunneling. This region is called the *tunnel capacitor* or the *tunnel junction*. One the other hand, the region between the island and the gate should be insulating so as not to allow electrons to flow between them, as in a conventional transistor. In the equivalent circuit, the double box symbolizes a tunnel capacitor, which is a special capacitor that allows quantum mechanical tunneling of electrons, as mentioned above. The region sandwiched by the tunnel capacitors corresponds to the island, which is designated by an oval for visualization. The region between the island and the gate can be expressed as a normal capacitor.

Figure 3.2 explains what happens when the gate voltage is varied with a fixed small source/drain voltage. When a positive voltage V_g is applied to the gate, positive charges are induced there, whose number is given by C_gV_g/e, where C_g is the gate capacitance. Then, in order to minimize the free energy of the system, the SET tries to induce the same number of negative charges (i.e. electrons) in the island, and these electrons are conveyed from the source or drain through the tunnel junctions. If C_gV_g/e is some integer N, the island obtains N electrons. After reaching this number, no more electron movement occurs. This is the *Coulomb blockade state* (the left equivalent circuit in Figure 3.2). When C_gV_g/e is not an integer, for example, a half integer $N + 1/2$, the number of electrons in the island changes with time so that it becomes $N + 1/2$ *on average*. What actually occurs in the SET is as follows. First, one electron enters the island from the source and the number of electrons in the island

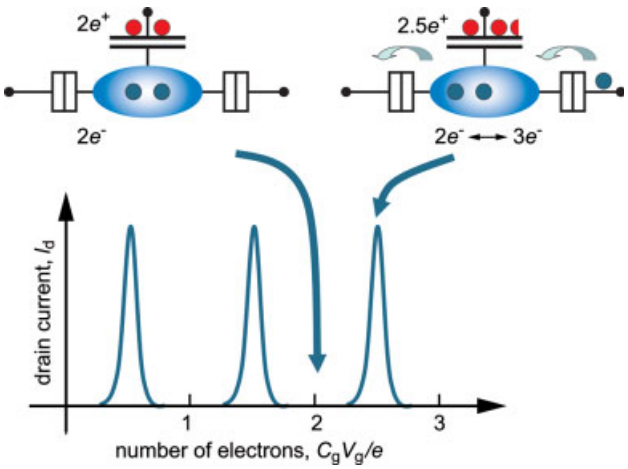


Figure 3.2 Drain current I_d as a function of number of charges C_gV_g/e induced in the gate. C_g and V_g are the capacitance and gate voltage, respectively. The equivalent circuits explain the Coulomb-blockade state (left) and single-electron-tunneling state (right).

becomes $N + 1$. Next, one electron emits to the drain from the island, resulting in N. This one-by-one electron transfer is repeated so that there is a net current between the source and drain. This is the *single-electron-tunneling state* (the right equivalent circuit in Figure 3.2). As a result, when the gate voltage is swept, the Coulomb-blockade state and the single-electron-tunneling state appears by turn, and the drain current versus gate voltage characteristics exhibit a repetition of sharp peaks, as shown in Figure 3.2. This is known as *Coulomb-blockade oscillation*. This ON-OFF characteristic indicates that the SET can function as a switching device.

For the complete description of the SET operation, consider the stability chart in the gate-voltage/drain-voltage plane in Figure 3.3. The rhombic-shaped regions colored in red are the region for the Coulomb-blockade state, and are known as *Coulomb diamonds*. Outside the Coulomb diamonds, the number of electrons in the island fluctuates between certain numbers. The degree of the fluctuation is determined by how far the voltage conditions are from the Coulomb diamonds. In the blue regions, the fluctuation is minimum – that is, the electron number changes only between two adjacent integers. These regions are for the single-electron-tunneling states. The shape and size of the Coulomb diamonds are determined only by the gate and junction capacitances. For example, the maximum drain voltage for the Coulomb blockade is given by e/C_Σ, where $C_\Sigma = C_g + C_d + C_s$ is the total capacitance of the island and C_d and C_s are junction capacitances at the drain and source. Each slope of the diamond is given by $-C_g/C_d$ and $C_g/(C_g + C_s)$.

A more detailed explanation of the SET operation can be found in textbooks [6, 7] and review articles [8, 9]. At this point, mention should be made of only one more item – which is what the SET *cannot* do. As explained above, the SET can convey electrons one by one, but the time interval of each transfer event is uncontrollable.

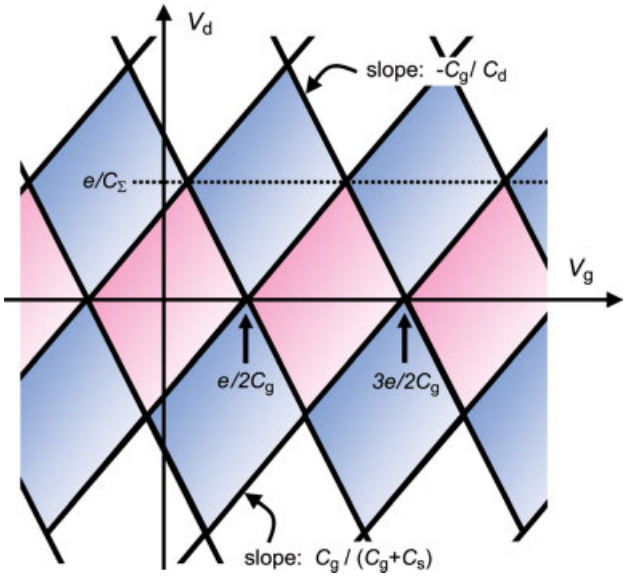


Figure 3.3 Stability chart of the SET in the gate/drain voltage plane.

This is because the transfer relies on tunneling, which is inherently stochastic. Therefore, it is difficult for the SET to transfer just one electron while preventing a second electron from being transferred. In other words, the transfer accuracy is quite low in the SET. In order to overcome this drawback, new single-electron devices have been invented and experimentally demonstrated. Sometimes called *single-charge-transfer devices*, these include the single-electron turnstile [10] and single-electron pump [11]. Although their structure is somewhat complicated (some of them possess two or more islands), they can transfer just one electron in one cycle of the gate clock, thereby providing high transfer accuracy. This function – the clocked single-electron transfer – is quite beneficial for implementing a certain level of logic architecture where one electron represents one bit. In this chapter, however, attention will be focused on the SET, and the single-charge-transfer devices and related logic styles will not be described in any detail. Very few experimental studies have been conducted on the single-charge-transfer logic circuits because of the difficulty of their fabrication. Hence, for single-charge-transfer logic, the reader is referred to review articles [8, 12].

3.3 SET Fabrication

When SETs are fabricated, two criteria should be borne in mind. First, the resistance of the tunnel junction must be sufficiently larger than the quantum resistance $R_q = h/e^2$ ($\sim$25.8 kΩ), where h is the Planck constant. Otherwise, the number of electrons in the island fluctuates because of the Heisenberg uncertainty principle. Because of this requirement, the current drivability is low, which is one major demerit of the SET as a logic element. Second, the energy for adding one electron to the island must be larger than the thermal energy. Otherwise, heated electrons tunnel through the barriers and the Coulomb blockade does not function. For example, as the temperature rises, each peak in Figure 3.2 broadens, and finally smears out. This relationship is expressed as $E \gg kT$, where E is the addition energy. The addition energy can be expressed in the form e^2/C_Σ, and thus a SET with a smaller island can operate at a higher temperature.

When the de Broglie wavelength of electrons is much smaller than the island size – which is the case of metal islands that are not too small ($\gg$1 nm) – charges are induced right at the island surface and E is determined only by the island size and the spatial configuration of the electrodes. However, when the de Broglie wavelength is comparable to the island size – typically as in the case of semiconductor islands with a nanometer size – the quantum size effect causes the kinetic energy of an electron to increase and hence E to increase. Thus, semiconductor islands can have larger addition energy than a metal island of the same size.

In order to explain how small the island must be made, Figure 3.4 shows the relationship between the island size (dot size) and the addition energy for a Si island embedded in SiO_2 dielectrics [13]. Both, the quantized level spacing and the charging energy (which can be defined as the addition energy for ideal metals) increase as the

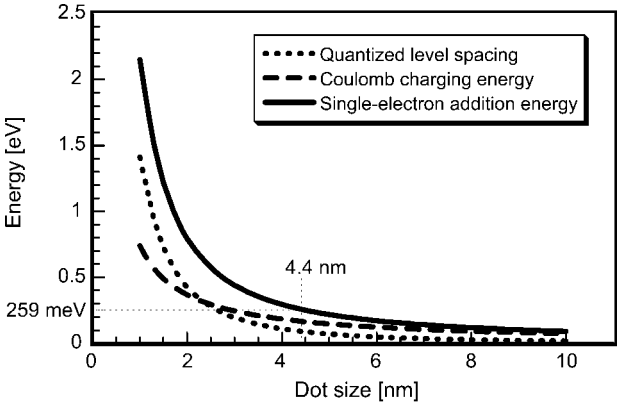


Figure 3.4 Relationship between dot size and addition energy [13].

island size decreases, and this leads to an increase in the resultant addition energy. If an addition energy 10 times larger than the room-temperature energy (25.9 meV) is required for the proper operation of a SET circuit, then an island as small as 4 nm is needed. The creation of such a small island and attaching tunnel junctions to it represents a technological challenge in SET fabrication. However, any conducting material can be used as long as the above criteria are satisfied, and an addition energy much larger than the room-temperature energy has already been demonstrated.

Historically, research into single-electron devices began with metals [4] and then expanded to semiconductors [14–18] and other materials, such as carbon nanotubes [19–22] and some molecules [23–28].

Among metals, a major material is aluminum, because its oxide functions as a good dielectric for tunnel junctions. Tunnel junctions are commonly made using Dolan's shadow evaporation technique [29]. The junction capacitance can be controlled in such structures to within 10% if the Al–AlO_x junctions have a relatively large capacitance of several hundred attofarads. Making smaller junctions is less easy, and thus electrical measurements with this material are commonly carried out below 1 K. However, it has been shown that making an extremely small SET, the island of which has an addition energy much larger than kT of room temperature, is possible [30].

Among semiconductors, Si is the most widely used material in research aimed at practical applications. Si SETs are commonly made on a certain type of Si substrate called silicon-on-insulator (SOI) [31]. In SOI substrate, a thin Si layer (typically 100–400 nm) is formed on a buried SiO_2 layer. Thinning the Si layer and reducing its size in the lateral direction by lithography enables small Si structures to be produced. It is possible to further miniaturize these structures by using thermal oxidation: Si is consumed during the oxidation, and thus the volume of the Si structures is reduced. With Si, it is relatively easy to make smaller islands compared with Al. Common measurement temperatures in Si SET research are 1 K to 100 K, and several groups have observed the Coulomb-blockade oscillation at higher temperatures (100–300 K) [13, 32–48].

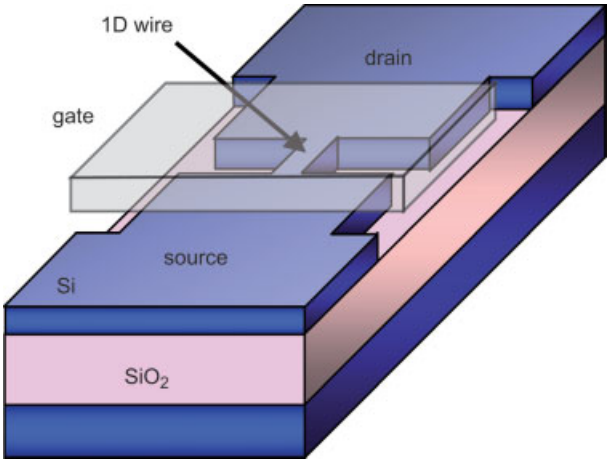


Figure 3.5 Basic device structure of PADOX SETs. The 1-D wire is converted to a Coulomb island and attaching tunnel junctions after thermal oxidation of the Si.

One reliable way of fabricating Si SETs is pattern-dependent oxidation, or PADOX [34, 49], and this has enabled the fabrication of a room-temperature-operating SET for the first time. It has been shown that the gate and junction capacitances are controllable even when their values are very small (a few attofarads) [34, 49, 50]. The PADOX method requires no special material for tunnel junctions as they are made of Si itself. PADOX exploits an oxidation-induced band modification [51], which makes it possible to produce a Coulomb island and tunnel junctions simultaneously during the gate oxidation step [52]. Figure 3.5 shows the basic structure for the PADOX SET. A one-dimensional (1-D) Si wire is converted to an island and tunnel junctions after thermal oxidation. As the name indicates, the final Si structure is dependent on the initial structures before oxidation. By elaborately designing the initial structures, a variety of SET configurations become possible [49, 53]. Figure 3.6 shows the drain current versus gate voltage characteristics of a PADOX SET measured at 27 K. Clear oscillation is observed. The PADOX method has contributed to the fabrication of many experimental SET circuits owing to its high controllability and the stability of the current characteristics.

Recent progress in the SET fabrication process has resulted in a very clear Coulomb blockade oscillation at room temperature; an example of this is shown in Figure 3.7, where a peak-to-valley ratio as large as 400 is achieved [48]. Although the working mechanisms underlining this excellent performance at room temperature are not satisfactorily understood at present, it is evident that room-temperature-operating SETs can be made. The focus of Si SET research is therefore moving from how to make room-temperature-operating SETs to how to control their size, which is still very difficult.

Carbon nanotubes are attractive for attaining small capacitance and thus high operation temperature because they have extremely small diameters of the order of a

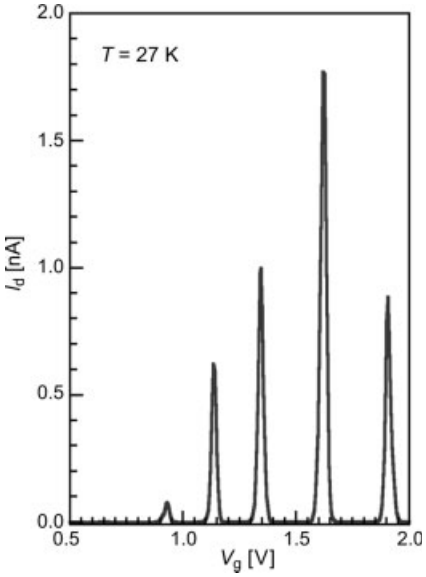


Figure 3.6 Current characteristics of a PADOX SET, measured at 27 K.

few nanometers. In the case of a 1-µm-long single-wall nanotube with a diameter of 1.4 nm suspended 100 nm above a ground plane, the addition energy E would be 8 meV, and this could be further increased by reducing the length. In fact, carbon nanotube SETs with E corresponding to this estimate have been reported [19, 20]. For practical applications, the nanotube diameter, chirality (i.e. electronic structure) and the locations of the tunnel junctions and nanotube itself must be controlled more precisely. Although these issues have already been partly addressed [21, 22], much further improvement is needed.

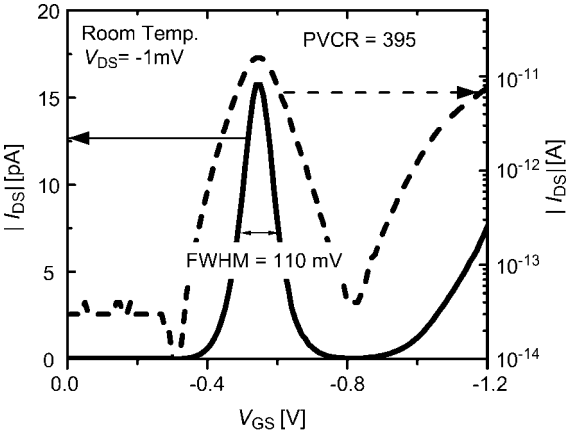


Figure 3.7 Current characteristics of a room-temperature-operating Si SET [48].

Fabrication methods that use molecules as building blocks are anticipated. In such methods, functions and characteristics are determined by chemical synthesis, without relying on lithographic techniques. Research into the charging effect or Coulomb blockade in a molecule began during the mid-1990s [23, 24], and more recently persuasive data showing conductance modulation by gate potential have been obtained [25–28]. Although the present understanding of transport in a molecule is improving, many issues of circuit integration, including architectural design, synthesis, and interfacing with external circuits, remain.

At this point, mention should be made of an infamous problem in SETs, known as the *background charge problem* [54]. Due to randomly distributed mobile and immobile charges in the dielectrics, the device characteristics may change over time and differ from one device to another. This is because SETs have a high sensitivity to charges due to their small size, and it makes the integration of SETs difficult. A typical case is seen in SETs made from metals and GaAs/AlGaAs heterostructures. For example, the characteristics of SEDs with Al–AlOx junctions change at least once a day. In order to stabilize such behavior, it may be necessary to wait for a long time after cooling down before measurements can be made [55]. The situation is similar for carbon nanotubes and some molecules, as these suffer from a large noise superimposed on the current characteristics, the origin of which is unknown.

The background charge problem is not specific to SETs, however, and may occur in any nanoscale *field-effect* device due to their high sensitivity to charges. In addition, the amount, location, and stability of the background charges are highly material- and process-dependent. In fact, it has already been shown that PADOX SETs have excellent long-term stability. The drift of the characteristics is less than $0.01e$ over a period of a week at cryogenic temperatures [56]. More practically, no noticeable change in the characteristics have been observed for more than eight years, during which time thermal cycling between room temperature and $\sim$20 K has occurred several times [57]. It has also been shown that the voltage at which the first Coulomb-blockade-oscillation peak appears is controllable [58]. These results demonstrate that PADOX SETs are not significantly influenced by slowly moving or immovable background charges, which indicates that the background charges problem is not intrinsic but rather can be solved. At present, no clear answers have been identified as to how seriously fewer fixed charges and/or faster motion of charges, which causes 1/f noise, will obstruct integration. However, it is believed that a circuit design with some degree of defect tolerance would relax the effects.

In summary, for room temperature operation, an island smaller than 10 nm is necessary. At present, Si is preferable for the SET fabrication from the viewpoints of operation stability and temperature. Some experimental data are available showing the control of the peak positions and the peak intervals in current characteristics. However, these parameters are still difficult to control in room-temperature-operating SETs. Also, there are no data showing the control of the resistance of room-temperature-operating SETs, and these points remain the subjects of future studies. A more complete description of the fabrication process for Si SETs can be found in Ref. [59].

3.4 Single-Electron Logic

Many logic styles have been proposed and analyzed for single-electron devices. Most of them can be categorized into two groups: charge-state logic; and voltage-state logic.

Charge-state logic [8, 12], which uses one electron to represent one bit, is highly specific to single-electron devices and might be in some sense the ultimate logic. Devices other than SETs, such as single-electron pumps, are building blocks. However, very few experimental studies have been reported regarding circuit operation based on this scheme because of the difficulty of the fabrication.

Voltage-state logic [60–88], which will be described here in detail, uses the SET as a substitute for the conventional MOS field-effect transistor (FET); hence it is referred to as SET logic. Although the circuit characteristics are predominated by the Coulomb blockade and single-electron tunneling, these phenomena are not directly employed for computation. Instead, the current produced by the sequence of single-electron tunnelings is used, and the bit is represented by the voltage generated by the accumulation of plural electrons. Actually, this is not a genuine single-electron logic because 10^1 to 10^3 electrons will be used in the operation. In many aspects, this logic is analogous to CMOS logic. The major advantage is that the accumulated technologies can be employed for CMOS circuit designs. However, the logic is not merely a copy of CMOS logic because the SET has completely different current characteristics from the MOSFET; that is, the current oscillates as a function of gate voltage. Some important elemental circuits such as an inverter, an exclusive-or (XOR) gate, a partial-sum/carry-out circuit, and an analog-to-digital converter, have been experimentally demonstrated [43, 46, 89–102]. Some of these will be introduced in the following three subsections.

3.4.1 Basic SET Logic

The voltage gain of the SET can be defined as for conventional transistors [5]. It is known from Figure 3.3 that when the drain voltage increases with a fixed gate voltage, a current begins to flow at the edge of the Coulomb diamond. As a result, the output drain voltage V_d for a fixed input drain current I_d exhibits a Coulomb diamond as a function of the gate voltage V_g. The measured characteristics for a Si SET are shown in Figure 3.8 (upper panel). The two slopes in the figure correspond to the inverting and non-inverting voltage gains G_I and G_{NI}. As shown in Figure 3.3, their values are determined by the capacitances as

$$G_I = C_g/C_d \tag{3.1}$$

$$G_{NI} = C_g/(C_g + C_s) \tag{3.2}$$

Although G_{NI} is always smaller than unity, G_I exceeds unity if $C_g > C_d$. Therefore, CMOS-like logic circuits can be prepared using SETs as substitutes for MOSFETs. From Eq. (3.1) it is clear that the SET must have a large C_g in order to obtain a high

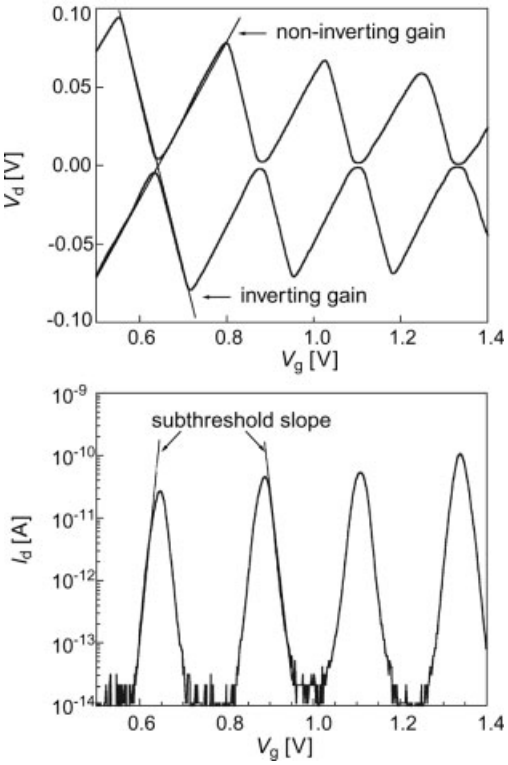


Figure 3.8 Electrical characteristics of a PADOX SET. Upper panel: Drain voltage V_d as a function of gate voltage V_g for a fixed drain current of ± 10 pA. Lower panel: Drain current I_d as a function of V_g for a fixed drain voltage of 10 mV. The measurement temperature was 27 K [50].

inverting gain, which in turn means that the total capacitance of the SET island will tend to increase. Therefore, the voltage gain and operating temperature are in a trade-off relationship and it is not easy to produce SETs with a larger-than-unity gain that operate at high temperatures. A G_I value larger than unity has been achieved in metal-based [103, 104], GaAs-based [105], and Si-based [48, 50, 106] SETs.

The current cut-off characteristics are determined by the subthreshold slope S in the I_d–V_g characteristics that rise and fall almost exponentially at the tails of the peaks. Figure 3.8 (lower panel) shows the output drain current I_d for a fixed input drain voltage V_d plotted as a function of V_g on a logarithmic scale. At a sufficiently low temperature and high tunnel resistance, S is given by

$$S = [d(\log_{10} I_d)/dV_g]^{-1} = \ln 10(C_\Sigma/C_g)kT/e \tag{3.3}$$

This equation is similar to that for a MOSFET. It also indicates that a high inverting voltage gain G_I is needed to obtain a steep subthreshold slope. Upon $C_s = C_d$, a G_I of 4 corresponds to a C_Σ/C_g of 1.5, or $S = 90\,\mathrm{mV\,dec^{-1}}$ at room temperature.

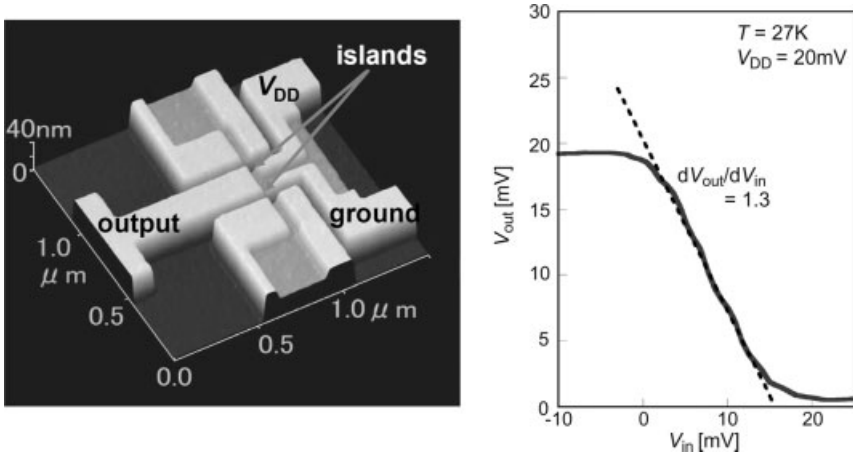


Figure 3.9 Si complementary single-electron inverter. Left: AFM image. Right: Input–output transfer curve measured with a power supply voltage V_{DD} of 20 mV. The measurement temperature was 27 K [89].

A logic circuit can be made by employing the above-mentioned voltage gain. The complementary inverter, which is the most fundamental logic element, was fabricated using Si SETs [89]. Figure 3.9 (left) shows an atomic force microscopy (AFM) image of an SET inverter made by Si. Two SETs with a voltage gain of about 2 were packed in a small (100×200 nm) area. As shown in Figure 3.9 (right), the input and output transfer curve attains both a larger than unity gain and a full logic swing at 27 K. Other complementary SET inverters, made from Al [90] and carbon-nanotube [91] have been reported, and resistive-loaded inverters have also been fabricated [92, 93].

3.4.2 Multiple-Gate SET and Pass-Transistor Logic

An important feature of SETs is that they can have *plural gates*. Such a multi-gate configuration enables the sum-of-products function to be implemented at the gate input level. That is, the total charges induced in the gates are expressed as $\Sigma C_{gi} V_{gi}$, where C_{gi} and V_{gi} are the gate capacitance and input voltage of the i-th gate. Provided that the gate input voltage V_{gi} is set to $e/2C_{gi}$, the SET is ON when the number of the ON gates is odd, and OFF when the number of the ON gates is even (Figure 3.10). This function is XOR. The SET XOR gate [65, 68] is a powerful tool for constructing arithmetic units such as adders and multipliers because XOR is nothing other than what is termed "half-sum", which is the lower order bit calculated by adding two one-bit binary numbers. The XOR gate has also been demonstrated experimentally using a Si dual-gate SET [94]. A scanning electron microscopy (SEM) image of the dual-gate SET is shown in Figure 3.11. The XOR function was confirmed in output drain current at 40 K, as shown in the figure.

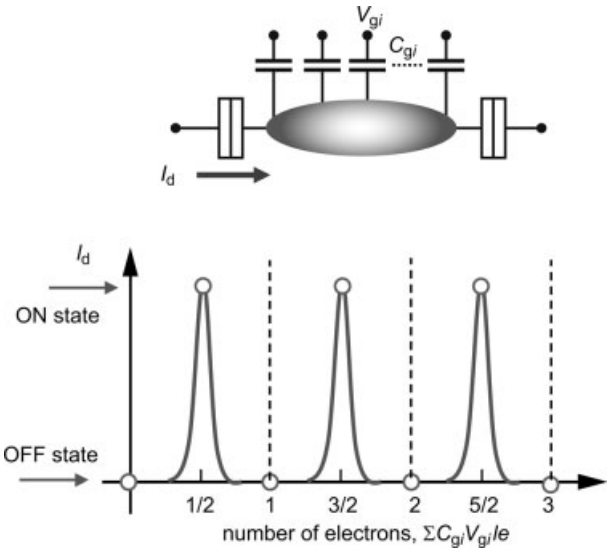


Figure 3.10 A multigate SET. Top: An equivalent circuit. Bottom: Current characteristics as a function of number of charges accumulated at the gates.

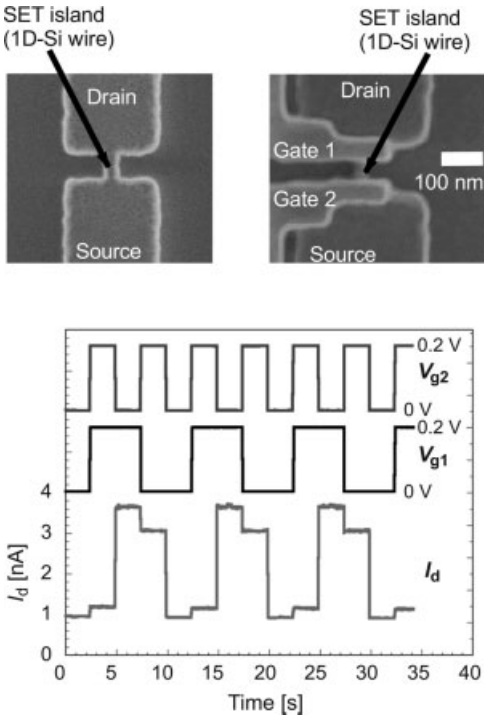


Figure 3.11 An experimental SET X-OR gate. Top: SEM images of the device before (left) and after (right) gate formation. Bottom: Output drain current for square-wave gate inputs [94].

In the multi-gate configuration, the gate capacitance for each gate is inherently smaller than that of the single-gate version. Therefore, it is more difficult to attain the larger-than-unity gain as the number of the gate increases. The CMOS-domino-type logic was proposed as a way of using SETs without a voltage gain [73]. A combinational logic circuit is built in a SET logic tree, where SETs are used as pull-down transistors. The point is that the tree is operated with a sufficiently small drain voltage in order to make the Coulomb blockade effective. The output signal is then amplified by using MOSFETs before being transferred to the next logic segment.

A single-electron pass-transistor logic, where SETs are used both as pull-up and pull-down transistors, has also been studied. The fundamental circuit of the single-electron pass-transistor logic was fabricated using PADOX SETs, and half-sum and carry-out for the half adder has been experimentally demonstrated [95, 96]. Figure 3.12 shows the equivalent circuits and the measurement data. Both half-sum and carry-out are correctly output at 25 K. What is significant here is that the gate and total capacitances, and even the peak positions of the used SETs, were well controlled for these operations. This is the first arithmetic operation ever performed by SET-based circuits. There have been attempts to construct logic elements [97–102] that operate based on the above-mentioned domino-type logic, pass-transistor logic, or the so-called binary-decision-diagram logic.

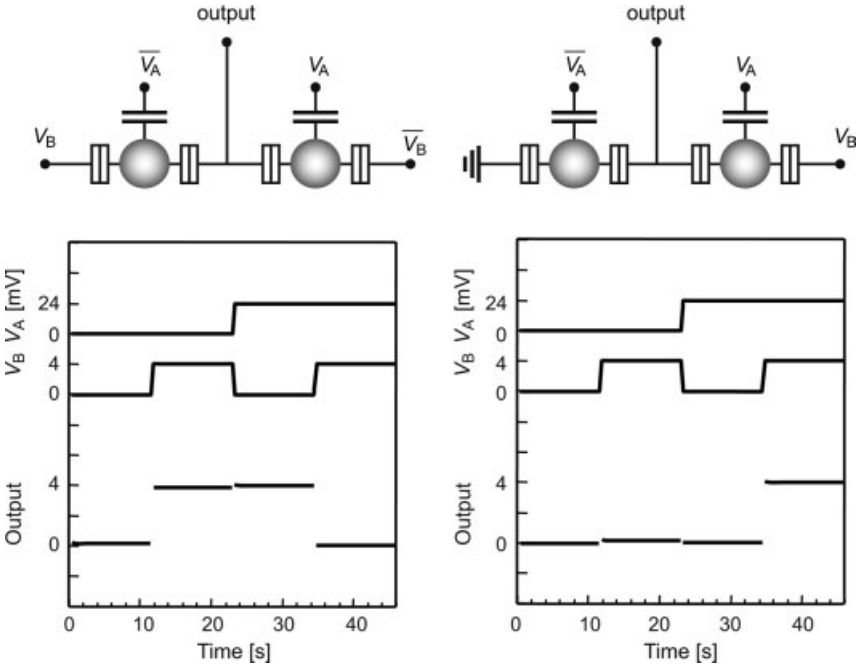


Figure 3.12 "Half-sum" and "carry-out" operations using SETs. In the equivalent circuits, V_A, V_B are inputs for addends and $\bar{V}_A$, $\bar{V}_B$ are their inverses [96].

3.4.3
Combined SET-MOSFET Configuration and Multiple-Valued Logic

In SETs, the applicable drain voltage is limited to a value smaller than e/C_Σ in order to maintain the Coulomb blockade. This may be an obstacle to driving a series of SETs and external circuits that require a high input voltage. A combined SET-MOSFET configuration has been proposed as a way to overcome this drawback [83, 85]. Figure 3.13(left) shows the equivalent circuit of the inverter based on this configuration. A MOSFET with a fixed gate bias V_{gg} is connected to the drain of a SET, and the inverter is driven by a constant current load, I_0. The MOSFET keeps the SET drain voltage sufficiently low, which helps to maintain the Coulomb blockade. As the drain voltage is almost independent of the output voltage, V_{out}, a large output voltage and voltage gain can be obtained.

The output voltage V_{out} and output resistance of the combined SET-MOSFET inverter are given by [86]:

$$V_{out} = -G_{m(SET)} R_{d(SET)} (1 + G_{m(MOS)} R_{d(MOS)}) V_{in}, \tag{3.4}$$

$$R_{out} = R_{d(MOS)} + (1 + G_{m(MOS)} R_{d(MOS)}) R_{d(SET)} \tag{3.5}$$

where $G_{m(SET)}$ is the transconductance of the SET, and $R_{d(SET)}$ and $R_{d(MOS)}$ are the drain resistances of the SET and MOSFET, respectively. The voltage gain of the SET is multiplied by that of the MOSFET, which means that the voltage gain of the SET-MOSFET inverter becomes very large because of the large voltage gain of the

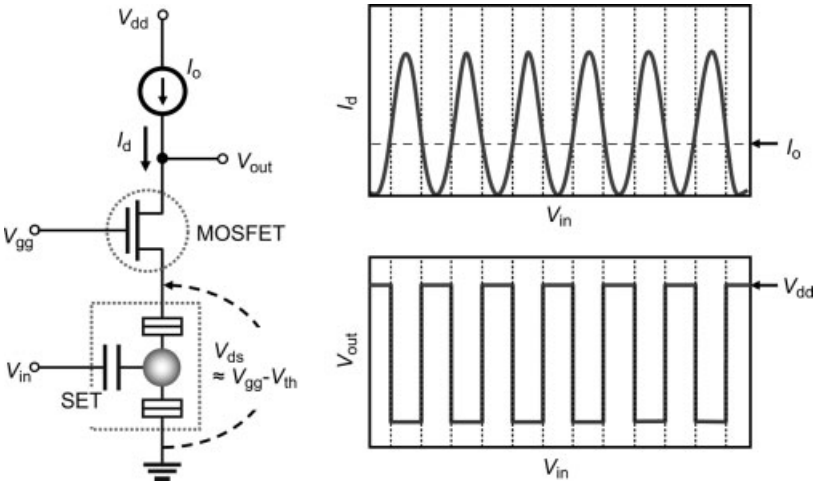


Figure 3.13 Left: Schematic of the universal literal gate comprising a SET, a MOSFET, and a constant-current load I_0. Right: $I_d - V_{in}$, and expected transfer ($V_{in} - V_{out}$) characteristics. $I_d - V_{in}$ characteristics are almost completely independent of V_{out} as the V_{ds} of the SET is kept nearly constant at ($V_{gg} - V_{th}$, the threshold voltage of the MOSFET [86].

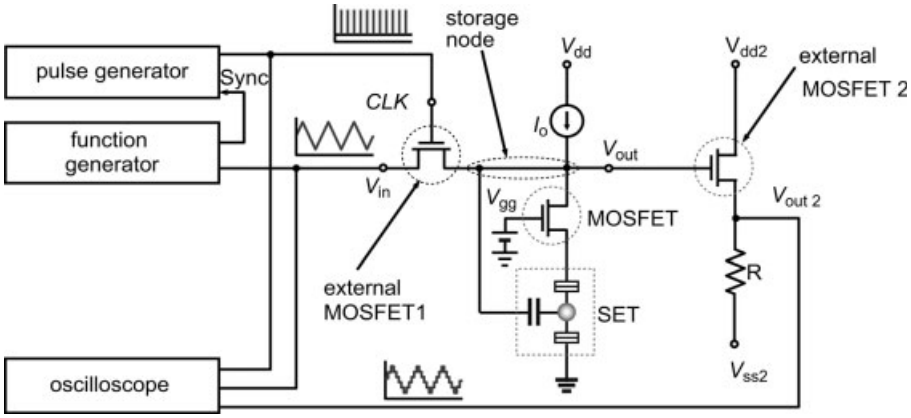


Figure 3.14 Measurement set-up for the single-electron quantizer [85]. CLK = clock.

MOSFET (see Figure 3.13, right). In fact, the measured voltage gain of the SET-MOSFET inverter was about 40 [86].

In this configuration, another important point is that the $I_d - V_g$ characteristics reflect the oscillatory I_d–V_g characteristics (Figure 3.13, right). This characteristic is referred to as the "universal literal", which is a basic unit for multiple-valued logic. Multiple-valued logics have potential advantages over binary logics with respect to the number of elements per function and operating speed. They are also expected to relax the interconnection complexity inside and outside of LSIs. These are advantageous, as they allow a further reduction in the power dissipation in LSIs and the chip sizes. However, success has been limited, partially because the devices that have been used (MOSFETs and negative-differential-resistance devices, such as resonant tunneling diodes) are inherently single-threshold or single-peak, and are not fully suited for multiple-valued logic. The oscillatory behavior seen in Figure 3.13 shows that the SET is suitable for implementing multiple-valued logic. By exploiting this behavior, a quantizer was fabricated. Figures 3.14 and 3.15 show the measurement set-up for the quantizer and the measured data, respectively. The triangular input V_{in} was successfully quantized into six levels.

3.4.4 Considerations on SET Logic

Many research groups have claimed that SETs could provide low-power circuits. In order to make clear the meaning of this claim, two parameters must first be discussed, namely information throughput I and the power density P. These parameters can be written in the following forms:

$$I = \alpha n f \tag{3.6}$$

$$P = E_{bit} n f \tag{3.7}$$

where n is the density of the binary switches, f the operating frequency, and E_{bit} the bit energy. A dimensionless parameter, α, was introduced which was referred to

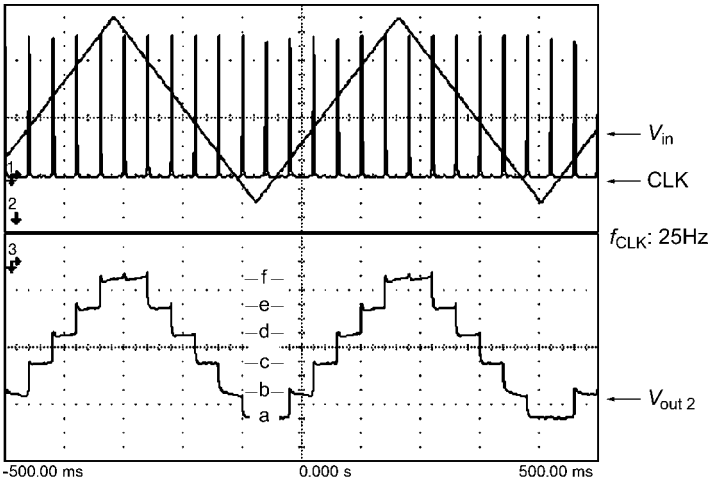


Figure 3.15 Experimental data for the single-electron quantizer [85]. CLK = clock.

as "functionality". The meaning of this parameter is simple; for example, if a transistor is available that can work more efficiently than a simple binary switch, then the information throughput can be increased. In such a case, the transistor has α larger than unity. Then, for low-power operations the aim would be to increase I while keeping P small. Therefore, an important parameter is the information throughput per power density, I/P, which is given by α/E_{bit}. A lower E_{bit} is better for larger I/P, but E_{bit} has a lower bound in order to avoid noise-induced bit errors. This minimum value is dependent on how many errors the system allows, and thus on the system architecture. However, the minimum E_{bit} will not change significantly unless the architecture is changed to an exotic version, like a neural network or fuzzy logic. Thus, the only option is to increase α. If the suggestion was to increase α on the device level, then there would be a need to depart from logic based on simple binary switches. This leads to a very important conclusion – that changing materials for the transistor channel, say, to carbon nanotubes or other molecules, is not the way to reduce dynamic power loss as long when transistors are used as binary switches. Hence, it is expected that the SET be an alternative device and would be highly functional, as shown previously in the chapter.

It should be mentioned, however, that the above discussion is rather too crude to draw any decisive conclusions. Actually, I/P does not include n and f, but indeed a large-scale integration and a fast device is needed in order to accomplish computation within acceptably short periods of time. Therefore, a more reasonable parameter may be $I^2/P\ (=\alpha^2 nf/E_{bit})$, and the size and the speed of the device need still to be discussed. At this time the static power loss should also be considered – that is, the loss due to leakage currents, which is independent of f. These points are discussed in the following sections.

It is important now to highlight once more the difference between voltage-state logic (or SET logic) and charge-state logic. The requirement for the addition energy – and hence for the island size – is different between the two. For charge-state logic, the

addition energy should be sufficiently large so as to avoid bit errors caused by thermal noise. As charge-state logic uses single electrons to represent bit information, the bit energy is given by $(1/2)C_\Sigma V^2$, where $V = e/C_\Sigma$. This is in effect the addition energy. If the bit error requirements for CMOS LSIs are assumed, then the bit energy will have to be 10^2 larger than the thermal noise energy. This means that an additional energy 10^2 larger than the room-temperature energy is needed. From Figure 3.4 it is clear that an island as small as 1 nm is needed. SET logic, on the other hand, uses the voltage generated at output terminals to represent bits, as do CMOS circuits. The bit energy is therefore given by $(1/2)C_L V^2$, where C_L is the load capacitance. If the term $V = e/C_\Sigma$ is adopted for the power supply voltage of SET logic circuits, then the bit energy is C_L/C_Σ larger than the case for charge-state logic. Therefore, for SET logic, the addition energy requirement does not come from the bit error requirement but rather from the static power loss because a small addition energy causes the valley current (i.e. the OFF-current) to increase. There is no clear guideline as to how small the OFF-current should be, because the acceptable static power loss depends on the degree of the power-saving ability of the system. If the requirement for low-operating-power (LOP) applications are adopted – as stated in the International Technology Roadmap for Semiconductors (ITRS) – the source/drain OFF-state leakage current should be on the order of 10^{-9} A μm^{-1}, which corresponds to 10 pA for 10-nm SETs. This will be achievable. It is also necessary to have a large ON/OFF current ratio and, again, considering the requirement for LOP applications, a ratio of 10^5 is needed. From this ratio, the addition energy should be about 16 kT or larger, which was derived based on the standard theory of single-electron tunneling. This estimation does not consider the quantum leakage current, which becomes significant as the junction resistance approaches the quantum resistance. However, as long as the junction resistance is not very close to the quantum resistance, the above criteria for the addition energy will be a reasonable basis for later discussions. With this requirement, an addition energy E as large as 0.4 eV is required for room-temperature operation. This addition energy corresponds to the total island capacitance ($C_\Sigma = e^2/E$) of 0.4 aF, an excitation voltage (e/C_Σ) of 0.4 V and, from Figure 3.4, an island size of about 3 nm.

There are two time scales for evaluating SET speed: one is the intrinsic switching time, and the other for the circuit speed. The intrinsic switching time defines how fast the SET changes its states, and is determined by the RC time constant of the tunneling, $C_\Sigma R_j$, where R_j is the junction resistance. If it is assumed that $C_\Sigma = 0.4$ aF and $R_j = 1$ MΩ ($\sim 4R_q$), the switching speed will be 0.4 ps and thus the SET is a fairly fast switching device. The problem is that only one electron is moved by the switching event, and it thus takes a much longer time to change the state of the output terminal with a larger capacitance. The time for changing the state of the output terminal is determined by $C_L R_{SET}$, where R_{SET} is the resistance of the SET ($\sim 4R_j$). If it is assumed that $C_L = 100$ aF and $R_{SET} = 4$ MΩ, the time is 0.4 ns or 2.5 GHz. It will also be helpful to compare the SET current density with that of the present nMOS transistor (for LOP applications), which is about 600 μA μm^{-1}. Assuming that the size, resistance, and drain voltage of the SET are 3 nm, 4 MΩ, and 0.4 V, respectively, the SET current density will be 33 μA μm^{-1}. Although this is not fatally bad, it implies that the use of SETs is restricted to a local communication with relatively small load capacitances. Crudely

speaking, the SET is inherently slower by the factor of at least 10^{-1} than FETs because the SET cannot operate with the resistance smaller than R_q, whereas FETs can.

The SET can be made very small, but this does not necessarily mean it will be the smallest. Ideally, a molecular-sized FET can be imagined, and could be made as small as the SET. Therefore, small size is not a major merit of the SET; rather, the main merit is that its operation is guaranteed even at the molecular level, and some parameters – such as the switching speed and current peak-to-valley ratio – can be improved owing to the reduced capacitance. At this point, it might be safe to say that SETs have no definite advantage over ultimately scaled-down MOSFETs from the viewpoint of the physical size itself.

Now, a return should be made to Eqs. (3.6) and (3.7). Considering the above arguments that SET size is comparable to that of ultimate future FETs, and that the circuit speed is 10^{-1} to 10^{-2} worse than its CMOS counterparts, the functionality α will need to be increased by 10^1 or more in order to make a drastic improvement in I/P while keeping I^2/P comparable to that for CMOS circuits.

Several ideas for improving the functionality have been reported. One is to use the SET XOR gates introduced in Section 3.4.2. Figure 3.16 shows the equivalent circuit of a full adder based on the SET XOR gate. A full adder can be constructed using six SETs, whereas this requires 28 MOSFETs in CMOS logic [106]. This can be interpreted as $\alpha = 4.7$. It has also been reported that, by integrating SET full adders, multi-bit adders can be constructed in a very area-efficient manner: there are no long wires in the carry-propagation path, which leads to fast operation in spite of a low drivability of the SET [80]. Another idea is based on the SET-MOSFET configuration introduced in Section 3.4.3. Based on this configuration, a SET logic gate family has been proposed [87, 88].

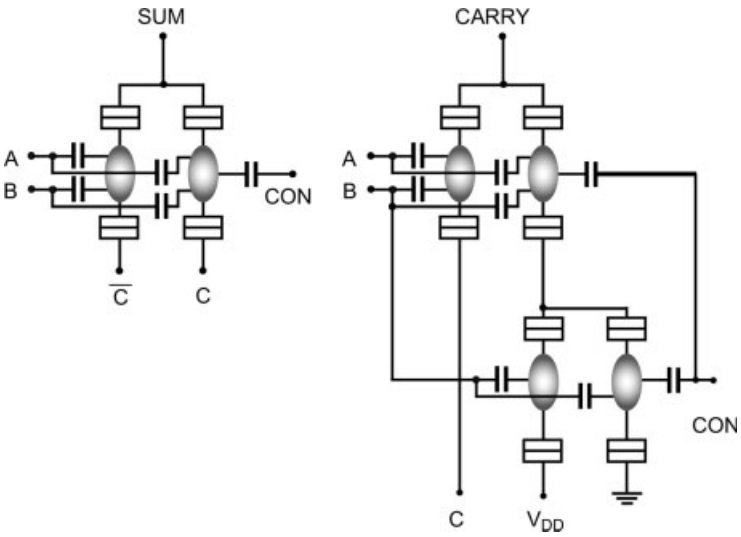


Figure 3.16 The SET full adder. A and B are addends and C is carry. CON is the control signal, which controls the phase of the Coulomb blockade oscillation [80].

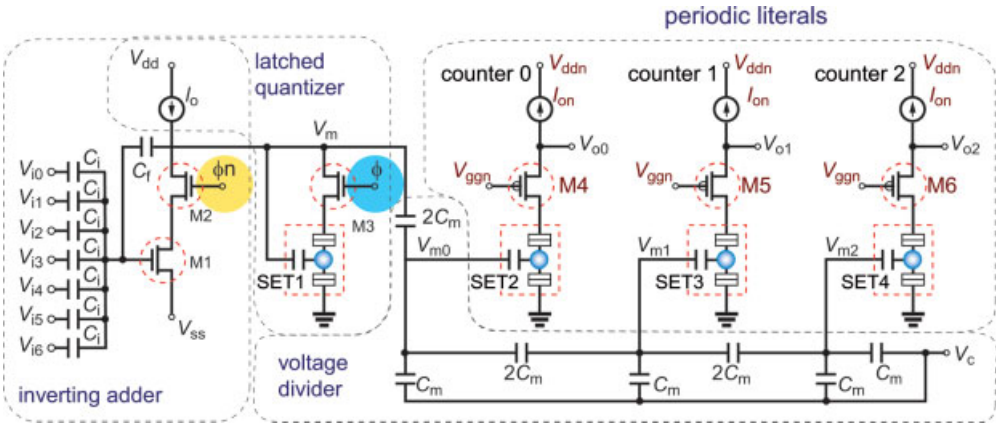


Figure 3.17 Circuit diagram of SET-based 7-3 counter. The circuit consists of five types of device: SETs (SET1–SET4), n-channel MOSFETs (M1–M3), p-channel MOSFETs (M4–M6), and constant current loads for the first stage and literals. No adjustment is required in the device parameters for devices of the same type. Clock ϕ and ϕn are complementary, and the multiple-valued data are latched when ϕ is high. V_{ddn} is set at a negative value to provide consistent voltage levels among the circuit blocks in and out of the counter [87].

These SET logic gates are useful for implementing binary logic circuits, multiple-valued logic circuits, and binary/multiple-valued mixed logic circuits in a highly flexible manner. As an example, a 7-3 counter is shown in Figure 3.17. This is a member of the M-N counters, which are generalized counters defined in the framework of Counter Tree Diagrams. Most adders, including those for redundant number systems, could be represented in this framework. The 7-3 counter can be constructed using four SETs and 10 FETs with some passive components, whereas 198 FETs are required in CMOS logic. The functionality α is of the order of 10^1 in this case.

The increase in α in the two examples is due to the application of the SET periodic function for implementing the operation "add". This is because the parity and the periodic function are the fundamentals of the arithmetic. These examples strongly suggest that the best use of the SETs will be in arithmetic units such as adders, while other arithmetic units such as multipliers are made from adders. Adders can be built by a repetition of relatively simple layouts, and require lesser amounts of long wiring, which will compensate for the low drivability of SETs. It is believed that, by pursuing this direction, a more efficient way to increase the functionality will be found. For this purpose, much larger-scale circuits should be investigated than have been studied to data.

In summary, the SET can function as a fairly fast switching device and, although the SET has low drivability, this will not prove fatal. The periodic function of the SET current characteristics suggests that it should be applied to arithmetic units such as adders and multipliers, which might reduce their dynamic power consumption. The most suitable applications for SET-based voltage-state circuits will be for LOP arithmetic units. However, there may be no merit in using SETs in terms of static

power consumption, and some elaborate system architectures would be required to reduce this. For low standby power applications, charge-state logic circuits should pursued where, in principle, there is no leakage current.

3.5 Conclusions

Despite concern persisting with regards to reducing dynamic power consumption, this problem will not be solved simply by changing the raw materials of transistor production. One way of reducing power requirements is to use functional devices rather than simple switches, and among the large numbers of emerging devices the SET is one of the best functional units, on the basis of its unique current characteristics.

However, two critical problems remain when applying SETs to logic circuits. The first – basically technological – problem is to control the size of the nanometer-scale islands and attached tunnel junctions. During the early 1990s, very few investigators considered that SETs operating at room temperature could be fabricated, yet today they can be prepared with good ON-OFF current ratios. Moreover, their performance continues to improve. With this in mind it is likely that, in the future, a new technology will emerge for integrating millions of room-temperature-operating SETs. The second problem is to identify the "killer" applications for SETs, and this is a more fundamental and critical question. During recent years, much effort has been expended in designing SET circuits, and those which utilize the periodic nature of the SET current characteristics appear to show the greatest promise for the construction of LOP circuits. Nonetheless, further studies will be necessary to develop SET circuits that are sufficiently powerful to surpass CMOS circuits, or at least to replace a proportion of them. For this purpose, a collaboration among system architects, circuit designers, and process engineers is clearly called for.

References

1 J. D. Meindl (Ed.), *Proc. IEEE* 2001, 89.

2 D. V. Averin, K. K. Likharev, *J. Low Temp. Phys.* 1986, **62**, 345.

3 L. S. Kuzmin, K. K. Likharev, *J. Exp. Theoret. Physics Lett. Lett.* 1987, **45**, 495.

4 T. A. Fulton, G. J. Doran, *Phys. Rev. Lett.* 1987, **59**, 109.

5 K. K. Likharev, *IEEE Trans. Magn.* 1987, **23**, 1142.

6 D. V. Averin, K. K. Likharev, *Mesoscopic Phenomena in Solids*, Chapter 6, B. L. Altshuler, P. A. Lee, R. A. Webb (Eds.), Elsevier, Amsterdam, 1991.

7 H. Grabert, M. H. Devoret (Eds.), *Single Charge Tunneling*, Plenum, New York, 1992.

8 K. K. Likharev, *Proc. IEEE* 1999, **87**, 606.

9 M. A. Kastner, *Rev. Mod. Phys.* 1992, **64**, 849.

10 L. J. Geerligs, V. F. Anderegg, P. A. M. Holweg, J. E. Mooij, *Phys. Rev. Lett.* 1990, **64**, 2691.

11 H. Pothier, P. Lafarge, P. F. Orfila, C. Urbina, D. Esteve, M. H. Devoret, *Physica B* 1991, **169**, 573.

12 Y. Ono, A. Fujiwara, K. Nishiguchi, H. Inokawa, Y. Takahashi, *J. Appl. Phys.* 2005, **97**, 031101.

13 M. Saitoh, N. Takahashi, H. Ishikuro, T. Hiramoto, *Jpn. J. Appl. Phys.* 2001, **40**, 2010.

14 J. H. F. Scott-Thomas, S. B. Field, M. A. Kastner, H. I. Smith, D. A. Antoniadis, *Phys. Rev. Lett.* 1989, **62**, 583.

15 U. Meirav, M. A. Kastner, S. J. Wind, *Phys. Rev. Lett.* 1990, **65**, 771.

16 C. de Graaf, J. Caro, S. Radelaar, V. Lauer, K. Heyers, *Phys. Rev. B* 1991, **44**, 9072.

17 H. Matsuoka, T. Ichiguchi, T. Yoshimura, E. Takeda, *Appl. Phys. Lett.* 1994, **64**, 586.

18 D. Ali, H. Ahmed, *Appl. Phys. Lett.* 1994, **64**, 2119.

19 M. Bockrath, D. H. Cobden, P. L. McEuen, N. G. Chopra, A. Zettl, A. Thess, R. E. Smalley, *Science* 1997, **275**, 1922.

20 S. J. Tans, M. H. Devoret, H. Dai, A. Thess, R. E. Smalley, L. J. Geerlings, C. Dekker, *Nature* 1997, **386**, 474.

21 K. Ishibashi, D. Tsuya, M. Suzuki, Y. Aoyagi, *Appl. Phys. Lett.* 2003, **82**, 3307.

22 K. Matsumoto, S. Kinoshita, Y. Gotoh, K. Kurachi, T. Kamimura, M. Maeda, K. Sakamoto, M. Kuwahara, N. Atoda, Y. Awano, *Jpn. J. Appl. Phys.* 2003, **42**, 2415.

23 H. Nejoh, *Nature* 1991, **353**, 640.

24 V. Mujica, M. Kemp, A. Roitberg, M. Ratner, *J. Chem. Phys.* 1996, **104**, 7296.

25 N. B. Zhitenev, H. Meng, Z. Bao, *Phys. Rev. Lett.* 2002, **88**, 226801.

26 J. Park, A. N. Pasupathy, J. I. Goldsmith, C. Chang, Y. Yaish, J. R. Petta, M. Rinkoski, J. P. Sethna, H. D. Abruna, P. L. McEuen, D. C. Ralph, *Nature* 2002, **417**, 722.

27 W. Liang, M. P. Shores, M. Bockrath, J. R. Long, H. Park, *Nature* 2002, **417**, 725.

28 S. Kubatkin, A. Danilov, M. Hjort, J. Cornil, J.-L. Bredas, N. Stuhr-Hansen, P. Hedegard, T. Bjørnholm, *Nature* 2003, **425**, 698.

29 G. J. Doran, *Appl. Phys. Lett.* 1977, **31**, 337.

30 Y. A. Pashkin, Y. Nakamura, J. S. Tsai, *Appl. Phys. Lett.* 2000, **76**, 2256.

31 J.-P. Colinge (Ed.), *Silicon-on-insulator technology: Materials to VLSI*, 2nd edn., Kluwer Academic Publishers, Boston, 1997.

32 Y. Takahashi, M. Nagase, H. Namatsu, K. Kurihara, K. Iwadate, Y. Nakajima, S. Horiguchi, K. Murase, M. Tabe, *International Electron Devices Meeting, Technical Digest, p.* 938, IEEE, Piscataway, NJ, 1994.

33 Y. Takahashi, M. Nagase, H. Namatsu, K. Kurihara, K. Iwadate, Y. Nakajima, S. Horiguchi, K. Murase, M. Tabe, *Electronics Lett.* 1995, **31**, 136.

34 Y. Takahashi, H. Namatsu, K. Kurihara, K. Iwadate, M. Nagase, K. Murase, *IEEE Trans. Electron Devices* 1996, **43**, 1213.

35 E. Leobandung, L. Guo, Y. Wang, S. Y. Chou, *Appl. Phys. Lett.* 1995, **67**, 938.

36 E. Leobandung, L. Guo, S. Y. Chou, *Appl. Phys. Lett.* 1995, **67**, 2338.

37 H. Ishikuro, T. Fujii, T. Saraya, G. Hashiguchi, T. Hiramoto, T. Ikoma, *Appl. Phys. Lett.* 1996, **68**, 3585.

38 H. Ishikuro, T. Hiramoto, *Appl. Phys. Lett.* 1997, **71**, 3691.

39 H. Ishikuro, T. Hiramoto, *Appl. Phys. Lett.* 1999, **74**, 1126.

40 D. H. Kim, J. D. Lee, B.-G. Park, *Jpn. J. Appl. Phys.* 2000, **39**, 2329.

41 D. H. Kim, S.-K. Sung, K. R. Kim, J. D. Lee, B.-G. Park, B. H. Choi, S. W. Hwang, D. A. Park, *IEEE Trans. Electron Devices* 2002, **49**, 627.

42 D. H. Kim, S.-K. Sung, K. R. Kim, J. D. Lee, B.-G. Park, *J. Vac. Sci. Technol.* 2002, **B20**, 1410.

43 K. Uchida, J. Koga, R. Ohba, A. Toriumi, *IEEE Trans. Electron Devices* 2003, **50**, 1623.

44 K. R. Kim, D. H. Kim, J. D. Lee, B.-G. Park, *Appl. Phys. Lett.* 2004, **84**, 3178.

45 M. Saitoh, T. Hiramoto, *Appl. Phys. Lett.* 2004, **84**, 3172.

46 M. Saitoh, H. Harata, T. Hiramoto, *Jpn. J. Appl. Phys.* 2005, **44**, L338.

47 H. Harata, M. Saitoh, T. Hiramoto, *Jpn. J. Appl. Phys.* 2005, **44**, L640.

48 K. Miyaji, M. Saitoh, T. Hiramoto, *Appl. Phys. Lett.* 2006, **88**, 143505.

49 Y. Ono, Y. Takahashi, K. Yamazaki, M. Nagase, H. Namatsu, K. Kurihara, K. Murase, *IEEE Trans. Electron Devices* 2000, **47**, 147.

50 Y. Ono, K. Yamazaki, Y. Takahashi, *IEICE Trans. Electron.* 2001, **E-84C**, 1061.

51 K. Shiraishi, M. Nagase, S. Horiguchi, H. Kageshima, M. Uematsu, Y. Takahashi, K. Murase, *Physica* 2000, **E 7**, 337.

52 S. Horiguchi, M. Nagase, K. Shiraishi, H. Kageshima, Y. Takahashi, K. Murase, *Jpn. J. Appl. Phys.* 2001, **40**, L29.

53 A. Fujiwara, Y. Takahashi, K. Murase, M. Tabe, *Appl. Phys. Lett.* 1995, **67**, 2957.

54 A. B. Zorin, F.-J. Ahlers, J. Niemeyer, T. Weimann, H. Wolf, V. A. Krupenin, S. V. Lotkhov, *Phys. Rev. B* 1996, **53**, 13682.

55 W. H. Huber, S. B. Martin, N. M. Zimmerman, *Proceedings of Experimental Implementation of Quantum Computation, p.* 176, Rinton Press, Princeton, 2001.

56 N. M. Zimmerman, W. H. Huber, A. Fujiwara, Y. Takahashi, *Appl. Phys. Lett.* 2001, **79**, 3188.

57 Y. Takahashi, Y. Ono, A. Fujiwara, K. Shiraishi, M. Nagase, S. Horiguchi, K. Murase, *Proceedings of Experimental Implementation of Quantum Computation,* p. 183, Rinton Press, Princeton, 2001.

58 A. Fujiwara, M. Nagase, S. Horiguchi, Y. Takahashi, *Jpn. J. Appl. Phys.* 2003, **42**, 2429.

59 Y. Takahashi, Y. Ono, A. Fujiwara, H. Inokawa, *J. Phys.: Condens. Matter* 2002, **14**, R995.

60 J. R. Tucker, *J. Appl. Phys.* 1992, **72**, 4399.

61 M. I. Lutwyche, Y. Wada, *J. Appl. Phys.* 1994, **75**, 3654.

62 M. Kirihara, N. Kuwamura, K. Taniguchi, C. Hamaguchi, *Ext. Abstracts 1994 International Conference on Solid State Devices and Materials*, p. 328, Business Center for Academic Societies Japan, Tokyo, 1994.

63 H. Fukui, M. Fujishima, K. Hoh, *Jpn. J. Appl. Phys.* 1995, **34**, 1345.

64 A. N. Korotkov, R. H. Chen, K. K. Likharev, *J. Appl. Phys.* 1995, **78**, 2520.

65 R. H. Chen, A. N. Korotkov, K. K. Likharev, *Appl. Phys. Lett.* 1996, **68**, 1954.

66 S. Amakawa, H. Fukui, M. Fujishima, K. Hoh, *Jpn. J. Appl. Phys.* 1996, **35**, 1146.

67 M. Fujishima, H. Fukui, S. Amakawa, K. Hoh, *IEICE Trans. Electron.* 1997, **E80-C**, 881.

68 M.-Y. Jeong, Y.-H. Jeong, S.-W. Hwang, D.-M. Kim, *Jpn. J. Appl. Phys.* 1997, **36**, 6706.

69 H. Iwamura, M. Akazawa, Y. Amemiya, *IEICE Trans. Electron.* 1998, **E81-C**, 42.

70 S. Amakawa, H. Majima, H. Fukui, M. Fujishima, K. Hoh, *IEICE Trans. Electron.* 1998, **E81-C**, 21.

71 M. Kirihara, K. Nakazato, M. Wagner, *Jpn. J. Appl. Phys.* 1999, **38**, 2028.

72 M. Akazawa, K. Kanaami, T. Yamada, Y. Amemiya, *IEICE Trans. Electron.* 1999, **E82-C**, 1607.

73 K. Uchida, K. Matsuzawa, A. Toriumi, *Jpn. J. Appl. Phys.* 1999, **38**, 4027.

74 S. Shimano, K. Masu, K. Tsubouchi, *Jpn. J. Appl. Phys.* 1999, **38**, 403.

75 Y. Takahashi, A. Fujiwara, Y. Ono, K. Murase, *Proceedings 30th IEEE International Symposium on Multi-Valued Logic,* p. 411, IEEE, Los Alamitos, CA, 2000.

76 K. Uchida, K. Matsuzawa, J. Koga, R. Ohba, S. Takagi, A. Toriumi, *Jpn. J. Appl. Phys.* 2000, **39**, 2321.

77 K. Uchida, J. Koga, R. Ohba, A. Toriumi, *IEICE Trans. Electron.* 2001, **E84-C**, 1066.

78 M.-Y. Jeong, B.-H. Lee, Y.-H. Jeong, *Jpn. J. Appl. Phys.* 2001, **40**, 2054.

79 K.-T. Liu, A. Fujiwara, Y. Takahashi, K. Murase, Y. Horikoshi, *Jpn. J. Appl. Phys.* 2002, **41**, 458.

80 Y. Ono, H. Inokawa, Y. Takahashi, *IEEE Trans. Nanotechnol.* 2002, **1**, 93.

81 Y. S. Yu, J. H. Oh, S. W. Hwang, D. Ahn, *Electronics Lett.* 2002, **38**, 850.

82 Y. Mizugaki, P. Delsing, *Jpn. J. Appl. Phys.* 2001, **40**, 6157.

83 H. Inokawa, A. Fujiwara, Y. Takahashi, *Appl. Phys. Lett.* 2001, **79**, 3618.

84 H. Inokawa, Y. Takahashi, *IEEE Trans. Electron Devices* 2003, **50**, 455.

85 H. Inokawa, A. Fujiwara, Y. Takahashi, *IEEE Trans. Electron Devices* 2003, **50**, 462.

86 H. Inokawa, A. Fujiwara, Y. Takahashi, *Jpn. J. Appl. Phys.* 2002, **41**, 2566.

87 H. Inokawa, Y. Takahashi, K. Degawa, T. Aoki, T. Higuchi, *IEICE Trans. Electron.* 2004, **E87-C**, 1818.

88 K. Degawa, T. Aoki, T. Higuchi, H. Inokawa, Y. Takahashi, *IEICE Trans. Electron.* 2004, **E87-C**, 1827.

89 Y. Ono, Y. Takahashi, K. Yamazaki, M. Nagase, H. Namatsu, K. Kurihara, K. Murase, *Appl. Phys. Lett.* 2000, **76**, 3121.

90 C. P. Heij, P. Hadley, J. E. Mooij, *Appl. Phys. Lett.* 2001, **78**, 1140.

91 D. Tsuya, M. Suzuki, Y. Aoyagi, K. Ishibashi, *Jpn. J. Appl. Phys.* 2005, **44**, 1588.

92 F. Nakajima, K. Kumakura, J. Motohisa, T. Fukui, *Jpn. J. Appl. Phys.* 1999, **38**, 415.

93 K. Nishiguchi, S. Oda, *Appl. Phys. Lett.* 2001, **78**, 2070.

94 Y. Takahashi, A. Fujiwara, K. Yamazaki, H. Namatsu, K. Kurihara, K. Murase, *Appl. Phys. Lett.* 2000, **76**, 637.

95 Y. Ono, Y. Takahashi, *International Electron Devices Meeting, Technical Digest*, p. 297, IEEE, Piscataway, NJ, 2000.

96 Y. Ono, K. Yamazaki, M. Nagase, S. Horiguchi, K. Shiraishi, Y. Takahashi, *Microelectron. Eng.* 2001, **59**, 435.

97 N. J. Stone, H. Ahmed, *Electronics Lett.* 1999, **35**, 1883.

98 A. Fujiwara, Y. Takahashi, K. Yamazaki, H. Namatsu, M. Nagase, K. Kurihara, K. Murase, *IEEE Trans. Electron Devices* 1999, **46**, 954.

99 N. Takahashi, H. Ishikuro, T. Hiramoto, International Electron Devices Meeting, Technical Digest, p. 371, IEEE, Piscataway, NJ, 1999.

100 S. Kasai, H. Hasegawa, *IEEE Electron Device Lett.* 2002, **23**, 446.

101 F. Nakajima, Y. Miyoshi, J. Motohisa, T. Fukui, *Appl. Phys. Lett.* 2003, **83**, 2680.

102 Y. Miyoshi, F. Nakajima, J. Motohisa, T. Fukui, *Appl. Phys. Lett.* 2005, **87**, 033501.

103 G. Zimmerli, R. L. Kautz, J. M. Martinis, *Appl. Phys. Lett.* 1992, **61**, 2616.

104 E. H. Visscher, S. M. Verbrugh, J. Lindeman, P. Hadley, J. E. Mooij, *Appl. Phys. Lett.* 1995, **66**, 305.

105 Y. Satoh, H. Okada, K. Jinushi, H. Fujikura, H. Hasegawa, *Jpn. J. Appl. Phys.* 1995, **38**, 410.

106 R. A. Smith, H. Ahmed, *Appl. Phys. Lett.* 1997, **71**, 3838.

107 R. Zimmermann, W. Fichtner, *IEEE J. Solid-State Circuits* 1997, **32**, 1079.

4
Magnetic Domain Wall Logic

Dan A. Allwood and Russell P. Cowburn

4.1
Introduction

The integrated circuit which, during recent years, has become the basis of modern digital electronics, functions by making use of electron charge. However, electrons also possess the quantum mechanical property of spin, which is responsible for magnetism. New "spintronic" technologies seek to make use of this electron spin, sometimes in conjunction with electron charge, in order to achieve new types of device. Several spintronic devices are currently being developed that outperform traditional electronics. Often, this results from an increased functionality, which means that a single spintronic element performs an operation that requires several electronic elements.

Different approaches to spintronics have been developed by the semiconductor and magnetism communities. Although there have been some very impressive demonstrations of spin-polarized charge transport and ferromagnetism in cooled semiconductors [1–3], the lack of a reliable room-temperature semiconductor ferromagnet has hampered their application. Within the magnetism community, however, considerable success has been achieved at room temperature by using common ferromagnetic materials such as $Ni_{81}Fe_{19}$ (Permalloy). This approach offers the benefits of low power operation, non-volatile data storage (no power required), and a high tolerance of both impurities and radiation.

A great success of electronics has been the ability to use groups of transistors for performing Boolean logic operations. Each type of operation has a particular relationship between its input and output states, each of which can take the value "1" or "0". These relationships are shown in the "truth tables" in Table 4.1 for the Boolean NOT, AND, and OR logic operations. Logical NOT has a single input and a single output, with the output having the opposite state of the input. Logical AND has two independent inputs and has a single output that is "1" for an input combination of "11" and "0" otherwise. Conversely, a logical OR output is "0" for an input combination of "00" and "1" otherwise. Importantly, a suitable combination of NOT

Nanotechnology. Volume 4: Information Technology II. Edited by Rainer Waser

ISBN: 978-3-527-31737-0

Table 4.1 Truth tables of some common Boolean logic functions.

NOT		AND			OR		
Input A	**Output B**	**Input A**	**Input B**	**Output C**	**Input A**	**Input B**	**Output C**
0	1	0	0	0	0	0	0
		0	1	0	0	1	1
1	0	1	0	0	1	0	1
		1	1	1	1	1	1

and AND operations, or NOT and OR operations, allows any computation to be performed. The CMOS architecture of NOT and AND logic gates and other circuit elements are listed in Table 4.2. Logical "1" and "0" are represented by the presence or absence of electrical charge, usually measured as either a high and low (zero) voltage. Signal splitting (fan-out) and signal cross-over is achieved by appropriate routing of wire tracks, although for signal cross-over this requires complex fabrication in three dimensions.

Magnetic logic seeks to perform the functions necessary for a logic system with ferromagnetic metals to make use of the advantages that these materials offer. One approach has been to use magnetic/non-magnetic/magnetic tri-layer structures known as *magnetic tunnel junctions* (MTJs) [4, 5]. These have an electrical resistance that depends on the relative orientation of magnetization of the two magnetic layers, and are commonly used in magnetic random access memory (MRAM) [6, 7]. Logic gates made from MTJs [8–12] can perform the logic operations described in Table 4.1, as well as others such as NAND, NOR and XOR. Alternatively, MTJs may be used to provide a switchable bias to CMOS transistors so that a single logic gate may be capable of performing one of two logic operations, say logical AND and OR, as desired [13]. Rather than the MTJ switching on every logic operation, it would simply toggle when the CMOS logic gate operation needs to be switched. This could dramatically increase the logic density of *field-programmable gate arrays* (FPGAs), which are used as flexible alternatives to printed circuit boards.

The other major approach to performing logic operations with magnetic materials has been to propagate magnetic solitons. This can be achieved using chains of isolated nanoscale dots, each element separated by a few tens of nanometers from its neighbors [14–17]. Each dot has uniform magnetization, but the proximity of adjacent dots means that they undergo magnetostatic interactions so that an ordered magnetization configuration is achieved. The same effect is achieved with a row of freely-rotating macroscopic bar magnets. However, defects in the magnetic ordering can be created in the chain of dots. These defects are magnetic solitons, and can be propagated through dot chains by the application of a suitable magnetic field. Furthermore, junctions of dots can be used to perform various Boolean logic functions [17]. The major challenge with this technology is to control the magnetostatic interactions between different pairs of dots in order to improve reliability and device yield.

Table 4.2 Common electronic circuit symbols and the equivalent CMOS and domain wall logic devices.

Symbol	CMOS Circuit	Domain Wall Logic Circuit
Vdd (+5 V)	Charge	Magnetization
0 V	No charge	Magnetization
Fan-out	Input, Output A, Output B	Input, Output A, Output B, 20°, 1 μm, 200 nm, 125 nm, 200 nm
Cross-over	Input A, Input B, Vias, Output A, Output B	Input A, Input B, Output A, Output B, 200 nm, 183 nm, 500 nm, 1 μm
NOT	Input, Output, Vdd	Input / Output, 500 nm, 225 nm, 20°, 100 nm
AND	Input A, Input B, Output, Vdd, NAND, Inverter	Input A, Input B, Output, 20°, 1 μm, 200 nm, 125 nm, 200 nm
NOT-fan hybrid	Sequential placement of fan and NOT elements above	A, B, C, 1 μm, 500 nm, 225 nm, 200 nm, 20°, 200 nm, 100 nm, 750 nm, 1 μm
Data input	Various	325 nm, 1 μm, 1 μm, 200 nm

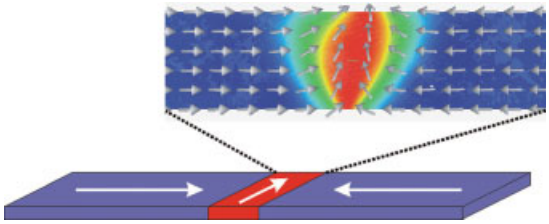


Figure 4.1 Schematic diagram of a transverse 'head-to-head' domain wall in a planar Permalloy nanowire 200 nm wide and 5 nm thick. The expanded region shows a numerically calculated and more detailed view of a domain wall's magnetic structure.

Alternatively, circuits made from planar magnetic nanowires can be used, with wires typically 100 to 250 nm wide and 5 to 10 nm thick. The *shape anisotropy* (geometry) of these wires creates a magnetic easy axis in the wire long axis direction that defines the stable orientations of magnetization (Figure 4.1). This system with two opposite stable magnetization orientations is ideal for representing logical "1" and "0" (see Table 4.2). Where opposite magnetizations meet they are separated by a transition region through which magnetization rotates by 180° (Figure 4.1). This is another form of a magnetic soliton, and is called a *domain wall*. For the wire dimensions relevant here, domain walls are typically approximately 100 nm wide. Domain walls can be moved by applying magnetic fields, and it is this ability which is exploited in magnetic domain wall logic. Domain walls travel down sections of nanowire between nanowire junctions where logic operations are performed. Crucially, the influence of nanowire imperfections on domain wall propagation is very significantly reduced compared with the propagation of solitons in interacting dots. Furthermore, very little power is required either to propagate a domain wall or to perform a logic operation, compared to the lowest power CMOS equivalents or magnetic alternatives. This combination makes magnetic domain wall logic a robust, low-power logic technology. The remainder of this chapter is devoted to explaining how magnetic domain wall logic functions, and what the future prospects of the technology might be.

4.2 Experimental

All of the magnetic structures shown here are fabricated by focused ion beam (FIB) milling [18] of 5 nm-thick Permalloy films. The films were thermally evaporated onto Si(0 0 1) substrates with a native oxide present in a chamber with a base pressure $<10^{-7}$ torr. FIB milling used 30 keV Ga^{+} ions which were focused to a diameter of $\sim$7 nm at the substrate. The Ga^{+} ions sputter the magnetic material and scatter within the film and substrate to implant at lateral distances up to 40 nm from the spot center [19]. Both of these processes lead to a loss of ferromagnetic order in the Permalloy film and allow nanostructures to be defined. A $150 \times 150\,\mu m$ square of magnetic material is cleared around each nanowire circuit to allow optical analysis

of the nanostructure without interference from the surrounding film. Sample images are obtained from secondary electron emission during a single FIB scan.

Magnetization measurements are performed using a magneto-optical Kerr effect (MOKE) magnetometer in the longitudinal MOKE configuration [20]. The instrument has a ~5 μm spatial resolution, given by the focused laser spot diameter, and is sensitive to single magnetization reversal events in individual nanowires [21]. The magnetometer also includes a facility for mapping the sample susceptibility using MOKE signals in order to select particular regions of a structure for measurement. In-plane magnetic fields are applied to samples at a frequency of 27 Hz using a quadrupolar electromagnet. This two-dimensional (2-D) field is characterized by orthogonal fields H_x and H_y, with amplitudes H_x^0 and H_y^0, respectively. The directions of H_x and H_y are defined in the images of each structure.

4.3 Propagating Data

Probably the most common measurement in magnetism is to determine the major hysteresis loop of a magnetic material. Figure 4.2a shows one such measurement from a 100 nm-wide Permalloy wire with field H_x applied along the wire length (along the magnetic easy axis) [22]. The sharpness of the transitions indicates that magnetization reversal most likely occurs by domain wall nucleation from one wire end, rapid propagation through the wire, and annihilation at the other wire end.

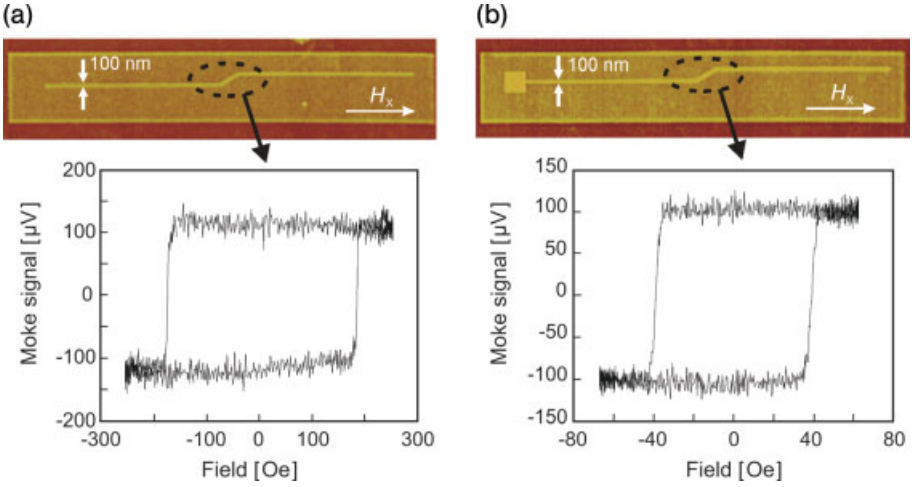


Figure 4.2 Atomic force microscopy images and magneto-optically measured magnetization hysteresis curves from: (a) 100 nm-wide wire and (b) 100 nm-wide wire with a 1 μm × 1 μm square domain wall "injection pad" [22]. Measuring either side of the kink does not change the observed switching field. The direction of the magnetic field H_x is indicated in both images.

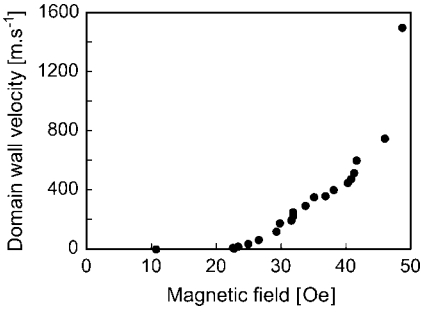


Figure 4.3 Measured domain wall velocity in a 200 nm-wide, 5 nm-thick $Ni_{81}Fe_{19}$ wire as a function of magnetic field along the wire long axis [23].

For this wire, magnetization reversal occurs at a coercive field $H_c = 180$ Oe. Unwanted domain wall nucleation from wire ends, junctions and corners must generally be avoided in logic circuits, as this will corrupt any existing data. It is imperative, therefore, domain walls can be introduced and propagated at magnetic fields lower than, in this case, 180 Oe. Figure 4.2b shows a structure with a similar wire to that in Figure 4.2a, but now with a 1 μm × 1 μm square pad attached to one end. MOKE measurement of the wire shows that $H_c = 39$ Oe. This reduction in H_c is a result of the square pad undergoing magnetization reversal at $H_c = 26$ Oe (not shown) before a domain wall is injected into the wire at $H_x = 39$ Oe. Different regions of the wire all have the same coercive field, even beyond the 30° kink, showing that domain walls can propagate in wires and through changes of wire direction at fields significantly lower than nucleation fields. To quantify this low-field propagation more precisely, the domain wall velocity was measured in a 200 nm-wide Permalloy wire (Figure 4.3), in an experiment described elsewhere [23]. Measured domain wall velocities exceeded 1500 m s^{-1} for certain fields applied along the wire long axis. Other studies [24–27] have shown that domain wall velocity does not increase continually with field, but rather reaches a maximum value before reducing at higher fields. Interestingly, domain wall propagation is still observed at fields as low as 11 Oe [23], albeit with very low velocities of 0.01 m s^{-1}. The data in Figures 4.2 and 4.3 indicate that nanowire devices operating by domain wall propagation will require $11\,\text{Oe} < H_x < 180\,\text{Oe}$, although these field values will change once wire junctions are introduced.

Simply using a unidirectional field will not allow domain walls to be separated reliably and, hence, normal data streams containing both "1"s and "0"s cannot be propagated. Instead, use is made of the orthogonal fields H_x and H_y to create a magnetic field vector that rotates in the plane of the sample to control domain wall propagation around smooth 90° wire corners. An important rule for understanding the nanowire circuits is that domain walls will propagate around corners of the same sense of rotation as the applied field – that is, a clockwise rotating field will lead to domain walls traveling around corners clockwise. In a correctly designed nanowire circuit, the sense of field rotation will define a unique direction of domain wall propagation and, hence, data flow. This is an essential feature of a Boolean logic

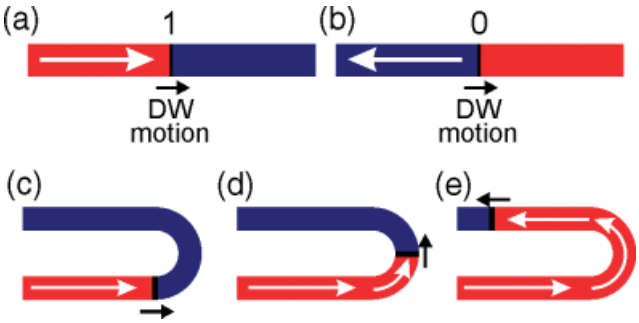


Figure 4.4 Schematic diagrams showing the definition of: (a) logical "1" and (b) logical "0" in magnetic domain wall (DW) logic. Panels (c–e) show sequentially how the wire magnetization changes when a domain wall propagates around a 180° wire corner.

system. Interestingly, the direction of data flow in magnetic domain wall logic can, in principle, be reversed simply by reversing the sense of field rotation.

At this point it must be considered how binary data are represented in magnetic domain wall logic. It was mentioned in Section 4.1 how the two opposite magnetization directions supported in magnetic nanowires can be used to represent the logic states "1" and "0". However, care must be taken here with the definition chosen to use with magnetic nanowire circuits, as the wires can change directions. Figure 4.4a and b show, schematically, two similar magnetic wires with opposite magnetizations containing a single domain wall. A simple approach would be to say that magnetization pointing to the right represents logical "1", and that pointing to the left represents logical "0". However, Figure 4.4c–e shows the magnetization following a domain wall that propagates around a 180° wire corner. In the final situation (Figure 4.4e), the magnetization is continuous up to the domain wall, meaning that there are no changes in logic state up to this point. However, the absolute directions of magnetization are opposite on either side of the turn, and so the simple definition cannot then be valid. Instead, the choice is made to define data representation in terms of the direction of magnetization relative to the direction of domain wall motion. In Figure 4.4c–e, the magnetization following the domain wall is always oriented in the direction of domain wall motion, so the logic state represented remains unchanged. This robust definition allows measurements from different parts of logic circuits to be interpreted correctly.

4.4 Data Processing

The NOT-gate was the first domain wall logic device to be introduced [28–30], and is foundational to the development of all other logic elements. Figure 4.5 shows the geometry of a NOT-gate, and illustrates its principle of operation. The NOT-gate is a junction formed by two wires. For a given field rotation, one wire will act as the input

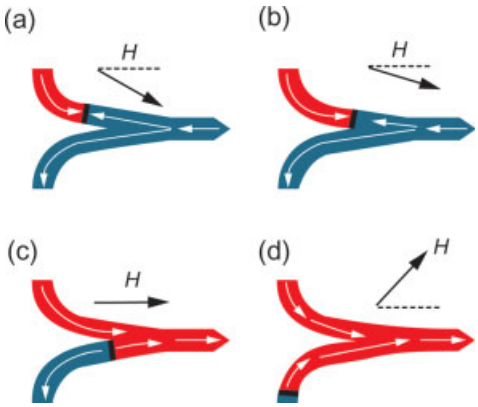


Figure 4.5 Schematic diagrams showing the operating principle of a magnetic nanowire NOT-gate [28]. The black arrows represent the instantaneous magnetic field vector, while the white arrows show the wire's internal magnetization configuration. A domain wall is shown as the black line that separates the oppositely magnetized (red and blue) magnetic domains.

and the other wire as the output. A small central "stub" which emerges from the wire junction is an important part of the device as it ensures there is sufficient shape anisotropy to maintain a magnetization component in the direction in which the stub points. The dimensions for an optimized NOT-gate design are given in Table 4.2. Under a rotating magnetic field, H, a domain wall enters the NOT-gate input wire (Figure 4.5a) before reaching the wire junctions (Figure 4.5b). The magnetization following the domain wall points in the direction of domain wall propagation. Provided that there is sufficient field, the domain wall expands over the junction and splits in two, with one domain wall traveling along the central stub, leaving the stub magnetization reversed, and the other free to propagate on the NOT-gate output wire (Figure 4.5c). As the field continues to rotate, the domain wall in the output wire leaves the NOT-gate. The magnetization following the domain wall is now pointing away from the direction of domain wall motion. The magnetization on either side of the wire junction is reversed, and the device has inverted the input logical state. The reversal in magnetization means that the inversion process would be expected to require a one-half cycle of field.

Figure 4.6a shows a structure containing a NOT-gate fabricated in a square loop of magnetic wire that has not been used to test the operation of the logic device [28–30]. The loop provides feedback to the NOT-gate by joining the output wire to the input wire. Having a single inverter in the loop guarantees that at least one domain wall will be present [28–30] and removes the need, at this stage, for explicit data input. If the principle of operation for a NOT-gate described above is correct, it should be possible to predict the switching period of the NOT-gate/loop structure in terms of field cycles. As mentioned above, a domain wall will take a one-half field cycle to propagate through the NOT-gate junction. Traveling around the 360° wire loop will then add another field cycle. So, magnetization reversal would be expected

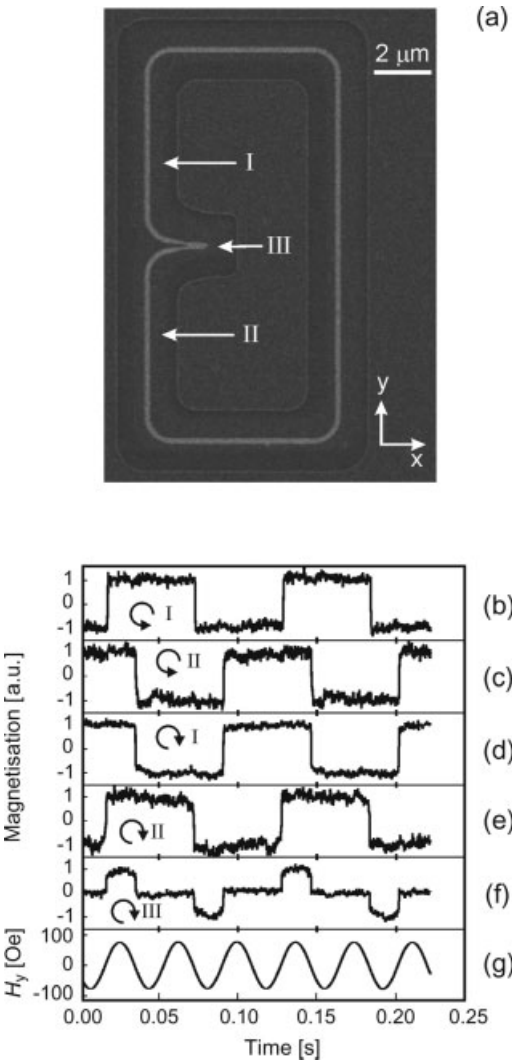


Figure 4.6 (a) Focused ion beam image of a ferromagnetic NOT-gate and feedback loop [29]. Also shown are the *x*- and *y*-directions, and the measurement regions I, II and III where the ∼5 μm diameter magnetometer laser spot was positioned. (b–e) Measured MOKE signals during application of an in-plane rotating magnetic field ($H_x^0 = 77$ Oe, $H_y^0 = 74$ Oe) for anti-clockwise field rotation at measurement position (b) I and (c) II, and clockwise field rotation at measurement position (d) I and (e) II. (f) A measured trace from position III with a clockwise field rotation; (g) the *y*-component of the rotating field.

to be seen at any position in the loop every 3/2 field cycles, giving a switching period of three field cycles.

Figure 4.6b shows a time-averaged MOKE trace obtained from position I on the structure (Figure 4.6a) under anti-clockwise field conditions [29]. As expected, a

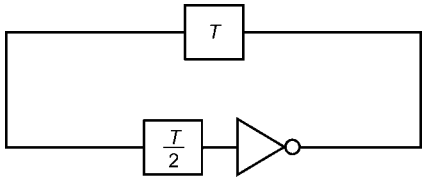


Figure 4.7 An equivalent electronic circuit used to model the ferromagnetic NOT-gate and feedback loop. The square boxes correspond to time delay elements, where T is the periodic time of the applied rotating magnetic field.

three-field cycle switching period was observed, indicating that the wire junction is performing as a NOT-gate. To validate this further, Figure 4.6c shows a MOKE trace obtained from position II, the other side of the NOT-gate (Figure 4.6a), with the same field conditions. This trace is inverted compared to that in Figure 4.6b, except for a one-half field cycle delay, consistent with the operation of a NOT-gate outlined above. As the NOT-gate has an equal number of input and output wires that have identical geometry, it may be operated reversibly. Figure 4.6d and e show measurements from positions I and II, respectively but now under a clockwise-rotating field. The observed phase relationship indicates that position II has now become the input and position I the output. Figure 4.6f shows a MOKE trace that is obtained with the laser spot positioned over the NOT-gate (position III, Figure 4.6a). A domain wall is observed to enter and leave the NOT-gate to correlate with the traces shown in Figure 4.6d and e. An important aspect of this initial demonstration is that the applied field acts as both power supply and clock to the magnetic circuit. The structure in Figure 4.6a can be thought of as having the equivalent electronic circuit shown in Figure 4.7. Electronic invertors do not have a delay in terms of a clock cycle, so a buffer must be introduced with a delay of $T/2$, where T represents a clock period. Another buffer with a delay of T is then introduced within a feedback loop to represent the time spent by a domain wall propagating around the wire loop. The signal from this circuit will replicate those observed in Figure 4.6.

One of the advantages of having input and output wires of identical forms is that logic gates can be directly connected together. Figure 4.8a shows a magnetic shift register circuit made of 11 NOT-gates within a wire loop [28]. The expected switching period can again be calculated by $11 \times 1/2$ field cycles for domain wall propagation through NOT-gates plus one field cycle for the loop to give a magnetization reversal every 6.5 field cycles, or a switching period of 13 field cycles. Figure 4.8b shows the MOKE trace obtained from the structure, in which the 13-field cycle period is clearly observed. The measurement was obtained over 30 min of averaging, indicating that almost 10^5 logical NOT operations were successfully performed. If there was a problem with a domain wall propagating through a NOT-gate on just one occasion, the resultant phase difference introduced would be clearly visible in the time-averaged trace (Figure 4.8b). Although in this case the topology of the shift register structure has been used to ensure the presence of a single domain wall, it will be seen later (in Section 4.5) how similar shift registers can support complex data sequences. The one-half cycle delay for domain wall propagation creates a natural buffer between

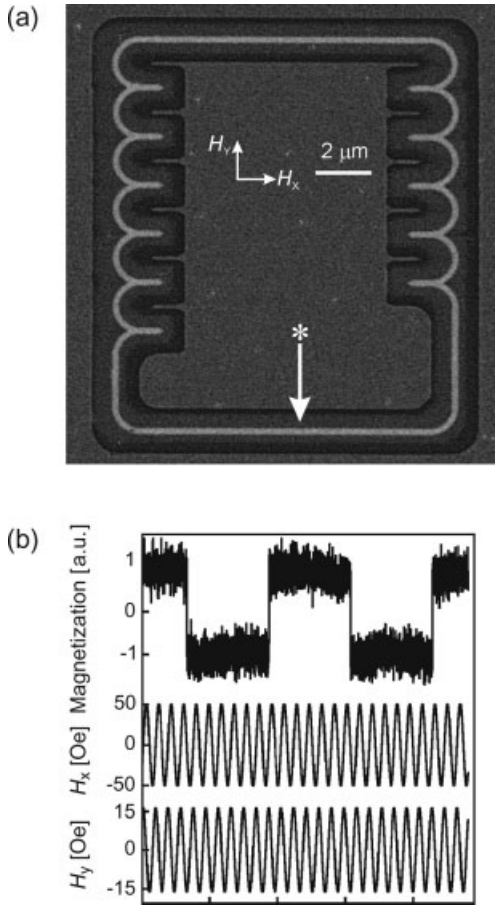


Figure 4.8 (a) Focused ion beam image of a magnetic ring including 11 NOT junctions, where the asterisk indicates the position of MOKE analysis [28]. The directions of the magnetic field components, H_x and H_y, are also indicated. (b) MOKE analysis of an identical structure within a clockwise rotating magnetic field ($H_x^0 = 50$ Oe and $H_y^0 = 15$ Oe).

data bits, removing the need for any complex circuitry, such as the flip-flop circuits that are commonly used in electronic memories.

Characterizing the performance of magnetic domain wall NOT-gates is essential for design optimization [29] and for integrating them with other types of nanowire junction. Here, structures similar to that in Figure 4.6a were used to assess a NOT-gate's operation as a function of in-plane rotating field amplitudes H_x^0 and H_y^0. Three types of operation are observed:

- When the field amplitudes are too low, domain walls experience pinning at the NOT-gate junction, and this leads either to no switching for very low fields or else

de-phasing of the time-averaged MOKE trace when domain walls are pinned even once.

- At high fields, additional domain wall pairs are nucleated in the structure and the magnetization reversal has a single field cycle period.
- At intermediate fields the three-field cycle operation described above is observed.

It should be noted that, after domain wall nucleation is observed at high fields, it is necessary to reduce the number of domain walls back to one by applying field conditions for occasional de-pinning [29]. This allows domain wall pairs to meet and annihilate. Figure 4.9 shows the resulting phase diagram describing NOT-gate operation as a function of field. There are two phase boundaries present, one

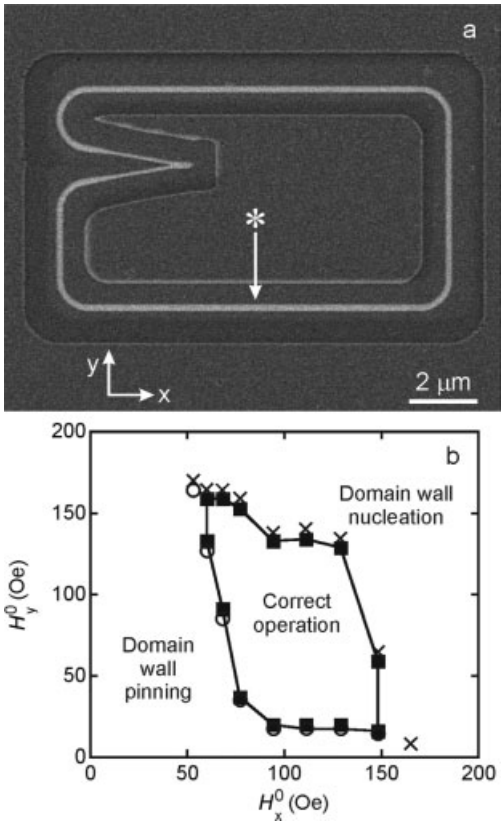


Figure 4.9 (a) Focused ion beam (FIB) image of NOT-gate and feedback loop [29]. The magnetic wire is the light-gray line, and all other features are a result of the FIB milling. Also shown are the measurement position (denoted by "*") and the *x*- and *y*-directions. (b) Experimentally determined phase diagram showing operation of the NOT-gate/feedback loop structure shown in (a) as a function of the rotating magnetic field component amplitudes H_x^0 and H_y^0. × = nucleating; ■ = correct operation; ○ = domain wall pinning. The region bounded by the solid line is the operating region for this device.

separating domain wall pinning from correct operation, and the other separating correct operation from domain wall nucleation. The two boundaries meet to define an area of field phase space in which the NOT-gate operates correctly. Figure 4.9 is taken for a NOT-gate with the optimized dimensions given in Table 4.2.

The other circuit elements that are required for a realistic logic system are a majority gate, signal fan-out, and signal cross-over. The NOT-gate operating phase diagram (Figure 4.9) provides a useful and necessary reference for comparing the performance of these additional elements to ensure compatibility. Figure 4.10a–c shows three structures used for testing the operating fields of majority gate junctions [31]. Each junction has two input wires and one output wire, with the structures having (a) no, (b) one, and (c) two input wires terminated by a domain wall "injection pad". The low field at which domain walls are introduced from an injection pad means that they provide a means of testing majority gate junction operation

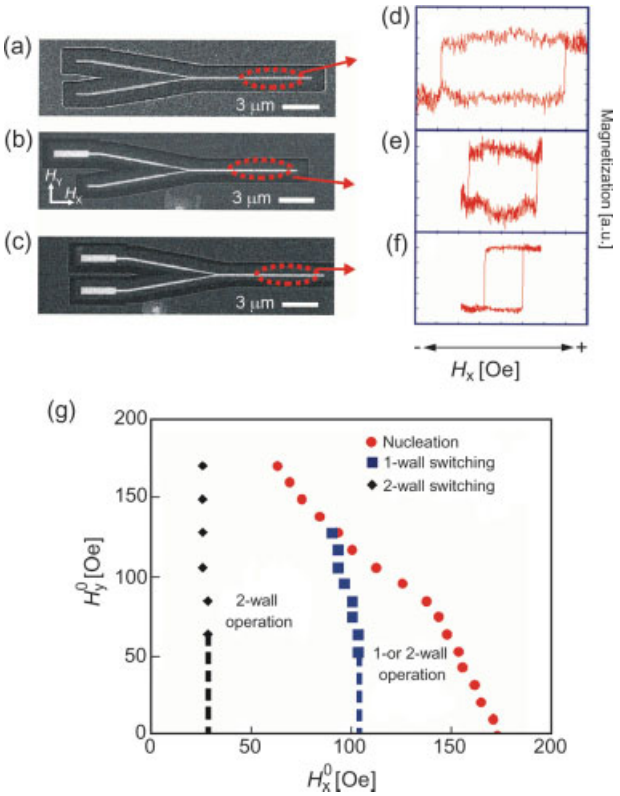


Figure 4.10 (a–c) Focused ion beam images of majority gate test structures with (a) zero, (b) one, and (c) two input wires connected to a 3 μm × 600 nm domain wall "injection pad" [31]. The directions of H_x and H_y are indicated in panel (b). (d–f) MOKE hysteresis loops from the output wires of panels (a)–(c), respectively. (g) Experimentally determined operating phase diagram of the majority gates in an in-plane rotating magnetic field as a function of the field amplitudes H_x^0 and H_y^0.

when 0, 1, or 2 domain walls are present in the input wires. Clearly, the output arm switching fields (Figure 4.10d–f) reduce as the number of domain walls present at the junction increases. In terms of magnetization dynamics, it is interesting that a single domain wall appears capable of expanding across a junction before propagating through the output wire, and that two domain walls are able to interact to enable very low output wire switching fields. Figure 4.10g shows a more detailed analysis of the operation of optimized majority gate junctions (dimensions given in Table 4.2) as a function of the in-plane rotating field amplitudes. The different input conditions now lead to two field-space regions of operation, depending on the number of domain walls present before switching. Crucially, comparison with Figure 4.9 shows that there is overlap between the NOT-gate operating region and both operating regions of the majority gate. The question remains, however, whether to use field amplitudes from the lower-field operating region of the majority gate, or the higher. The answer is to use both. For a majority gate aligned parallel to H_x, field conditions $H_x^0 = 120$ Oe and $H_y^0 = 50$ Oe will mean that the output wire will switch whenever there is just one domain wall present. This corresponds to an input condition of "01" or "10". Clearly, this should not happen for either an AND-gate or an OR-gate. However, by examining the truth tables for each (see Table 4.1), it becomes obvious that a "10" input should always lead to a "0" output for an AND-gate and a "1" output for an OR-gate. However, H_x need not be symmetric; instead, a dc field H_x^{DC} can additionally be applied to bias H_x so that for one sense of H_x the majority gate reverses with a domain wall in either input, while in the other sense of H_x the majority gate requires domain walls in both input wires. The AND/OR function of the gate is then selected by the polarity of H_{dc}.

Signal fan-out and signal cross-over junctions were developed in a similar manner [32], with optimized geometries shown in Table 4.2. Figure 4.11a shows a circuit that integrates all of the structures necessary for performing logic operations: a NOT-gate, a majority gate, two signal fan-out junctions and a signal cross-over element [33]. An anti-clockwise rotating field with amplitudes of $H_x^0 = 75$ Oe and $H_y^0 = 88$ Oe was used with $H_x^{DC} = -5$ Oe (Figure 4.11b) in order to circulate domain walls in an anti-clockwise direction and select logical AND operation for the majority gate. The NOT-gate/loop is similar to those discussed above, and will contain a single domain wall and a magnetization switching period of three field cycles, as before. The difference from Figure 4.11a, however, is that a fan-out structure is incorporated within the loop to split a domain wall each time it is incident on the junction. Part of the domain wall will continue propagating around the loop, while the other part exits the loop to the rest of the circuit. When used in this manner, the NOT-gate/loop acts as a three-field cycle period signal generator for testing the circuit. A domain wall that exits the NOT-gate/loop is then split again at a second fan-out junction. MOKE measurement at position I in Figure 4.11a shows that the three-field cycle period from the NOT-gate loop is preserved through both fan-out junctions (Figure 4.11b, trace I). The domain walls from the second fan-out junction now have separate paths before reaching the AND-gate inputs. The domain wall that passes position I simply has to propagate through two 90° corners and some straight wire sections. The resulting half-field cycle delay between domain walls passing position 1 and arriving at the AND-gate is indicated by trace II in Figure 4.11b. The other domain wall from

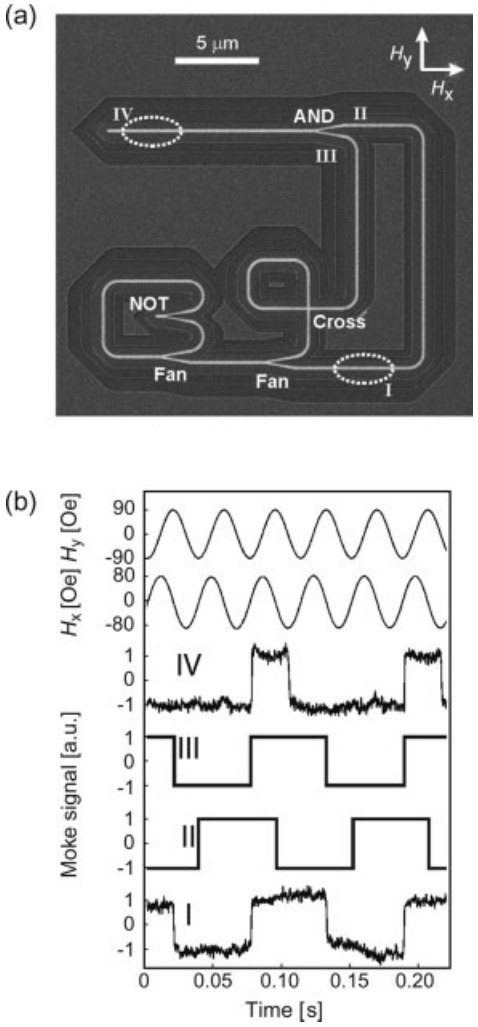


Figure 4.11 (a) Focused ion beam image of a magnetic nanowire circuit containing one NOT-gate, one AND-gate, two fan-out junctions and a cross-over junction [33]. MOKE measurements were made at positions I and IV, while positions II and III denote the inputs to the AND-gate. Also indicated are the directions of field components, H_x and H_y. (b) MOKE traces describing the operation of the magnetic circuit within an anti-clockwise rotating field with $H_x^0 = 75$ Oe, $H_y^0 = 88$ Oe and $H_x^{DC} = -5$ Oe. Experimental MOKE measurements from positions I and IV of the circuit are shown. Traces II and III are inferred from trace I, and show the magnetization state of the AND-gate's input wires.

the fan-out junction has to negotiate a cross-over junction and an additional 360° loop before arriving at an AND-gate input at position III (Figure 4.11a). The inclusion of the loop tests the operation of the cross-over element and will create a one-field cycle delay between domain walls arriving at positions II and III, as indicated in the

inferred trace III in Figure 4.11b. Measurement at position IV in Figure 4.11a shows that the output is high only when both inputs are high, showing that the majority gate is operating correctly as an AND-gate. Furthermore, this demonstrates that all four of the element types can operate under identical field conditions in a single circuit.

4.5 Data Writing and Erasing

In the previous section, domain walls were introduced to nanowire junctions either by using topological constraints of a nanowire circuit or domain wall injection from a large area pad. These are both valid methods for developing logic elements, although a method of entering user-defined data is still required to create a viable logic system. Here, an element for data input is presented that is integrated with a domain wall shift register [33]. Furthermore, it is shown how data can be deleted from the shift register.

The design of the optimized data input element is shown in Table 4.2. Figure 4.12 shows the operating phase diagram of this element, obtained from simple test structures, overlaid with that of a NOT-gate. A single phase boundary for the data input element bisects the NOT-gate field operating area. Above the phase boundary, a domain wall is nucleated from the data input element, whereas below the phase boundary no magnetic reversal occurs. Two sets of field amplitudes can then be identified for operating both NOT-gates and the data input element. Below the data input element phase boundary are the *read* or *no-write* field conditions of

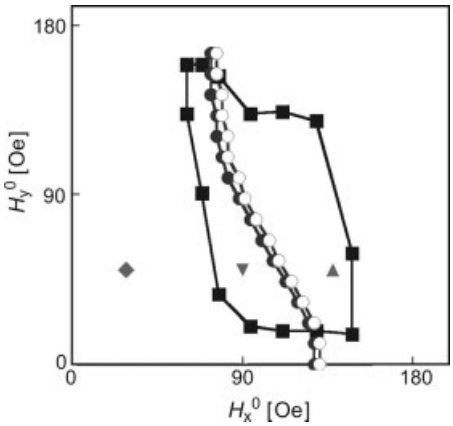


Figure 4.12 Operating field phase diagram of optimized NOT-gate and data input elements. Symbols represent the limits of the NOT-gate operating region (■), maximum field for no domain wall injection (●) and minimum field for reliable domain wall injection (○) from the data input element, and selected *write* field (▲), *read*/*no-write* field (▼) and *erase* field (◆) conditions. The lines are provided only as guides to the eye.

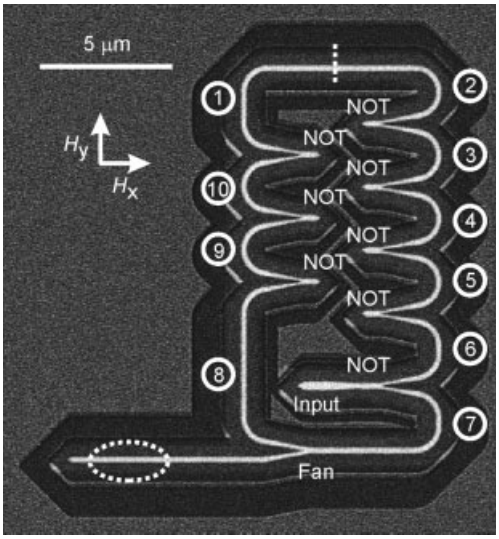


Figure 4.13 Focused ion beam image of a continuous shift register made of eight NOT-gates, a fan-out junction and a data input element connected to the central wire of one of the NOT-gates [33]. The shift register is divided into ten labeled cells, each separated from its neighbors either by a wire junction (NOT or fan-out) or a straight horizontal wire (indicated by dotted line). The field directions H_x and H_y are shown; the position of magneto-optical measurement is indicated by the dotted ellipse.

$H_x^{\text{no-write}} = 90$ Oe and $H_y^0 = 50$ Oe, and above the phase boundary are the *write* field conditions of $H_x^{\text{write}} = 138$ Oe and $H_y^0 = 50$ Oe (Figure 4.12).

Figure 4.13 shows an image of a shift register containing eight NOT-gates and one fan-out junction [33]. In addition, one NOT-gate has a data input element attached to its central stub. The fan-out element provides a monitor arm for MOKE measurement, as used in Section 4.4 above. The shift register can be divided into ten cells, each capable of holding a single domain wall and separated from its neighbors by a total of 180° of wire turn. Due to topological restrictions, domain walls can only be introduced or removed in pairs. Therefore, a data bit is represented by the presence or absence of a domain wall pair, so the shift register in Figure 4.13 contains five data bits.

Figure 4.14a–d shows, schematically, the operating principle of the data input element connected to the NOT-gate [33]. Initially, no domain walls are present and the two connecting wires to the NOT-gate have opposite magnetizations (Figure 4.14a). As the field rotates, the *write* field amplitude is used (Figure 4.14b) so that a domain wall is nucleated at the end of the data input element. This domain wall will propagate to the NOT-gate junction, where it will split into domain walls DW 1 and DW 2, one in each of the input/output wires (Figure 4.14b). As the field rotates further (Figure 4.14c), both domain walls follow the field rotation and propagate clockwise around corners. DW 1 propagates away from the NOT-gate, while DW 2 returns to the junction (Figure 4.14c). Finally, the field rotates to be oriented 180° from when nucleation occurred, but now with *no-write* conditions (Figure 4.14d). DW 1 has

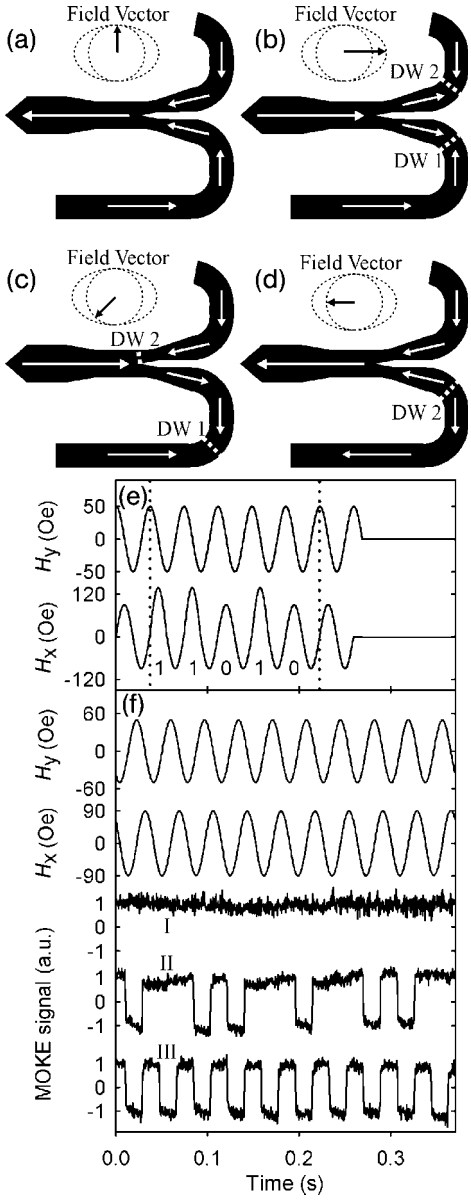


Figure 4.14 (a–d) Schematic diagrams describing the operation of a data input element [33], including instantaneous field vectors (black arrows), magnetization directions (white arrows) and position of domain walls (white dotted line). (e) *Write* field pattern with field amplitudes $H_x^{\text{no-write}} = 90$ Oe, $H_x^{\text{write}} = 138$ Oe and $H_y^0 = 50$ Oe. The five-bit sequence "11010" is generated during the interval between the dotted lines. (f) MOKE measurements ($H_x^0 = 90$ Oe and $H_y^0 = 50$ Oe) from the shift register in Figure 4.13 in reset configuration (trace I), after applying the *write* field pattern shown in (e) (trace II), and after a 1.85-ms duration half-sinusoid field pulse of amplitude $H_x^0 = 234$ Oe (trace III).

propagated out of the section shown in Figure 4.14, while DW 2 has propagated through the NOT-gate, to leave the NOT-gate and data input element's magnetization back in their initial configuration (Figure 4.14d). One single half-cycle of *write* field conditions has created a pair of domain walls – that is, a single data bit has been written.

Figure 4.14e shows a field sequence that is used to write data to the shift register in Figure 4.13. A combination of *write* and *no-write* field conditions is used to write the five-bit data sequence "11010". Time-averaged MOKE measurements were performed during continual application of the *read* field conditions. Trace I in Figure 4.14f shows the MOKE signal obtained prior to the application of the *write* field sequence in Figure 4.14e. No transitions are observed, meaning that no domain walls are present. After a single application of the *write* field sequence, the MOKE signal changes to show that pairs of domain walls are propagating continuously around the shift register (Figure 4.14f, trace II). Crucially, the pattern of domain wall pairs matches the original input data sequence of "11010", although the phase of the measurement is such that the MOKE trace starts part-way through this sequence. Note that in this case logical "1" is represented by a low MOKE signal, due to the 180° wire turn between the data input element and the measurement position. This observation confirms the principle of operation for a data input element outlined above. Delays of an hour between writing and successfully reading data have been seen, demonstrating the intrinsic non-volatility of the data storage. The whole shift register can be filled with domain walls, destroying any data present, by applying an *over-write* half-sinusoid field pulse of amplitude $H_x^0 = 243$ Oe and 1.85 ms pulse length (Figure 4.14f, trace III).

Individual domain wall pairs can also be removed from the shift register in Figure 4.13. This represents a selective bitwise delete operation. Almost all of the ten cells shown in Figure 4.13 are separated from their neighbors by a nanowire junction. The exceptions are cells 1 and 2, which are separated by a straight section of wire. Domain walls require *read* field conditions to propagate successfully through the shift register. However, when *erase* field conditions of $H_x^{\text{erase}} = 24$ Oe and $H_y^0 = 50$ Oe (see Figure 4.12) are used, domain walls cannot overcome the pinning potentials associated with the nanowire junctions. The only possible domain wall motion will be between cells 1 and 2, where there are no wire junctions. Figure 4.15a shows the field sequence for erasing a single pair of domain walls. The first half-cycle has *erase* field conditions, so the only domain wall propagation will be from cell 1 to cell 2. All other domain walls will remain pinned at the junctions between cells. The next half-cycle has *read* field amplitudes, so all domain walls will propagate forward by one cell, with the exception of the pair of domain walls in cell 2 which will meet and annihilate. The second full field cycle continues with *read* field conditions to move all domain walls on by two cells and allowing the field sequence to be repeated on the next domain wall pair. Figure 4.15b shows MOKE traces obtained from the shift register following an *over-write* half-sinusoid pulse and between 0 and 5 erase field sequences (Figure 4.15a). The MOKE traces have a five-cycle period and each erase sequence removes a pair of domain walls, validating the operating principle described above.

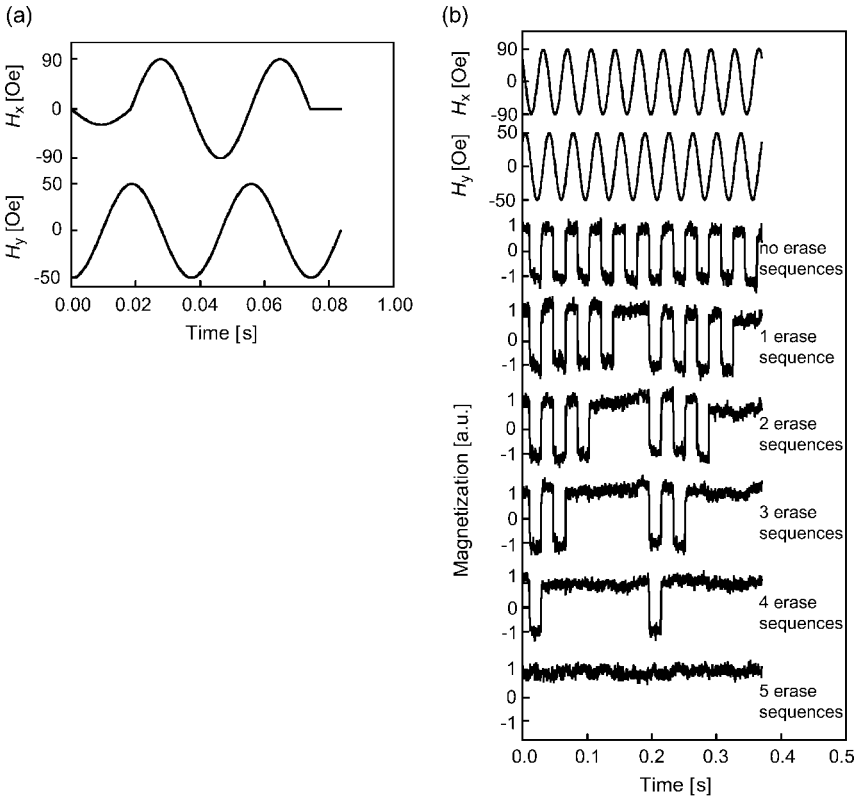


Figure 4.15 (a) *Erase* field sequence ($H_x^0 = 90$ Oe, $H_x^{erase} = 24$ Oe, $H_y^0 = 50$ Oe) applied to the structure in Figure 4.13. (b) *Read* field sequence $H_x^0 = 90$ Oe, $H_y^0 = 50$ Oe) and MOKE signals measured from the structure in Figure 4.13 following a saturating pulse and between zero to five 5 *erase* field sequences, as indicated.

4.6 Outlook and Conclusions

Domain wall logic is not a contender for a wholesale replacement of CMOS microelectronics. CMOS is a highly mature technology with many advantages, and still has many years of scaling available to it. The limited operational speed of domain wall logic does not render it suitable for many applications. However, a strong trend in microelectronics which is expected to apply to the relationship between CMOS and to many other areas of nanotechnology in the future, is to combine multiple technologies on a single platform: the System on Chip (SoC).

So – what does domain wall logic do well? First, it provides a high level of functionality to relatively simple structures. To implement an AND gate in CMOS would take six transistors, but domain wall logic achieves this simply by bringing

two nanowires together. Similarly, the other high-level properties that have been highlighted in this chapter – such as input–output isolation and signal/power gain – are all intrinsic to the nanowire and do not have to be explicitly created.

The power dissipation per logic gate is extremely low. Microelectronic engineers usually measure dissipation from a gate by the power–delay product; that is to say, the product of how much power is dissipated multiplied by how long the gate takes to process a single function. The units of this quantity are energy, corresponding to the energy dissipated during the evaluation of the function performed by the gate. The power–delay product of CMOS depends on the size of the devices. Hence, in order to compare like with like, a 200 nm minimum feature size CMOS value of 10^{-2} pJ is considered [34]. On very general magnetic grounds, it can be said that an upper bound for the power–delay product for domain wall logic is $2M_sVH$, where M_s is the saturation magnetization of the magnetic material, V is the volume of magnetic material in a gate, and H is the amplitude of the applied field. Applying the parameters for a typical 200 nm domain wall logic gate gives 10^{-5} pJ – that is, 1000 times lower than the equivalent CMOS device. Because of the inefficiencies inherent in the generation of high-speed magnetic fields (see above), this does not necessarily mean that domain wall logic chips will not consume much power. What it does mean, however, is that the waste heat will be generated from the global field generator and not from the logic devices themselves. This is of particular relevance if the devices are to be stacked into three-dimensional (3-D), neural-like circuits. The two key technical difficulties in doing this in CMOS are: (i) distributing the power and clock to everywhere inside the volume of network; and (ii) extracting the waste heat from the center of the network so that the device does not melt. It is believed that domain wall logic is an excellent choice of primitive for 3-D architectures.

Non-volatility comes as standard. In a world of mobile computing and portable (or even wearable) devices, the concept of "instant-on" is becoming increasingly important. Users accept that devices cannot be expected to operate when there is no power, but as soon as the power becomes available they want the device to be ready, and not have to undergo a long boot process, or to have forgotten what it was doing when the power last failed. As there are currently very few non-volatile memory technologies available which can be embedded directly into CMOS, a data transfer process is usually required between a high-speed, volatile memory register in the heart of the CMOS logic and an off-chip, low-speed, non-volatile store where the state variables of the system are stored. With domain wall logic, all of this becomes redundant. Provided that the rotating field is properly controlled so that it stops gracefully as power fails, and does not apply intermediate levels of field leading to data corruption, the domain wall logic circuit should simply stop and retain all of its state variables. Then, as soon as the power returns, the logic continues from where it left off.

Domain wall logic can make use of redundant space on top of CMOS. Because no complex heterostructures are required, the logic elements can sit in a single layer fabricated as a Back End Of Line process after the CMOS has been laid down. This can improve the efficiency of the underlying CMOS by farming out some space-consuming task to the domain wall logic on top. As this space was never accessible to CMOS itself anyway, it all counts as a gain.

Being metals, the basic computational elements of domain wall logic are automatically radiation-hard, and so are suitable for use in either space or military applications.

Domain wall logic is very good at forming high-density shift registers that could be used as non-volatile serial memory for storing entire files, and so would not require high-speed random access. The hard disk drive and NAND Flash devices – for example, as used to store photographs in a digital camera – are examples of non-volatile serial memory. At present, both of these devices are 2-D in form, but registers made from domain wall logic elements have the potential to be stacked into three dimensions, without incurring extra wiring complexity, as the data and power can be transmitted remotely through magnetic fields (see above). In a hard disk drive the data are stored as rows of magnetic domains, and this would remain the same in a domain wall logic serial memory. What would differ is that, in a hard disk, the domains are mechanically rotated on their disk underneath a static sensor, whereas in domain wall logic the domains themselves would move under the action of an externally applied magnetic field along static domain wall conduits, potentially stacked into an ultrahigh-density, 3-D array.

Acknowledgments

The research studies described in this chapter were funded by the European Community under the Sixth Framework Programme Contract Number 510993: MAGLOG. The views expressed are solely those of the authors, and the other Contractors and/or the European Community cannot be held liable for any use that may be made of the information contained herein. D.A.A. acknowledges the support of an EPSRC Advanced Research Fellowship (GR/T02942/01).

References

1 T. Dietl, H. Ohno, F. Matsukura, J. Cibert, D. Ferrand, *Science* 2000, **287**, 1019.

2 H. Ohno, D. Chiba, F. Matsukura, T. Omiya, E. Abe, T. Dietl, Y. Ohno, K. Ohtani, *Nature* 2000, **408**, 944.

3 Y. Ohno, D. K. Young, B. Beschoten, F. Matsukura, H. Ohno, D. D. Awschalom, *Nature* 1999, **402**, 790.

4 G. A. Prinz, *Science* 1998, **282**, 1660.

5 S. A. Wolf, D. D. Awschalom, R. A. Buhrman, J. M. Daughton, S. von Molnar, M. L. Roukes, A. Y. Chtchelkanova, D. M. Treger, *Science* 2001, **294**, 1488.

6 S. A. Wolf, D. Treger, A. Chtchelkanova, *MRS Bulletin* 2006, **31**, 400.

7 R. W. Dave, G. Steiner, J. M. Slaughter, J. J. Sun, B. Craigo, S. Pietambaram, K. Smith, G. Grynkenich, M. DeHerrera, J. Åkerman, S. Tehrani, *IEEE Trans. Magn.* 2006, **42**, 1935.

8 R. Richter, L. Bar, J. Wecker, G. Reiss, *Appl. Phys. Lett.* 2002, **80**, 1291.

9 A. Ney, C. Pampuch, R. Koch, K. H. Ploog, *Nature* 2003, **425**, 485.

10 C. Pampuch, A. Ney, R. Koch, *Europhys. Lett.* 2004, **66**, 895.

11 G. Reiss, H. Brückl, A. Hütton, H. Koop, D. Meyners, A. Thomas, S. Kämmerer, J. Schmalhorst, M. Brzeska, *Phys. Stat. Sol. A* 2004, **201**, 1628.

12 D. Meyners, K. Rott, H. Brückl, G. Reiss, J. Wecker, *J. Appl. Phys.* 2006, **99**, 023907.

13 W. C. Black, B. Das, *J. Appl. Phys.* 2000, **87**, 6674.

14 R. P. Cowburn, M. E. Welland, *Science* 2000, **287**, 1466.

15 R. P. Cowburn, *Phys. Rev. B* 2002, **65**, 092409.

16 M. C. B. Parish, M. Forshaw, *Appl. Phys. Lett.* 2003, **83**, 2046.

17 A. Imre, G. Csaba, L. Ji, A. Orlov, G. H. Bernstein, W. Porod, *Science* 2006, **311**, 205.

18 G. Xiong, D. A. Allwood, M. D. Cooke, R. P. Cowburn, *Appl. Phys. Lett.* 2001, **79**, 3461.

19 D. Ozkaya, R. M. Langford, W. L. Chan, A. K. Petford-Long, *J. Appl. Phys.* 2002, **91**, 9937.

20 A. Hubert, R. Schäfer, *Magnetic Domains. The Analysis of Magnetic Microstructures*, Springer-Verlag, Berlin, 1998.

21 D. A. Allwood, G. Xiong, M. D. Cooke, R. P. Cowburn, *J. Phys. D Appl. Phys.* 2003, **36**, 2175.

22 R. P. Cowburn, D. A. Allwood, G. Xiong, M. D. Cooke, *J. Appl. Phys.* 2002, **91**, 6949.

23 D. Atkinson, D. A. Allwood, G. Xiong, M. D. Cooke, C. C. Faulkner, R. P. Cowburn, *Nature Mater.* 2003, **2**, 85.

24 Y. Nakatani, A. Thiaville, J. Miltat, *Nature Mater.* 2003, **2**, 521.

25 Y. Nakatani, A. Thiaville, J. Miltat, *J. Magn. Magn. Mater.* 2005, **290–291**, 750.

26 D. G. Porter, M. J. Donahue, *J. Appl. Phys.* 2004, **95**, 6729.

27 G. S. D. Beach, C. Nistor, C. Knutson, M. Tsoi, J. L. Erskine, *Nature Mater.* 2005, **4**, 741.

28 D. A. Allwood, G. Xiong, M. D. Cooke, C. C. Faulkner, D. Atkinson, N. Vernier, R. P. Cowburn, *Science* 2002, **296**, 2003.

29 D. A. Allwood, G. Xiong, M. D. Cooke, C. C. Faulkner, D. Atkinson, R. P. Cowburn, *J. Appl. Phys.* 2004, **95**, 8264.

30 X. Zhu, D. A. Allwood, G. Xiong, R. P. Cowburn, P. Grütter, *Appl. Phys. Lett.* 2005, **87**, 062503.

31 C. C. Faulkner, D. A. Allwood, M. D. Cooke, G. Xiong, D. Atkinson, R. P. Cowburn, *IEEE Trans. Magn.* 2003, **39**, 2860.

32 D. A. Allwood, G. Xiong, R. P. Cowburn, *J. Appl. Phys.* 2007, **101**, 024308.

33 D. A. Allwood, G. Xiong, C. C. Faulkner, D. Atkinson, D. Petit, R. P. Cowburn, *Science* 2005, **309**, 1688.

34 R. Waser, *Nanoelectronics and Information Technology*, Wiley VCH, Weinheim, 2003.

5
Monolithic and Hybrid Spintronics

Supriyo Bandyopadhyay

5.1
Introduction

An electron has three attributes: mass; charge; and spin. An electron's mass is too small to be useful for practical applications, but the charge is an enormously useful quantity that is utilized universally in every electronic device extant. The third attribute – spin – has played mostly a passive role in such gadgets as magnetic disks and magneto-electronic devices, where its role has been to affect the magnetic or the electrical properties in useful ways – for example, in the giant magnetoresistance devices used to read data stored in the magnetic disks of laptops and Apple iPods. Only recently has a conscious effort been made to utilize spin – either singly or in conjunction with the charge degree of freedom – to store, process, and transmit information. This field is referred to as modern "spintronics".

There are two distinct branches of spintronics:

- *Hybrid spintronics:* these devices are very much conventional electronic devices, as information is still encoded in the charge (ultimately detected as voltage or current), but spin augments the functionality of the device and *may* improve device performance. Examples of hybrid spintronic devices are *spin field effect transistors* (SPINFETs) [1] and *spin bipolar junction transistors* (SBJTs) [2], where information is still processed by modulating the charge current flowing between two terminals via the application of either a voltage or a current to the third terminal. The process by which the third terminal controls the voltage or current is spin-mediated – hence the term "spin transistors".
- *Monolithic spintronics:* here, charge has no direct role whatsoever. Rather, the information is encoded entirely in the spin polarization of an electron, which may be made to have only two *stable* values: "upspin" and "downspin", by placing the electron in a static magnetic field. "Upspin" will correspond to polarizations anti-parallel to the magnetic field, while "downspin" will be parallel. These two

Nanotechnology. Volume 4: Information Technology II. Edited by Rainer Waser

ISBN: 978-3-527-31737-0

polarizations can encode binary bits 0 and 1 for digital applications. Toggling a bit merely requires flipping the spin, *without any physical movement of charge*. It has recently been argued that as no charge motion (or current flow) is required, there can be tremendous energy savings during switching [3]. As a result, monolithic spintronic devices are far more likely to yield low-power signal processing units than are hybrid spintronic devices. An example of monolithic spintronic devices is the Single Spin Logic (SSL) paradigm that is described in Section 5.3.

In this chapter, the two most popular hybrid spintronic devices – the SPINFET and the SBJT – will be described, and evidence provided that neither device is likely to produce significant advantages in terms of speed or power dissipation over conventional charge-based transistors. The concept of single spin logic (SSL) will then be discussed, and its significant advantages in power dissipation over SPINFET or SBJT outlined. SSL also has significant advantages over any charge-based paradigm where charge, rather than spin, is used as the state variable to encode information. Finally, it will be shown that the *maximum* energy dissipation in switching a bit in SSL is the Landauer–Shannon limit of $kT\ln(p)$ per bit operation, where $1/p$ is the bit error probability. Some gate operations dissipate less energy than this because of *interactions* between spins, which may reduce dissipation [4], because many spins function collectively, as a single unit, to effect gate operation. This collective, cooperative dynamics is conducive to energy efficiency. Any discussion of adiabatic switching [5], which can reduce energy dissipation even further, is avoided as it is very slow, error-prone, and therefore impractical. The discussion of devices in non-equilibrium statistical distribution, where energy dissipation can be reduced below the Landauer–Shannon limit [6] is also avoided, simply because energy is required to maintain the non-equilibrium distributions over time, and that energy must be dissipated in the long term. The final section includes a very brief engineer's perspective on spin-based quantum computing (included at the request of the editor).

5.2 Hybrid Spintronics

Hybrid spintronic devices are those where spin is used to "enhance" the performance of charge but does not itself play a direct role in storing, processing or communicating information. The two most popular hybrid spintronic devices are the SPINFET and the SBJT.

5.2.1 The Spin Field Effect Transistor (SPINFET)

A schematic representation of the SPINFET, as proposed in the seminal studies of Ref. [1], is shown in Figure 5.1a. This device exactly resembles a conventional metal-

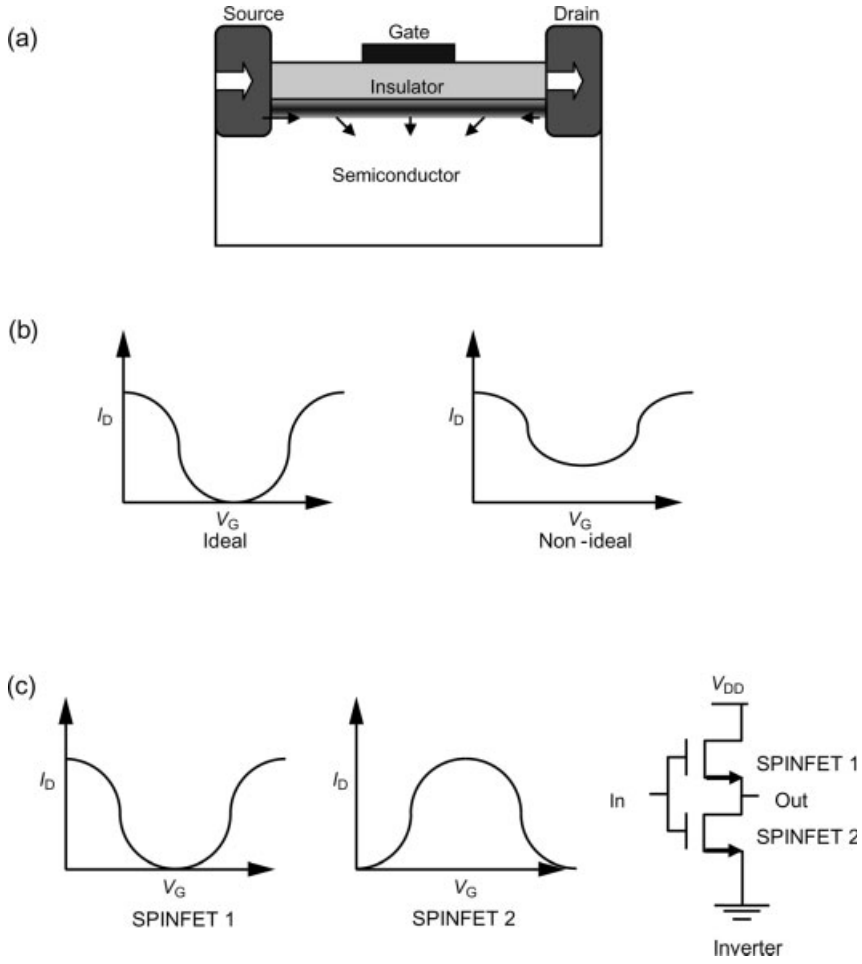


Figure 5.1 (a) Schematic of a spin field effect transistor (SPINFET). (b) Ideal and non-ideal transfer characteristic of a SPINFET. (c) Transfer characteristics of two SPINFETs with different threshold shifts and realization of a CMOS-analog inverter by connecting these two SPINFETs in series.

oxide-semiconductor-field-effect-transistor (MOSFET), except that the source and drain contacts are ferromagnetic. It will be assumed that the channel is strictly one-dimensional (1-D) (quantum wire), and only the lowest transverse subband is occupied by electrons. Both, source and drain contacts are magnetized so that their magnetic moments are parallel and point along the direction of current flow ($+x$-direction). As a result, when the source-to-drain voltage is turned on, the ferromagnetic source injects carriers into the channel with $+x$ polarized spins (the majority spins in the ferromagnet). It will also be assumed that the spin injection efficiency is 100% so that *only* majority spins ($+x$-polarized spins) are injected

from the source contact, and absolutely no minority spin (i.e. $-x$-polarized spin) is injected. Immediately after injection into the channel, all spins are polarized along the $+x$ direction. When the gate voltage is switched on, it induces an electric field in the y-direction that causes a Rashba spin–orbit interaction [7] in the channel. This spin–orbit interaction acts like an effective magnetic field in the z-direction (which is the direction mutually perpendicular to the electron's velocity in the channel and the gate electric field). This pseudo-magnetic field B_{Rashba} causes the spins to precess in the x-y plane, as they travel towards the drain. The angular frequency of spin precession (which is essentially the Larmor frequency) is given by $\Omega = eB_{\text{Rashba}}/m^*$, where e is the electronic charge and m^* is the effective mass of the carrier. The pseudo-magnetic field B_{Rashba} depends on the magnitude of the gate voltage and the carrier velocity along the channel according to

$$B_{\text{Rashba}} = \frac{2(m^*)^2 a_{46}}{e\hbar^2} E_y v_x \tag{5.1}$$

where a_{46} is a material constant, E_y is the gate electric field, and v_x is the electron velocity.[1)]

The spatial rate of spin precession is

$$\frac{d\phi}{dx} = \frac{d\phi}{dt}\frac{1}{\frac{dx}{dt}} = \Omega/v_x = 2\frac{m^* a_{46}}{\hbar^2} E_y \tag{5.2}$$

which is *independent* of the carrier velocity and depends only on the gate voltage (or gate electric field). The total angle by which the spins precess in the x-y plane as they travel through the channel from source to drain is

$$\Phi = \frac{2m^* a_{46}}{\hbar^2} E_y L \tag{5.3}$$

where L is the channel length. This angle is independent of the carrier velocity and therefore is the *same* for every electron, regardless of its initial velocity or scattering history in the channel. If the gate voltage (and E_y) is of such magnitude that Φ is an odd multiple of π, then *every* electron has its spin polarization anti-parallel to the drain's magnetization when it arrives at the drain. These electrons are blocked by the drain, and therefore the source to drain current falls to zero. Here, it has been assumed that the drain is a perfect spin filter that allows only majority spins to transmit, while completely blocking the minority spins. Without a gate voltage, the

1) Some authors assume incorrectly that the Rashba field B_{Rashba} is proportional to wavevector k_x and not the velocity v_x. This makes a difference since, in the presence of the Rashba interaction, $v_x = \hbar k_x/m* \pm \eta/\hbar$, where η is the strength of the Rashba interaction. Following the derivation in this chapter, the reader can easily convince herself/himself that the SPINFET would not work as claimed if B_{Rashba} were proportional to wavevector k_x and not the velocity v_x. As the magnetic field associated with spin–orbit interaction is proportional to $\frac{\vec{v} \times \vec{E}}{2c^2}$ (where c is the speed of light in vacuum and $\vec{E}$ is the electric field seen by the electron), it is obvious that B_{Rashba} should be proportional to v_x and not to k_x.

spins do not precess[2)] and the source to drain current is non-zero. Thus, the gate voltage causes current modulation via spin precession and this realizes transistor action.[3)] This device is also briefly discussed in Chapter 3 in Volume III of this series.

It should be clear that (in this device) although "spin" plays the central role in current modulation, it plays no direct role in information handling. Information is still encoded in "charge" which carries the current from the source to the drain. The transistor is switched between the "on" and "off" states by changing the current with the gate potential, or by controlling the motion of charges in space. The role of spin is only to provide an alternate means of changing the current with the gate voltage. Thus, this device is a quintessential hybrid spintronic device.

5.2.1.1 The Effect of Non-Idealities

The operation of the SPINFET described above is an idealized description. In a real device, there will be many non-idealities. First, there will be a magnetic field along the channel because of the magnetized ferromagnetic contacts. This will cause problems, as it will add to B_{Rashba} and the total effective magnetic field will be $|\vec{B}_{\text{Rashba}} + \vec{B}_{\text{channel}}|$, which is no longer linearly proportional to carrier velocity v_x. As a result, the precession rate in space will no longer be given by Eq. (5.2)[4)] and will *not* be independent of the carrier velocity (or energy). Therefore, at a finite temperature, different electrons having different velocities due to the thermal spread in carrier energy, or because of different scattering history, will suffer different amounts of precession Φ. As a result, when the current drops to a minimum, not all spins at the drain end will have their polarizations exactly anti-parallel to the drain's magnetization. Those that do not, will be transmitted by the drain and contribute to a *leakage current* in the off state [8]. This is extremely undesirable as it decreases the ratio of on- to off-current and will lead to standby power dissipation when the device is off.

A more serious problem is that the magnetic field changes the energy dispersion relations in the channel. In Figure 5.2, the energy dispersion relation (energy versus wavevector) is shown schematically with and without the channel magnetic field [9]. Without the magnetic field, the Rashba interaction lifts the spin degeneracy at any non-zero wavevector, but each spin-split band still has a fixed spin quantization axis (meaning that the spin polarization in each band is always the same and independent of wavevector) (Figure 5.2a). The spin polarizations in the two bands are anti-parallel and the eigenspinors in the two bands are orthogonal. Because of this orthogonality, there can be no scattering between the two bands. Electrons can scatter elastically

2) Even without the gate voltage, there is obviously some Rashba interaction in the channel due to the electric field associated with the hetero-interface. This field exists because the structure lacks inversion symmetry along the direction perpendicular to the hetero-interface. This vestigial interaction will cause some spin precession even at zero gate voltage, but this effect is simply equivalent to causing a fixed threshold shift.

3) The device described here is a "normally on" device. If the magnetizations of the source and drain are anti-parallel instead of parallel, or if the spin polarizations in the source and drain contacts have *opposite* signs (e.g. iron and cobalt), then the device will be a "normally off" device.

4) In Eq. (5.2), B_{Rashba} must be replaced by $|\vec{B}_{\text{Rashba}} + \vec{B}_{\text{channel}}|$ in Ω.

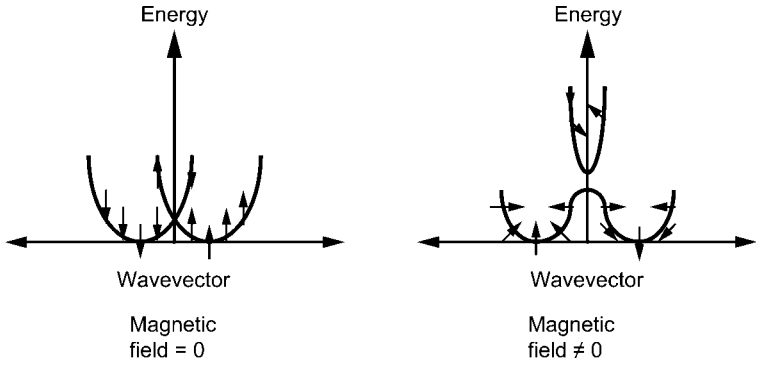


Figure 5.2 Schematic energy dispersion relationships for electrons in the channel of a one-dimensional SPINFET channel, with and without an axial magnetic field. The arrowheads indicate the spin polarization of a carrier in the corresponding wavevector state.

or inelastically only within the same band, but this does not alter the spin polarization since every state in the same band has exactly the same spin polarization. However, if a magnetic field is present in the channel, then the spin polarizations in both bands become *wavevector-dependent* and neither subband has a fixed spin polarization. Two states in two subbands with different wavevectors[5)] will have different spin polarizations that are not completely anti-parallel (orthogonal). Therefore, the matrix element for scattering between them is non-zero, which means that there is finite probability that an electron can scatter between them. Therefore, any momentum randomizing scattering event (due to interactions with impurities or phonons) will rotate the spin as the initial and final states have different spin polarizations. This rotation is random in time or space as the scattering event is random; therefore, it will cause spin relaxation. This is a new type of spin relaxation, and it is introduced solely by the channel magnetic field [10]. It is similar to the Elliott–Yafet spin relaxation mechanism [11] in the sense that it is associated with momentum relaxation. Any such spin relaxation in the channel will randomize the spin polarizations of electrons arriving at the drain and thus give rise to a significant leakage current. Therefore, the channel magnetic field causes leakage current in two different ways, both of which are harmful.

In Ref. [1], where the ideal SPINFET was analyzed, it was assumed that there is no spin relaxation in the channel. The transfer characteristic shown Figure 5.1b, which shows zero leakage drain current in the OFF-state, is predicated on this assumption. A question might arise as to whether the usual spin relaxation mechanisms are operative in the channel even without a channel magnetic field. For the ideal SPINFET, the answer is in the negative. The two spin relaxation mechanisms of concern are the Elliott–Yafet mode [11] and the D'yakonov–Perel' mode [12]. The

5) Eigenspinors in the two bands having the same wavevector are still orthogonal.

former is absent if the eigenspinors are wavevector-independent (as is the case with the ideal SPINFET), and the D'yakonov–Perel' mode is absent if carriers occupy only a single subband [13]. Thus, in the ideal 1-D SPINFET, there can be no significant spin relaxation (even if there is scattering due to interactions with non-magnetic impurities and phonons). If any spin relaxation occurs, it will be due to hyperfine interactions with nuclear spins. Since such interactions are very weak, they can be ignored for the most part. However, if there is an axial magnetic field in the channel, then all this changes and scattering with non-magnetic impurities or phonons will cause spin relaxation (and therefore a large leakage current). Consequently, it is extremely important to eliminate the channel magnetic field.

There are two ways to eliminate (or reduce) the channel magnetic field. One way is to magnetize the contacts in the y-direction instead of the x-direction. Since B_{Rashba} is in the z-direction, it makes no difference as to whether the spins are initially polarized in the x- or y-direction, as they precess in the x-y plane. The SPINFET works just as well if the source and drain contacts are magnetized in the $+y$ direction instead of the $+x$-direction. The advantage is that the magnetic field lines emanating from one contact, and sinking in the other, are no longer directed along the channel. Consequently, the channel magnetic field will be a fringing field, which is much weaker.

A more sophisticated approach is to play off the Dresselhaus spin–orbit interaction [14] against the channel magnetic field. This spin–orbit interaction is present in any zinc-blende semiconductor that lacks crystallographic inversion symmetry. In a 1-D channel oriented along the [100] crystallographic direction, the Dresselhaus spin–orbit interaction gives rise to a pseudo-magnetic field along the channel (x-axis), just as the Rashba spin–orbit interaction gives rise to a pseudo-magnetic field perpendicular to the channel (in the z-direction). The Dresselhaus field $B_{\text{Dresselhaus}}$ can be used to offset the channel magnetic field due to the contacts. Since $B_{\text{Dresselhaus}}$ depends on the carrier velocity v_x, the ensemble average velocity $<v_x>$ (which is the Fermi velocity for a degenerate carrier concentration) can be tuned with a backgate to make $B_{\text{Dresselhaus}}$ equal and opposite to the channel magnetic field, thereby offsetting the effect of the channel field. This was the approach proposed in Ref. [8].

In Figure 5.1b, the transfer characteristic (drain current versus gate voltage) is shown for an ideal SPINFET and a non-ideal SPINFET (with a channel magnetic field), ignoring any fixed threshold shift caused by a non-zero Rashba interaction in the channel at zero gate voltage. It should be noted that the transfer characteristic is "oscillatory" and therefore non-monotonic. As a result, the transconductance ($\partial I_D/\partial V_G$), where I_D is the drain current and V_G is the gate voltage, can be either positive or negative depending on the value of V_G (gate bias). A fixed threshold shift can be caused in any SPINFET by implanting charges in the gate insulator. Imagine now that there are two ideal SPINFETs with transfer characteristics, as shown in Figure 5.1c. These two can be connected in series to behave as a complementary metal oxide semiconductor (CMOS) like inverter where an appreciable current flows only during switching. This would be a tremendous advantage as CMOS like operation can be achieved with just n-type SPINFETs where the majority carriers are electrons. In conventional CMOS technology, both an n-type and a p-type device would have

normally have been needed. Here, we need only n-type devices. However, all this advantage is defeated if there is significant leakage current flowing through the SPINFET when it is "off". Thus, the leakage current is a rather serious issue and care must be taken to eliminate it as much as possible.

5.2.1.2 The SPINFET Based on the Dresselhaus Spin–Orbit Interaction

The Dresselhaus spin–orbit interaction can also be gainfully employed to realize a different kind of SPINFET [15]. In a 1-D channel, the strength of the Dresselhaus interaction depends on the physical width of the channel. If the 1-D channel is defined by a split-gate, then the voltages on the split gate can be varied to change the channel width and therefore the strength of the Dresselhaus interaction. As with the Rashba interaction, the Dresselhaus interaction also causes spins to precess in space at a rate independent of the carrier velocity because it gives rise to a pseudo-magnetic field $B_{\text{Dresselhaus}}$ along the x-axis. The spins precess in the y-z plane. By varying the split gate voltage it is possible to change $B_{\text{Dresselhaus}}$ and the precession rate, and therefore the angle Φ, by which the spins precess as they traverse the channel from source to drain. As the split gate voltage can be used to modulate the source to drain current, transistor action can be realized. A schematic representation of a Dresselhaus-type SPINFET is shown in Figure 5.3.

The advantage of the Dresselhaus-type SPINFET over the Rashba-type is that, in the former, there is never a strong magnetic field in the channel due to contacts [15], as $B_{\text{Dresselhaus}}$ is in the x-direction and therefore the ferromagnetic contacts must be magnetized in the y-z plane (see Figure 5.3). By contrast, B_{Rashba} is in the z-direction, and therefore in the Rashba-type device the ferromagnetic contacts must be magnetized in the x-y plane. As mentioned above, the Rashba-type device could avoid a strong channel magnetic field if the contacts were to be magnetized in the y-direction, but this is difficult to do as the ferromagnetic layer thickness in the y-direction is much smaller than that in the x- or z-directions. Thus, the Rashba-type device will typically have some magnetic field in the channel, while the Dresselhaus-type device will not, except for the fringing fields. This feature eliminates many of the problems

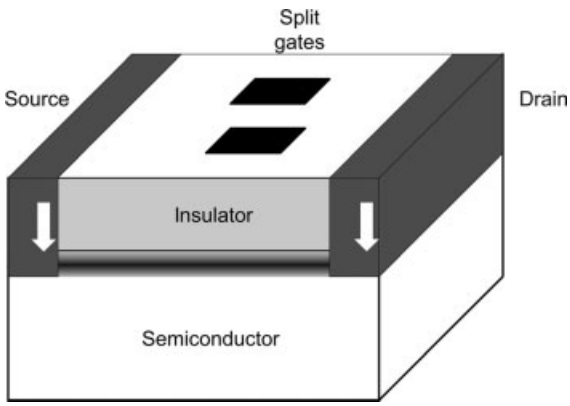


Figure 5.3 Schematic of a SPINFET based on the Dresselhaus spin–orbit interaction.

associated with the channel magnetic field in the Dresselhaus-type SPINFET, and could lead to a reduced leakage current. A comparison between the Rashba-type and Dresselhaus-type SPINFETs is provided in Ref. [15].

5.2.2 Device Performance of SPINFETs

In the world of electronics, the universally accepted benchmark for the transistor device is the celebrated metal-oxide-semiconductor-field-effect-transistor (MOSFET) which has been – and still is – the "workhorse" of all circuits. Therefore, the SPINFET must be compared with an equivalent MOSFET to determine if there are any advantages to utilizing spin. Surprisingly, in spite of many papers extolling the perceived merits of SPINFETs, this elementary exercise was not carried out until recently. When an ideal SPINFET was compared with an equivalent ideal MOSFET at low temperatures [16], the results were quite illuminating.

According to Ref. [1], the switching voltage necessary to turn a 1-D Rashba-type SPINFET from on to off, or vice versa, is given by

$$V_{switching}^{SPINFET} = \frac{\pi\hbar^2}{2m^* L\xi} \tag{5.4}$$

where m^* is the effective mass, L is the channel length (or source-to-drain separation), and ξ is the rate of change of the Rashba interaction strength in the channel per unit change of the gate voltage. This is given by [16]:

$$\xi = \frac{\hbar^2}{2m^*}\frac{\Delta(2E_g+\Delta)}{E_g(E_g+\Delta)(3E_g+2\Delta)}\frac{2\pi e}{d} \tag{5.5}$$

where E_g is the bandgap of the channel semiconductor, Δ is the spin orbit splitting in the valence band of the semiconductor, e is the electronic charge, and d is the gate insulator thickness. If $d = 10$ nm,[6] then it can be calculated that in an InAs channel, $\xi = 10^{-28}$ C-m. This is the theoretical value, but an actual measured value is much less than this [17]. The compound InAs has strong spin–orbit interaction and therefore is an ideal material for SPINFETs.

Now, imagine that the same structure is used as a MOSFET. Then, the switching voltage that turns the MOSFET device off (depletes the channel of mobile carriers) is E_F/e, where E_F is the Fermi energy in the channel.[7] Thus, the ratio of the switching

6) The ideal semiconductor is a narrow gap semiconductor such as InAs, which has strong Rashba spin–orbit interaction. In that case, the gate insulator will probably be AlAs, which is reasonably lattice-matched to InAs. Because the conduction band offset between these two materials is not too large, a minimum of 10 nm gate insulator thickness may be necessary to prevent too much gate leakage.

7) An accumulation mode MOSFET has been assumed that is "normally-on". It has also been assumed that the normal channel carrier concentration is large enough that $E_F \gg kT$, where E_F is the Fermi energy and kT is the thermal energy. Therefore, the comparison is strictly valid at very low temperatures. It is believed that, at higher temperatures, the fundamental conclusions from this comparison will not be significantly altered.

voltages is

$$\frac{V_{switching}^{SPINFET}}{V_{switching}^{MOSFET}} = \frac{\pi\hbar^2 e}{2m^* L\xi E_F} \tag{5.6}$$

In order to maintain single subband occupancy in an InAs 1-D channel of reasonable width, E_F must be less than ~3 meV. Therefore, from Eq. (3.6) it is found that the SPINFET will have a lower threshold voltage than a comparable MOSFET (at low temperature) only if its channel length exceeds 2.4 μm! Thus, no submicron SPINFET has any advantage over a comparable MOSFET in terms of switching voltage or dynamic power dissipation during switching. Currently, MOSFETs with 90 nm channel length are in production [18]. For a SPINFET with this channel length to have a lower threshold voltage than a comparable MOSFET, the channel must be made from a material in which the product $m^*\xi$ is 26-fold larger than it is in InAs, assuming that E_F is still 3 meV. Such materials are not currently available, but of course could become available in the future. The unfortunate spoiler is that materials which have large effective mass also tend to have weak spin–orbit interaction, which makes it difficult to increase the product $m^*\xi$.

One issue that requires some thought here is that the switching voltage of a MOSFET depends on E_F – and therefore the carrier concentration in the channel – while the switching voltage of a SPINFET depends on the channel length. It is not clear which quantity is easier to control in batch processing, but that would determine which device has an advantage in terms of threshold variability and, ultimately, of yield.

5.2.3 Other Types of SPINFET

Slightly different types of SPINFET ideas have also been reported in the literature, with names such as "Non-ballistic SPINFET" [19] or the "Spin Relaxation Transistor" [20–22].

5.2.3.1 The Non-Ballistic SPINFET

The channel of the so-called "non-ballistic SPINFET" has a two-dimensional (2-D) electron gas, like an ordinary MOSFET. Unlike in a 1-D structure (quantum wire), the spin split bands in a 2-D structure (quantum well or 2-D electron gas) do not have a fixed spin quantization axis (meaning that the spin eigenstates are wavevector-dependent; recall Figure 5.2b), even if there is no magnetic field. The only exception to this situation is when the Rashba and Dresselhaus interactions in the channel have exactly the same strength. In that case, each band has a fixed spin quantization axis, and the spin eigenstate in either band is wavevector-independent.

In the non-ballistic SPINFET, the Rashba interaction is first tuned with the gate voltage to make it exactly equal to the Dresselhaus interaction, which is gate voltage-independent in a 2-D electron gas. This makes the spin eigenstates wave-

vector-independent. Electrons are then injected into the channel of the transistor from a ferromagnetic source with a polarization that corresponds to the spin eigenstate in one of the bands. All carriers enter this band. As the spin eigenstate is wavevector-independent, any momentum relaxing scattering in the channel, which will change the electron's wavevector, will not alter the spin polarization (recall the discussion in Section 5.2.1.1). Scattering can couple two states within the same band, but not in two different subbands, as the eigenspinors in two different subbands are orthogonal. Therefore, regardless of how much momentum-relaxing scattering takes place in the channel, there will be no spin relaxation via the Elliott–Yafet mode. There will also be no D'yakonov–Perel' relaxation as it can be shown that the pseudo-magnetic field due to the Rashba and Dresselhaus interactions will be aligned exactly along the direction of the eigenspin. As all spins are initially injected in an eigenstate, they will always be aligned along the pseudo-magnetic field. Consequently, there will be no spin precession which would have caused D'yakonov–Perel' relaxation. As no major spin relaxation mechanism is operative, the carriers at the source will arrive at the drain with their spin polarization intact. If the drain ferromagnetic contact is magnetized parallel to the source magnetization, then all these carriers will exit the device and contribute to current. In order to change the current, the gate voltage is detuned; this makes the Rashba and Dresslhaus interaction strengths unequal, thereby making the spin eigenstates in the channel wavevector-dependent. In that case, if the electrons suffer momentum-relaxing collisions due to impurities, defects, phonons, and surface roughness, their spin polarizations will also rotate and this will result in spin relaxation. Thus, the carriers that arrive at the drain no longer have all their spins aligned along the drain's magnetization. Consequently, the overall transmission probability of the electrons decreases, and the current drops. This is how the gate voltage changes the source-to-drain current and produces transistor action.

One simple way of viewing the transistor action is that when the Rashba and Dresselhaus interactions are balanced, the channel current is 100% spin polarized (all carriers arriving at the drain have exactly the same spin polarization). However, when the two interactions are unbalanced, then the spin polarization of the current decreases owing to spin relaxation. The spin polarization can decrease to zero – *but no less than zero* – which means that, on average, 50% of the spins will be aligned and 50% anti-aligned with the drain's magnetization when the minimum spin polarization is reached. The "aligned" component in the current will transmit and the "anti-aligned" component will be blocked. Thus, the minimum current (off-current) of this transistor *is only one-half* of the maximum current (on-current). As the maximum ratio of the on-to-off conductance is only 2, this device is clearly unsuitable for most – if not all – mainstream applications. A recent simulation has shown that the on-to-off conductance ratio is only about 1.2 [23], which precludes use in any fault-tolerant circuit.

The situation can be improved dramatically if the source and drain contacts have *anti-parallel* magnetizations, instead of parallel. In that case, when the Rashba and Dresselhaus interactions are balanced, the transmitted current will be exactly zero, but when they are unbalanced, it is non-zero. Therefore, the on-to-off conductance

becomes infinity. However, there is a caveat. This argument pre-supposes that the ferromagnetic contacts can inject and detect spins with 100% efficiency, meaning that only the majority spins are injected and transmitted by the ferromagnetic source and drain contacts, respectively. That has never been achieved, and even after more than a decade of research the maximum spin injection efficiency demonstrated to date at room temperature is only about 70% [24]. That means

$$\frac{I_{\mathrm{maj}} - I_{\mathrm{min}}}{I_{\mathrm{maj}} + I_{\mathrm{min}}} = 0.7 \tag{5.7}$$

where $I_{\mathrm{maj(min)}}$ is the majority (minority) spin component of the current. If the spin injection efficiency is only 70%, then 15% of the injected current is due to minority spins. These minority spins will transmit through the drain, even when the Rashba and Dresselhaus interactions are balanced. Thus, the off-current is 15% of the total injected current ($I_{\mathrm{maj}} + I_{\mathrm{min}}$), whereas the on-current is still at best 50% of the total injected current. Therefore, the on-to-off ratio of the conductance is $0.5/0.15 = 3.3$, which is not much better than 2.[8] Consequently, achieving a large conductance ratio is very difficult, particularly at room temperature when the spin injection efficiency tends to be small. In order to make the conductance ratio 10^5 – which is what today's transistors have – the spin injection efficiency must be 99.999% at room temperature. This is indeed a tall order, and may not be possible in the near term. If the off (leakage) current is nearly one-third of the on-current (which is what it will be with present-day technology), then the standby power dissipation will be intolerable and the noise margin unacceptable.

The device described in Ref. [20] is identical to that in Ref. [19], and therefore the same considerations apply.

5.2.3.2 **The Spin Relaxation Transistor**

The proposed "spin relaxation transistor" [21, 22] is very similar to the non-ballistic SPINFET. With zero gate voltage, the spin–orbit interaction in the channel is either weak or non-existent, which makes the spin relaxation time very long. When the gate voltage is turned on, the spin–orbit interaction strength increases, which makes the spin relaxation time short. Thus, with zero gate voltage, the spin polarization in the drain current is large (maximum 100%), while with a non-zero gate voltage it is small (minimum 0%). This device cannot be any better than the non-ballistic SPINFET. If the drain and source magnetizations are parallel, then it is easy to see that in the "on" state, the transmitted current is at best 100% of the injected current, while in the "off" state, it is no less than 50% of the injected current (the same arguments as in Section 5.2.3.1 apply). Therefore, the on-to-off conductance ratio is less than 2. With anti-parallel magnetizations in the source and drain contacts, the off-current approaches 0, and the conductance ratio approaches infinity, but only if the contacts inject and detect spin with 100% efficiency. If the injection efficiency is only 70%

8) Here it has been assumed that the drain is a perfect spin filter, or equivalently, the spin detection efficiency at the drain is 100%.

then, following the previous argument, the conductance ratio is no more than 3.3 [25]. Therefore, this device, too, is unsuitable for mainstream applications.

5.2.4 The Importance of the Spin Injection Efficiency

In every proposal for SPINFETs discussed here [1, 15, 19–22] there has always been the tacit assumption that spin injection and detection efficiencies at the source/channel and drain/channel interfaces are 100%. This is of course unrealistic. It can be shown easily that the ratio of on- to off-conductance of SPINFETs of the type proposed in Refs. [1, 15] is

$$r = \frac{1 + \xi_S \xi_D}{1 - \xi_S \xi_D} \tag{5.8}$$

where ξ_S is the spin injection efficiency at the source/channel interface and ξ_D is the spin detection efficiency at the drain/channel interface. For SPINFETs of the types proposed in Refs. [19–22],

$$r = \frac{1}{1 - \xi_S \xi_D} \tag{5.9}$$

Therefore, the spin injection/detection efficiency ξ is critical in order to obtain a large enough value of r. If the spin injection and detection efficiencies fall from 100% to 90%, the conductance on-off ratio drops from infinity to 5.2! Therefore, without a very high spin injection efficiency, approaching 100%, none of the SPINFETs will have a sufficiently high on-to-off conductance ratio to be of much use for anything. Thus, a closer look at spin injection efficiency is clearly required.

5.2.4.1 Spin Injection Efficiency

The science of "spin injection efficiency" (what controls it, what improves it, what does it depend on, etc.) is controversial (see Chapter 3 in Volume III), and there is scant agreement among research groups in this field. However, to the authors' knowledge nobody has claimed in the open literature that the spin injection efficiency can be 100% (even theoretically), particularly at elevated temperatures, except for Ref. [26]. These authors hold that, as there has been rapid experimental progress in improving spin injection efficiency over the past seven years (as of this writing), 100% efficient spin injection at room temperature should be around the corner. This optimism is not shared by the present author, since the two mechanisms suggested [26] as possible routes to achieving 100% spin injection efficiency are to use 100% spin-polarized half-metallic ferromagnets as spin injectors, and spin-selective barriers. Unfortunately, there are no 100% spin-polarized half-metals at any temperature above 0 K because of phonons and magnons [27], and even at 0 K the 100% spin polarization is destroyed by surfaces and inhomogeneities [27]. Thus, 100% spin-polarized half-metals simply do not exist. The best spin-selective barriers are resonant tunneling devices [28] that have spin-resolved energy levels. When the carrier energy is resonant with one of these levels, only the corresponding spin transits through the barrier. However, at any

temperature exceeding 0 K, the thermal spread in the electron energy will cause injected electrons to tunnel through *both* levels if the level separation is less than kT (which it usually is). Therefore, both spins will be transmitted. This happens even if the spin levels themselves are not broadened by $\sim kT$ because of weak spin–phonon coupling. As long as both spins are transmitted, the spin "selection" suffers and that makes the spin injection efficiency considerably less than 100%.

To summarize, as yet there is no known method to suggest – even theoretically – the possibility of 100% spin injection efficiency. As a result, all SPINFETs discussed in this chapter suffer from the malady of a low on-to-off conductance ratio, and this alone may make them non-competitive with MOSFETs.

5.2.5
Spin Bipolar Junction Transistors (SBJTs)

The SBJT is identical with the normal bipolar junction transistor, except that the base is ferromagnetic and it has a non-zero spin polarization. The conduction energy band diagram of a heterojunction n-p-n SBJT is shown in Figure 5.4. Assuming that the carrier concentration in the base is non-degenerate, so that Boltzmann statistics apply, the spin polarization in the base is

$$\alpha_b = \tanh(\Xi/2kT) \tag{5.10}$$

where 2Ξ is the energy splitting between majority and minority spin bands in the base. Based on a small signal analysis, it was shown that the voltage and current gains afforded by the SBJT is about the same as a conventional BJT [29], but the short-circuit current gain β has a dependence on the degree of spin polarization in the base which can be altered with an external magnetic field using the Zeeman effect. Thus, the external magnetic field can act as a "fourth terminal", and this can lead to non-linear circuits such as mixers/modulators. For example, if the ac base current is a sinusoid with a frequency ω_1 and the magnetic field is an ac field which is another sinusoid with frequency ω_2, then the collector current will contain frequency components

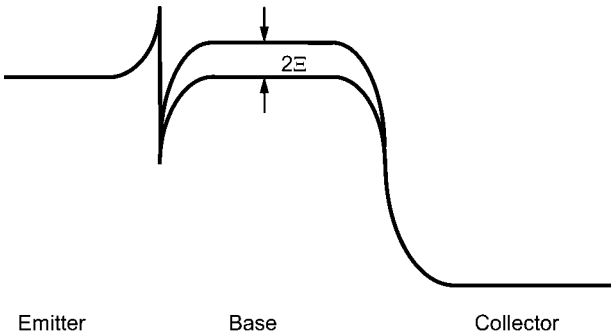


Figure 5.4 Conduction band energy profile for an n-p-n spin bipolar junction transistor (SBJT). The base is spin-polarized (ferromagnetic) and the spin splitting energy in the base is 2Ξ.

$\omega_1 \pm \omega_2$. This is one example where "spin" augments the role of "charge", making the SBJT another "hybrid spintronic" device.

5.2.6 The Switching Speed

The switching delay of any of the SPINFETs discussed above is limited by the transit time of carriers through the channel (or base). This is entirely due to the fact that information is encoded by charge (or current), and therefore the charge transit time is the bottleneck that ultimately limits the switching speed. Thus, hybrid spintronic devices do not promise any better speed than their charge-based counterparts.

5.3 Monolithic Spintronics: Single Spin Logic

At this point the discussion centers on "monolithic spintronics" where charge plays no role whatsoever and spin polarization is used to store, process and transmit information. In 1994, the idea was proposed of "single spin logic" (SSL) where a single electron acts as a binary switch and its two orthogonal (anti-parallel) spin polarizations encode binary bits 0 and 1 [30]. Switching between bits is accomplished by simply flipping the spin *without physically moving charges*. To the author's knowledge, this is the first known logic family (classical or quantum) based on single electron spins.

5.3.1 Spin Polarization as a Bistable Entity

The first step in SSL is to make the spin polarization of an electron a *bistable* quantity that has only two stable "values" that will encode the bits 0 and 1. In charge-based electronics, the state variables representing digital bits (voltage, current or charge), are not bistable but are continuous variables. So, why is the spin polarization required to be bistable? The reason is that in the world of electronics, there are analog-to-digital converters that can convert a continuous variable (analog signal) to a discrete variable (digital signal). More importantly, logic gates act as amplifiers with power gain and can automatically restore digital signal at logic nodes [31] if the signal is corrupted by noise. There are no equivalent analog-to-digital converters for spin polarization and no spin amplifiers, and therefore *Nature* must be relied upon to digitize spin polarization and make signal degeneration impossible. This can happen if *Nature* permits only two values of spin polarization – that is, it inherently makes it "bistable". That can be accomplished by placing an electron in a static magnetic field. The Hamiltonian describing a single electron in a magnetic field is

$$H = (\vec{p} - q\vec{A})^2/2m^* - (g/2)\mu_B \vec{B} \cdot \vec{\sigma} \tag{5.11}$$

where $\vec{A}$ is the vector potential due to the magnetic flux density $\vec{B}$, μ_B is the Bohr magneton, g is the Lande g-factor, and $\vec{\sigma}$ is the Pauli spin matrix. If the magnetic

field is directed in the x-direction ($\vec{B} = B\hat{x}$), then diagonalization of the above Hamiltonian immediately produces two mutually orthogonal eigenspinors [1,1] and [1, − 1] which are the $+x$ and $-x$-polarized spins – that is, states whose spin quantization axes are parallel and anti-parallel to the x-directed magnetic field. Thus, the spin quantization axis (or spin polarization) has only two stable values and therefore becomes a *binary* variable. The "down" (parallel) or "up" (anti-parallel) states can encode logic bits 0 and 1, respectively.

5.3.2
Stability of Spin Polarization

Although the binary bits 0 and 1 can be encoded in the two anti-parallel spin polarizations, there remains a problem in that random bit flips caused by coupling of spins to the environment will cause bit errors and corrupt the data. The probability of a "bit flip" within a clock cycle is $1 - e^{-\frac{T}{\langle\tau\rangle}}$, where T is the clock period and $\langle\tau\rangle$ is the mean time between random spin flips. In order to make this probability small, it must be ensured that $\langle\tau\rangle \gg T$.

If the host for the spin is a "quantum dot", then indeed $\langle\tau\rangle$ can be quite long. In InP quantum dots, the single electron spin flip time has been reported to exceed 100 μs at 2 K [32]. More recently, several experiments have been reported claiming spin flip times (or so-called longitudinal relaxation time, T_1) of several milliseconds, culminating in a recent report of 170 ms in a GaAs quantum dot at low temperature (see the last reference in Ref. [33]). An extremely surprising result is that spin relaxation time $\langle\tau\rangle$ in organic semiconductors can be incredibly long. It was found that the spin relaxation time in tris(8-hydroxyquinolinolate aluminum) – popularly known as Alq_3 – can be as long as 1 s at 100 K [34]. If the clock frequency is 5 GHz, then the clock period T is 200 ps, which is 5×10^9 times smaller than the spin flip time. Therefore, the probability that an unintentional spin flip will occur between two successive clock pulses is $1 - e^{-1/(5\times 10E9)} = 2 \times 10^{-10}$, which can be handled by modern error correction algorithms [35].

Typical error probabilities encountered in today's integrated circuits range from 10^{-10} to 10^{-9}. If a 5-GHz clock is used and an error probability $1/p$ of 10^{-9} is required, the spin flip time needs to be only 200 ms, which is fairly easy to achieve today at low temperatures (77 K).

5.3.3
Reading and Writing Spin

"Spin" has one major disadvantage compared to "charge". Whereas, charge is extremely easy to read (or measure) with voltmeters, ammeters, electrometers, and so on, and extremely easy to write (or inject) with voltage and current sources, spin is much more difficult to read or write. Although reading is more difficult than writing, even single spins have been "read" recently. The following section discusses reading and writing.

5.3.3.1 Writing Spin

Spin bits in SSL are represented by the spin polarizations of single electrons, with each electron being hosted in a quantum dot. There are exotic methods of "writing" spin bits in quantum dots [3], but the easiest and conceptually most simple is to use local magnetic fields generated with inductive loops. These fields will orient the spins in the target dots along the direction of the field and write the chosen bit (see Chapter 4 in Volume III of this series). Local magnetic fields are generated on chip in magnetic random access memory (MRAM), but not with the spatial resolution required in the case of SSL. However, carbon nanotube-based inductive loops may be up to the task in the near future.

The next task is to estimate what amount of energy will be dissipated during the "writing" operation. Previously [3], the idea of writing spin using Rabi oscillation, as is done in electron spin resonance spectroscopy, was proposed. That mode of writing will not dissipate any energy whatsoever, but has no error tolerance as an ac magnetic pulse must be applied for precisely the correct duration. However, if overshoot or undershoot occurs, an error will be incurred that can build up and ultimately cause a bit-writing error. Error-free writing will require some dissipation, though this can be arbitrarily small.

5.3.3.2 Reading Spin

Reading spin bits is more difficult than writing spin bits. Here, the aim is to ascertain the spin polarization of a single electron in a solid (quantum dot), a feat which has been accomplished recently using three different methods [36–38]. The technique reported in Ref. [37] is eminently suitable for application in electronics.

5.3.4 The Universal Single Spin Logic Gate: The NAND Gate

The basic idea behind implementing logic gates in SSL is to engineer the interactions between input and output spin bits in such a way that the input–output relationship represents the "truth table" of the desired logic gate. This approach can be illustrated by showing how a NAND gate can be realized. The NAND gate is a universal gate with which any arbitrary combinational or sequential logic circuit may be implemented, and it is realized with a linear chain of three electrons in three quantum dots (see Figure 5.6). It will be assumed that only nearest-neighbor electrons interact via exchange as their wavefunctions overlap. Second nearest-neighbor interactions are negligible as exchange interaction decays exponentially with distance.

For NAND gate implementation, the leftmost and rightmost spins in Figure 5.5 must be regarded as the two "inputs bits", and the center spin as the corresponding "output bit". Assume that the downspin state ($\downarrow$) represents bit 1, and the upspin state ($\uparrow$) is bit 0. The global magnetic field, defining the spin quantization axes, is in the direction of "downspin". It has been shown recently, using a Heisenberg Hamiltonian to describe the 3-spin array, that as long as the Zeeman splitting energies caused by the local magnetic fields writing bits in the input dots is much larger than the

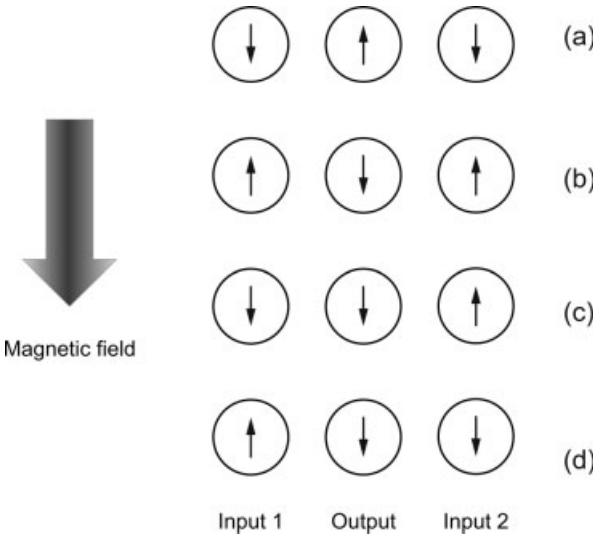


Figure 5.5 Single spin realization of the NAND gate. (a) When two inputs are [1 1]; (b) when two inputs are [0 0]; (c) when two inputs are [1 0]; and (d) when two inputs are [0 1]. Reproduced from Ref. 3 with permission from American Scientific Publishers: http://www.aspbs.com.

exchange coupling strength between neighboring dots, the ground-state spin configurations (determined by the directions of the local magnetic fields) are precisely those shown in Figure 5.5 [39]. In other words, if the input bits are written with local magnetic fields and the array is allowed to relax to the ground state in the presence of the local magnetic fields, then the output bit conforms to the diagrams in Figure 5.5a–d. It is evident that these configurations represent the truth table of the NAND gate:

Input 1	Input 2	Output
1	1	0
0	1	1
1	0	1
1	1	1

Therefore, if there is a 3-spin array, with nearest-neighbor exchange coupling, placed in a global magnetic field, and local magnetic fields align the spins in the peripheral dots to desired input bits, the output bit in the central dot will always be the NAND function of the input bits according to the truth table above, as long as two conditions are fulfilled:

- The array is in the thermodynamic ground state.

- The Zeeman splitting in the input dots caused by the local magnetic fields writing input data is much larger than the exchange coupling strength, which is roughly the energy difference between the triplet and singlet states in two neighboring dots.

Independent quantum mechanical calculations to confirm the working of the NAND gate were carried out by Molotkov and Nazin [40], while further investigations in this area have been conducted by Bychkov and coworkers [41].

Once the NAND gate is realized, only one other component is needed to implement any arbitrary combinational or sequential Boolean logic circuit. That element is a "spin wire" (with fan out) which will ferry spin logic signal from one stage to another *unidirectionally*. A spin wire (see Figure 5.6) consists of a linear array of quantum dots with clock pads between them. When the clock signal at a given pad is high, the potential barrier between the two surrounding dots is lowered and these dots are exchange-coupled because their wavefunctions overlap. This makes the spin states in these dots anti-parallel [42]. Therefore, by sequentially clocking the barriers, the spin bit can be replicated in every other dot, thus moving the spin signal along unidirectionally. Both, Sarkar et al. [43] and Bose et al. [44] have implemented a large array of combinational and sequential logic circuits using this approach. Two examples of their investigations are shown in Figure 5.7. The issue of unidirectionality and clocking will be revisited in Section 5.3.7.

5.3.5
Bit Error Probability

The NAND gate operates by relaxation to the thermodynamic ground state. It is the natural tendency of any physical system to gravitate towards the minimum energy

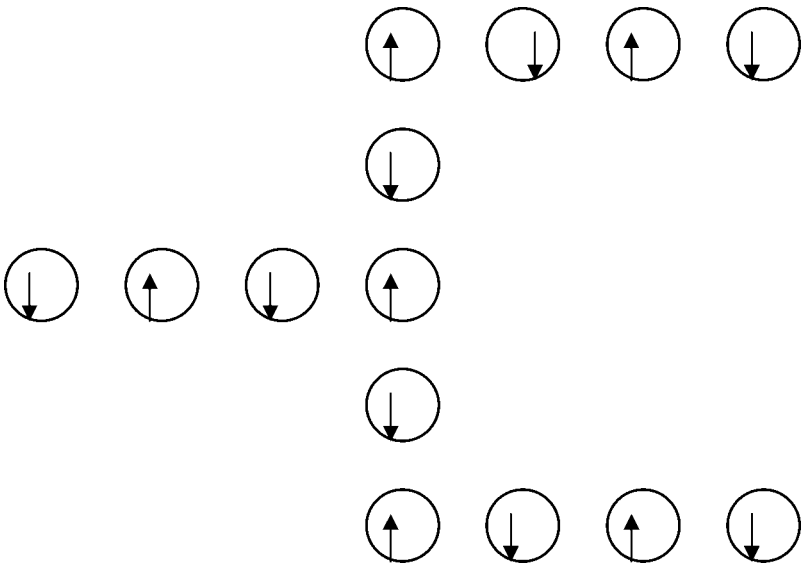


Figure 5.6 Realization of a "spin wire" with fan out.

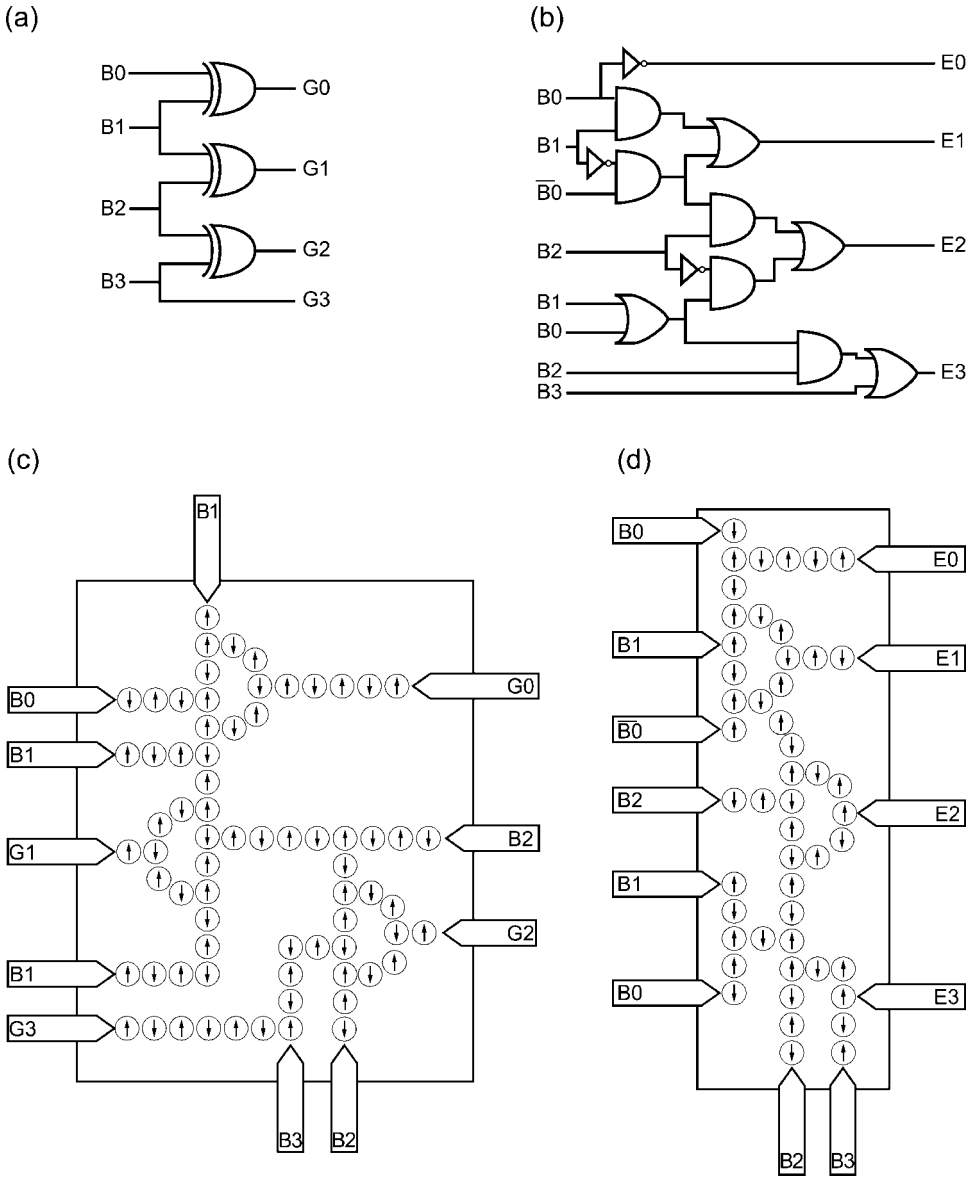


Figure 5.7 Single spin realizations of code converters. (a) Logic diagram for binary to Gray code converter; (b) logic diagram for binary to Excess-3 converter; (c) SSL realization of binary-to-Gray-code converter; and (d) SSL realization of binary-to-Excess-3 converter. The clock pads between successive cells are not shown for the sake of clarity. The input binary code is '1010'. Note that adjacent cells have anti-parallel polarizations indicating anti-ferromagnetic ordering. Reproduced from Ref. 43, with permission from Institute of Solid State Physics, Chernogolovka, Russia.

state (ground state), this being the law of thermodynamics. However, when a system achieves the ground state it need not stay there forever, as noise and fluctuations can take it to an excited state. If that happens and the NAND gate strays from the ground state, the results will be in error. This error probability is calculated next.

The NAND gate reaches ground state by exchanging phonons with the thermal environment (phonon bath). This brings it into thermodynamic equilibrium with the surrounding thermal bath. In that case, the occupation probability of any eigenstate of the gate will be given by Fermi–Dirac statistics. The ground-state occupation probability is $1/[\exp(E_{\text{ground}} - E_F)/kT + 1]$ (where E_F is the Fermi energy) and the excited state occupation probability is $1/[\exp(E_{\text{excited}} - E_F)/kT + 1]$. As the occupation probability of the ground state is *not* unity, the gate does not always work correctly with 100% certainty – in other words, the error probability $1/p$ is never zero. However, it does decrease with increasing energy separation between the excited and ground states.

The error probability $1/p$ is the sum of the ratios of the probabilities of being in the excited and ground state, summed over all excited states. This quantity is approximately $\sum_{\text{excited states}} e^{-(E_{\text{excited}} - E_{\text{ground}})/kT}$, if the Fermi–Dirac statistics are approximated with Boltzmann statistics. It transpires that, in the case of the NAND gate, the second and higher excited states are far above in energy than the first excited state [39]. Therefore, only the first excited state E_1 can be retained in the sum above, and hence $E_1 - E_{\text{ground}} = kT\ln(p)$. It was also shown rigorously in Ref. [39] that $E_1 - E_{\text{ground}}$ is: (i) $4J - 2Z$ when the input bits are [1 1]; (ii) $4J + 2Z$ when the input bits are [0 0]; and (iii) $2Z$ when the input bits are [0 1] or [1 0]. Here, J is the exchange coupling strength between two neighboring dots, and $2Z$ is the Zeeman splitting energy in any dot due to the global magnetic field. Therefore, in order to attain an error probability of $1/p$ at a temperature T, it must be ensured that: (a) $2Z = kT\ln(p)$; and (b) $4J - 2Z = kT\ln(p)$, or $2J = kT\ln(p)$. The maximum values of J or Z are usually limited by technological constraints; for example, J is usually limited to 1 meV in gate-defined quantum dots [45]. These limits will then determine the maximum temperature of operation if a certain error probability $1/p$ is insisted on (this issue will be revisited in Section 5.3.10).

5.3.6
Related Charge-Based Paradigms

A similar idea for implementing logic gates using single electron charges confined in "quantum dashes" was proposed by Pradeep Bakshi and coworkers in 1991 [46]. There, logic bits were encoded in bistable charge polarizations of elongated quantum dots known as "quantum dashes". Coulomb interaction between nearest-neighbor quantum dashes pushes the electrons in neighboring dashes into antipodal positions, making the ground-state charge configuration "anti-ferroelectric", just as the exchange interaction in the present case tends to make the ground-state spin configuration almost "anti-ferromagnetic". Three Coulomb-coupled quantum dashes would realize a NAND gate in a way very similar to what was described here.

Bakshi's idea inspired a closely related idea known as "quantum cellular automata" [47], which uses a slightly different host, namely a four- or five-quantum dot "cell" instead of a quantum dash to store a bit. Here too the charge polarization of the cell is bistable and encodes the two logic bits. The only difference from Ref. [46] is that coulomb interaction makes the ground-state charge configuration ferroelectric, instead of anti-ferroelectric.

In the schemes of Refs. [46, 47] it is difficult to implement only nearest-neighbor interactions, at the exclusion of second-nearest-neighbor interactions, mainly because the Coulomb interaction is long range. The interaction in Refs. [46, 47] drops off as a polynomial of distance, but never exponentially with distance, unless strong screening can be implemented. If second-nearest-neighbor interactions are not much weaker than their nearest-neighbor counterparts, the ground-state charge configuration is weakly stable and not sufficiently robust against noise. In this respect, the spin-based approach has an advantage. As exchange interaction is short range (it always drops off exponentially with distance), it is much easier to make the second-nearest-neighbor interactions considerably weaker than the nearest-neighbor interactions.

A second issue is that in Refs. [46, 47], there is internal charge movement within each cell during switching, causing "eddy currents". This is a source of dissipation that is absent in the spin-based paradigm, as there is never any charge movement.

5.3.7
The Issue of Unidirectionality

The "unidirectional" propagation of logic signal was briefly mentioned in Section 5.3.5. This is a vital issue as the input signal should always influence (and determine) the output signal, but not the other way around. Unidirectionality is an important requirement for logic circuits [31]. A transistor is inherently unidirectional as there is *isolation* between its input and output terminals that guarantees unidirectionality. Therefore, it is easy to make logic circuits with transistors. In SSL, there is unfortunately *no isolation* between the input and output of the logic gate as exchange interaction is "bidirectional". Consider just two exchange coupled spins in two neighboring quantum dots; they will form a singlet state and therefore act as a natural NOT gate if one spin is the input and the other is the output (the output is always the logic complement of the input) [42]. However, exchange interaction, being bidirectional, cannot distinguish between which spin is the input bit and which is the output. The input will influence the output just as much as the output influences the input, and the master–slave relationship between input and output is lost. As the input and output are indistinguishable, it becomes ultimately impossible for logic signal to flow *unidirectionally* from an input stage to an output stage, and not the other way around. This issue has been discussed at length elsewhere [30, 48].

It is of course possible to forcibly impose unidirectionality in some (but not all) cases by holding the input cell in a fixed state until the desired output state is produced in the output cell. In that case, the input signal itself enforces unidirectionality because it is a symmetry-breaking influence. This approach was

actually used to demonstrate a "magnetic cellular automaton", where the input enforced unidirectionality and produced the correct output [49]. However, there are problems with this approach. First, it can only work for a small number of cells before the influence of the input dies out. Second – and more important – the input cannot be changed until the final output has been produced, since otherwise the correct output *may not be produced at all*. That makes such architectures *non-pipelined* and therefore unacceptably slow. There may also be additional problems associated with random errors when this approach is employed; these are discussed in Ref. [50].

5.3.8
Unidirectionality in Time: Clocking

If unidirectionality cannot be imposed in *space*, then it must be imposed in *time*. This is accomplished by using clocking to activate successive stages sequentially in time [51, 52]. This strategy is well known and routinely adopted in bucket-brigade devices, such as charge-coupled-device (CCD) shift registers,[9] where a push clock and a drop clock are used to lower and raise barriers between neighboring devices and thus steer a charge packet unidirectionally from one device to the next. The same can be done in single-spin circuits. A gate pad will be delineated between every two neighboring quantum dots, and a clock signal applied to this pad. During the positive clock cycle, a positive potential will appear over the potential barrier, thus isolating two neighboring quantum dots; this will lower the barrier temporarily to exchange-couple the two spins and result in the two spins assuming antiparallel polarizations [42]. Then, during the negative clock cycle, the barrier is raised again to decouple the two spins. In this way, pairs of spin bits can be coupled sequentially in time, and the logic information transferred unidirectionally from one dot to the next in a bucket-brigade fashion. It has been shown previously [52] that a single-phase clock does not work, and that a three-phase clock is required to carry out this task.

The clocking circuit, however, introduces additional dissipation. The energy dissipated in the clock pad is $(1/2)CV^2$ if the clock pulse is applied non-adiabatically, and much less if applied adiabatically [6]. So, what should be the value of V? This is determined by noise considerations. The noise voltage on a capacitor is [53]:

$$U_n = \sqrt{\frac{kT}{C}} \tag{5.12}$$

and Ref. [53] prescribes that $V = 12\ U_n$ for reasonable error rates at clock frequencies over 10 GHz. In that case, the energy dissipated in the clock pads is $72\,kT$ if

9) CCDs also have no inherent unidirectionality. There, a push clock and a drop clock are used to steer charge packets from one device to the next. See, for example, D. K. Schroeder, in: G. W. Neudeck, R. F. Pierret (Eds.), *Advanced MOS Devices, Modular Series in Solid State Devices*, Chapter 3, Addison-Wesley, Reading, MA, **1987**.

applied non-adiabatically. If the clock signal is applied from a sinusoidal voltage source of amplitude V, then the energy dissipated per clock cycle is

$$E_{diss} = \frac{1}{2} CV^2 \frac{\omega}{\omega_{RC}} \tag{5.13}$$

where $\omega_{RC} = 1/RC$. The above formula holds if the clock circuit is modeled as a capacitor C in series with a resistor R representing the resistance in the charging path. Assuming that $R = 1\,\text{k}\Omega$ and $C = 1\,\text{aF}$, $\omega_{RC} = 10^{15}\,\text{rad}\,\text{s}^{-1}$. If the clock frequency is 5 GHz, then $\omega = 3.45 \times 10^{10}\,\text{rad}\,\text{s}^{-1}$. Therefore, $E_{diss} = 2.5 \times 10^{-3} kT$, which is negligible.

When clock pads are used, the most attractive feature of SSL is removed, namely the *absence* of interconnects (or "wires") between successive devices. "Wireless" exchange interaction plays the role of physical wires to transmit signals between neighboring devices, but in order to transmit "unidirectionally", each stage will need to be clocked and this requires a physical interconnect. A clock pad must be placed between pairs of quantum dots and wires must be attached to them to ferry the clock signal. Of course, a clock signal is also needed in traditional digital circuits involving CMOS, so that it is not an additional burden. Nevertheless, it still detracts from the appeal of a "wireless architecture", or the so-called "quantum-coupled architecture".

5.3.9
Energy and Power Dissipation

Most likely, by merely examining Figure 5.5, the reader can understand that the maximum energy dissipated when the gate switches between any two states is $2Z$, which is the Zeeman splitting in the output dot caused by the global magnetic field (this result was proved rigorously in Ref. [39]). Since it was shown in Section 5.3.5 that $2Z = kT\ln(p)$, the *maximum* energy dissipated during switching is $kT\ln(p)$, which was expected from the Landauer–Shannon result. The interesting point however is that the energy dissipated can be *less* than $kT\ln(p)$, depending on the initial and final states – that is, depending on the old and new input bit strings. If the gate switches from the state in Figure 5.5c to that in Figure 5.5a, the energy dissipated is actually $(2/3)\,kT\ln(p)$, while if it switches from the state in Figure 5.5b to that in Figure 5.5c, the energy dissipated is $(1/3)kT\ln(p)$ [39]. The reductions by factors of $1/3$ and $2/3$ are due to *interactions* between spins. Some implications of "interactions" were discussed in Ref. [4] in the context of reducing energy dissipation. What really happens here is that the three spins act collectively in unison as a single unit, because of the exchange-coupling between them, which reduces the total energy dissipation.

The maximum energy dissipation occurs when the gate switches from the state in Figure 5.5b to that in Figure 5.5a. That energy is $kT\ln(p)$. With $p = 10^{-9}$, this energy is $\sim 21\,kT$. By contrast, modern-day logic gates dissipate more than $50\,000\,kT$ when they switch [54].

5.3.10
Operating Temperature

In Section 5.3.5, the following result was established:

$$2J = kT\ln(p)$$
$$2Z = kT\ln(p)$$

The maximum value of J in semiconductor quantum dots is ~1 meV [45], while in molecules it is about 6 meV [55]. Therefore, if an error probability p of 10^{-9} is required, $T = 1.1$ K if semiconductor quantum dots are employed. Room-temperature operation with this error probability will require J to be 270 meV, which is clearly unattainable at present.

Since $Z = J$ from the above relationships, the strength of the global magnetic field required can be estimated. That strength is found by setting $2Z = g\mu_B B_{\text{global}} = 2J = 2$ meV. If a material is used with a g-factor of 15 (e.g. InAs), then $B_{\text{global}} = 2.3$ Tesla, which is very reasonable.

As the energy separation between the spin levels ($g\mu_B B_{\text{global}}$) is 2 meV and kT at 1.1 K = 0.1 meV, it might be considered that a low temperature of 1.1 K was required in order to make the thermal broadening of the spin split levels much less (in this case, 20-fold less) than the level separation. However, this line of thinking would be entirely wrong, as the low temperature was needed for a small error probability ($1/p = 10^{-9}$), and *not* because there was a need to reduce the thermal broadening of levels. Spin, unlike charge, does not couple strongly to phonons, and therefore the thermal broadening of spin levels is much less than kT. If this were not the case, then electron spin resonance experiments at microwave frequencies could not be carried out at room temperature.

Unfortunately, this fact is not often understood or appreciated. For example, in Ref. [56] the authors state (contradicting Ref. [6]) that: "Energy barriers for a spin system play the same role as for a charge transfer system. The barrier must be large enough to make the different bit states distinguishable in a thermal environment". This argument is at best half-correct. For charge-based representation, the energy barrier separating logic states indeed need to be large enough that the different bit states are distinguishable in a thermal environment, since charge couples very strongly to the thermal environment, so that the thermal broadening of the energy states is $\sim kT$. This is *not* true for spin which, unlike charge, does *not* couple strongly with the thermal environment, so that the spin energy states are not broadened by $\sim kT$. As a result, the energy barrier separating logic states (i.e. $|g\mu_B B|$) can be much less than kT if bit error probability were not a concern. The energy barrier merely needs to be equal to $kT\ln(p)$, which would be *less than* kT if an error probability $1/p > 1/e = 0.367$ can be handled.

5.3.11
Energy Dissipation Estimates

It is now possible to estimate how much energy is dissipated in various actions, assuming $T = 1.1$ K:

Clocking	Bit flip
$2.48 \times 10^{-3}\,kT$ (sinusoidal) $= 4 \times 10^{-26}$ Joules ~ 0 (adiabatic)	$kT\ln(p)$ $[1/p = 10^{-9}]$ $= 3 \times 10^{-22}$ Joules

With a bit density of $10^{11}\,\mathrm{cm}^{-2}$, the dissipation per unit area is 3×10^{-22} Joules $\times\, 5\,\mathrm{GHz} \times 10^{11}\,\mathrm{cm}^{-2} = 0.15\,\mathrm{W\,cm}^{-2}$. In comparison, the Pentium IV chip, with a bit density three orders of magnitude smaller, dissipates about $50\,\mathrm{W\,cm}^{-2}$ [57]. SSL dissipates 300 times less power with a bit density three orders of magnitude larger.

5.3.12 Other Issues

In charge-based devices such as MOSFETs, logic bits 0 and 1 are encoded by the *presence* and *absence* of charge in a given region of space. This region of space could be the channel of a MOSFET. When the channel is filled with electrons, the device is "on" and stores the binary bit 0. When the channel is depleted of electrons, the device is "off" and stores the binary bit 1. Switching between bits is accomplished by the physical motion of charges in and out of the channel, and this physical motion consumes energy.[10] There is no easy way out of this problem since charge is a "scalar" quantity and therefore has only a *magnitude*. Bits can be demarcated only by a difference in the magnitude of charge or, in other words, by the relative presence and absence of charge. Therefore, to switch bits the magnitude of charge must be changed, and this invariably causes motion of the charges with an associated current flow. Spin, on the other hand, has a polarization which can be thought of as a "pseudo-vector" with two directions: "up" and "down". Switching between bits is accomplished by simply flipping the direction of the pseudo-vector with no change in magnitude of anything. As switching can be accomplished without physically moving charges (and causing a current to flow), spin-based devices could be inherently much more energy-efficient than charge-based devices, a point which was highlighted in Ref. [3].

It is because of this reason that SSL is much more energy efficient than the Pentium IV chip.

5.4 Spin-Based Quantum Computing: An Engineer's Perspective

Quantum computing is based on the idea of encoding information in a so-called "qubit", which is very different from a classical bit. It is a coherent superposition of

10) In fact, the physical motion of charges causes a current I to flow with an associated I^2R dissipation (where R is the resistance in the path of the current).

classical bits 0 and 1:

$$\begin{aligned} &qubit = a|0\rangle + b|1\rangle \\ &|a|^2 + |b|^2 = 1 \end{aligned} \tag{5.14}$$

The coefficients a and b are complex numbers, and the phase relationship between them is important. The "qubit" is the essential ingredient in the quantum Turing machine first postulated by Deutsch to elucidate the nuances of quantum computing [58].

The power of quantum computing accrues from two attributes: *quantum parallelism* and *entanglement*. Consider two electrons whose spin polarizations are made bistable by placing them in a magnetic field. If the system is classical, these two electrons can encode only two bits of information: the first one can encode either 0 or 1 (downspin or upspin), and the second one can do the same. By analogy, N number of electrons can encode N bits of information, as long as the system is classical.

But now consider the situation when the spin states of the two electrons are quantum mechanically entangled so that the two electrons can no longer be considered separately, but rather as one coherent unit. In that case, there are four possible states that this two-electron system can be in: both spins "up"; first spin "up" and the second spin "down"; the first spin "down" and the second spin "up"; and both spins "down". The corresponding "qubit" can be written as:

$$\begin{aligned} &qubit = a|\uparrow\uparrow\rangle + b|\uparrow\downarrow\rangle + c|\downarrow\uparrow\rangle + d|\downarrow\downarrow\rangle \\ &|a|^2 + |b|^2 + |c|^2 + |d|^2 = 1 \end{aligned} \tag{5.15}$$

Obviously, this system can encode four information states, as opposed to two. By analogy, if N qubits can be quantum mechanically entangled, then the system can encode 2^N bits of information, as opposed to simply N. This becomes a major advantage if N is large. Consider the situation when $N = 300$. There is no way in which a classical computer can be built that can handle 2^{300} bits of information as that number is larger than the number of electrons in the known universe. However, if just 300 electrons could be taken and their spins entangled (a very tall order), then 2^{300} bits of information could be encoded. Thus, entanglement bestows on a quantum computer tremendous information-handling capability.

The above must not be misconstrued to imply that a quantum computer is a "super-memory" that can store 2^N bits of information in N physical objects such as electrons. When a bit is "read", it always collapses to either 0 or 1. In Eq. (5.17), the probability that the qubit will be read as 0 is $|a|^2$, and the probability that it will be read as 1 is $|b|^2$. As either a 0 or a 1 is always read, a quantum computer allows *access* to no more than N bits of information. Thus, it is no better than a classical memory – in fact, it is worse! Because of the probabilistic nature of quantum mechanics, a stored bit will sometimes be read as 0 (with probability $|a|^2$) and sometimes as 1 (with probability $|b|^2 = 1 - |a|^2$). If repeated measurements are made of the stored bits, the exact same result will never be achieved every time, and thus the quantum system is not even a reliable memory.

The power of *entanglement* does not result in a super memory, but it is utilized in a different way, and is exploited in solving certain types of problem super efficiently.

Two well-known examples are Shor's quantum algorithm for factorization [59] and Grover's quantum algorithm for sorting [60]. Factorization is an *NP*-hard problem, but using Shor's algorithm, the complexity can be reduced to *P*. Grover's algorithm for sorting has a similar advantage. Suppose that there is a need to sort an *N*-body ensemble to find one odd object. By using the best classical algorithm, this will take $N/2$ tries, but using Grover's algorithm it will take only $\sqrt{N}$ tries. Thus, entanglement yields super-efficient algorithms that can be executed in a quantum processor where qubits are entangled. That is the advantage of quantum computing.

5.4.1 Quantum Parallelism

In classical computing, "parallelism" refers to the parallel (simultaneous) processing of different information in different processors. Quantum parallelism refers to the simultaneous processing of different information or inputs in the *same* processor. This idea, due to Deutsch, refers to the notion of evaluating a function once on a superposition of all possible inputs to the function to produce a superposition of outputs. Thus, all outputs are produced in the time taken to calculate one output classically. Of course, not all of these outputs are accessible since a measurement on the superposition state of the output will produce only one output. However, it is possible to obtain certain joint properties of the outputs [61], and that is a remarkable possibility.

Quantum parallelism may be illustrated with an example. Consider the situation when *M* inputs $x_1, x_2, x_3, \ldots x_M$ are provided to a computer, and their functions $f(x_1)$, $f(x_2)$, $\ldots f(x_M)$ are to be computed. The results are then fed to another computer to calculate the functional $F(f(x_1), f(x_2), \ldots f(x_M))$.

With a classical computer, $f(x_1), f(x_2), \ldots f(x_M)$ will be calculated serially, one after the other. However, with a quantum computer, the initial state will be prepared as a superposition of the inputs:

$$|I\rangle = \frac{1}{\sqrt{M}}(|x_1\rangle + |x_2\rangle + \ldots |x_M\rangle) \tag{5.16}$$

This input will then be fed to a quantum computer and allowed to evolve in time to produce the output

$$|O\rangle = \frac{1}{\sqrt{M}}(|f(x_1)\rangle + |f(x_2)\rangle + \ldots |f(x_M)\rangle) \tag{5.17}$$

The output $|O\rangle$ has been obtained in the time required to perform a single computation. Now, if the functional $F(f(x_1), f(x_2), \ldots f(x_M))$ can be computed from $|O\rangle$, then a quantum computer will be extremely efficient. This is an example where "quantum parallelism" can speed up the computation tremendously.

There are two questions, however. First, can the functional be computed from a knowledge of the superposition of various $f(x_i)$ and not the individual $f(x_i)$s? The answer is "yes", but for a small class of problems – the so-called *Deutsch–Josza* class of problems which can benefit from quantum parallelism. Second, can the functional be computed correctly with unit probability? The answer is "no". However, if the answer obtained in the first trial is wrong (hopefully, the computing entity can

distinguish right from wrong answers), then the experiment or computation is repeated until the right answer is obtained. The probability of getting the right answer within k iterations is $(1 - p^k)$ where p is the probability of getting the wrong answer in any iteration. The mean number of times the experiment should be repeated to get the correct answer is $M^2 - 2M - 2$.

The following is an example of the Deutsch–Josza class of problems. For integer $0 \leq x \leq 2L$, given that the function $f_{k(x)} \in [0,1]$ has one of two properties – (i) either $f_{k(x)}$ is independent of x; or (ii) one-half of the numbers $f_{k(0)}, f_{k(1)}, \ldots f_{k(2L-1)}$ are zero – determine which type the function belongs to, using the fewest computational steps.

The most efficient classical computer will require $L + 1$ evaluations whereas, according to Deutsch and Josza, a quantum computer can solve this problem with just two iterations.

5.4.2
Physical Realization of a Qubit: Spin of an Electron in a Quantum Dot

The all-important question that stirs physical scientists and engineers is: Which system is most appropriate to implement a qubit? It must be one where the phase relationship between the coefficients a and b in Eq. (5.17) are preserved for the longest time. Charge has a small phase-coherence time which saturates to about 1 ns as the temperature is lowered towards 0 K because of coupling to zero point motion of phonons [62]. Spin has a much longer phase-coherence time as it does not couple to phonons efficiently. As a result, the phase-coherence time may not rapidly decrease with increasing temperature. Measurements of the phase-coherence time (also called the transverse relaxation time, or T_2 time) in an ensemble of CdS quantum dots has recently been shown actually to increase with increasing temperature [63]. It is believed that this is because the primary phase relaxation mechanism for electron spins in these quantum dots is hyperfine interactions with nuclear spins. The nuclear spins are increasingly depolarized with increasing temperature, and that leads to an actual increase in the electron's spin coherence time with increasing temperature. Therefore, it is natural to encode a qubit by the coherent superposition of two spin polarizations of an electron.

In 1996, the idea was proposed of encoding a "qubit" by the spin of an electron in a quantum dot [42]. A simple spin-based "quantum inverter" was designed which utilized two exchange-coupled quantum dots. This was not a universal quantum gate, but relied on quantum mechanics to elicit the Boolean logic NOT function. The spin of the electron was used as a qubit. To the present author's knowledge, this was the first instance where the spin of an electron in a quantum dot was used to implement a qubit in a gate. This idea was inspired by single spin logic and so in many ways SSL could be regarded as the progeny of spin-based quantum computing.

Unlike SSL – which is purely classical and does not require "phase coherence" – quantum computing relies intrinsically on phase coherence and is therefore much more delicate. While the phase coherence of single spins can be quite long, it is doubtful that several entangled spins will have sufficiently long-lived phase coherence to allow a

significant number of computational steps to be carried out, particularly when gate operations are performed on them. At the time of writing, the realization of practical scalable quantum processors based on electron spins appears to be decades away.

5.5 Conclusions

In this chapter, an attempt has been made to distinguish between the two approaches to spintronics, namely "hybrid spintronics" and "monolithic spintronics". It is unlikely that the hybrid approach will bring about significant advances in terms of energy dissipation, speed, or any other metric. The monolithic approach, on the other hand, is more difficult, but also more likely to produce major advances. The SSL idea revisited here is a paradigm that may begin to bear fruit with the most recent advances in manipulating the spins of single electrons in quantum dots [64–71]. This is a classical model and does not require the phase coherence that is difficult to maintain in solid-state circuits. There is also no requirement to "entangle" several bits, but rather a need to exchange-couple two bits pairwise. Thus, SSL is much easier to implement than quantum processors based on single electron spins.

Acknowledgments

The author acknowledges useful discussions with Profs. Marc Cahay and Supriyo Datta. These studies were supported by the National Science Foundation under grant ECCS-0608854, and by the Air Force Office of Scientific Research under grant FA9550-04-1-0261.

References

1 S. Datta, B. Das, *Appl. Phys. Lett.* 1990, **56**, 665.
2 (a) J. Fabian, I. Zutic, S. Das Sarma, *Appl. Phys. Lett.* 2004, **84**, 85; (b) M. E. Faltte, Z. G. Yu, E. Johnston-Halperin, D. D. Awschalom, *Appl. Phys. Lett.* 2003, **82**, 4740.
3 S. Bandyopadhyay, *J. Nanosci. Nanotechnol.* 2007, **7**, 168.
4 S. Salahuddin, S. Datta, *Appl. Phys. Lett.* 2007, **90**, 093503.
5 (a) V. V. Zhirnov, R. K. Cavin, J. A. Hutchby, G. I. Bourianoff, *Proc. IEEE* 2003, **91**, 1934; (b) R. K. Cavin, V. V. Zhirnov, J. A. Hutchby, G. I. Bourianoff, *Fluctuations Noise Lett.* 2005, **5**, C29. See also Chapter 4 in Volume III of this series.
6 D. Nikonov, G. I. Bourianoff, P. Gargini, www.arXiv.org/cond-mat/0605298.
7 E. I. Rashba, *Sov. Phys. Semicond.* 1960, **2**, 1109.
8 S. Bandyopadhyay, M. Cahay, *Physica E* 2005, **25**, 399.
9 M. Cahay, S. Bandyopadhyay, *Phys. Rev. B* 2003, **68**, 115316.
10 M. Cahay, S. Bandyopadhyay, *Phys. Rev. B* 2004, **69**, 045303.
11 R. J. Elliott, *Phys. Rev.* 1954, **96**, 266.

12 (a) M. I. D'yakonov, V. I. Perel', *Sov. Phys. JETP* 1971, **33**, 1053; (b) M. I. D'yakonov, V. I. Perel', *Sov. Phys. Solid State* 1972, **13**, 3023.

13 S. Pramanik, S. Bandyopadhyay, M. Cahay, *IEEE Trans. Nanotech.* 2005, **4**, 2.

14 G. Dresselhaus, *Phys. Rev.* 1955, **100**, 580.

15 S. Bandyopadhyay, M. Cahay, *Appl. Phys. Lett.* 2004, **85**, 1814.

16 S. Bandyopadhyay, M. Cahay, *Appl. Phys. Lett.* 2004, **85**, 1433.

17 J. Nitta, T. Takazaki, H. Takayanagi, T. Enoki, *Phys. Rev. Lett.* 1997, **78**, 1335.

18 Suman Datta, Intel Corporation, private communication.

19 J. Schliemann, J. C. Egues, D. Loss, *Phys. Rev. Lett.* 2003, **90**, 146801.

20 X. Cartoixa, D. Z.-Y. Ting, Y.-C. Chang, *Appl. Phys. Lett.* 2003, **83**, 1462.

21 (a) K. C. Hall, W. H. Lau, K. Gundogdu, M. E. Flatte, T. F. Boggess, *Appl. Phys. Lett.* 2003, **83**, 2937; (b) K. C. Hall, K. Gundogdu, J. L. Hicks, A. N. Kocbay, M. E. Flatte, T. F. Boggess, K. Holabird, A. Hunter, D. H. Chow, J. J. Zink, *Appl. Phys. Lett.* 2005, **86**, 202114.

22 K. C. Hall, M. E. Flatte, *Appl. Phys. Lett.* 2006, **88**, 162503.

23 E. Safir, M. Shen, S. Siakin, *Phys. Rev. B* 2004, **70**, 241302(R).

24 G. Salis, R. Wang, X. Jiang, R. M. Shelby, S. S. P. Parkin, S. R. Bank, J. S. Harris, *Appl. Phys. Lett.* 2005, **87**, 262503.

25 S. Bandyopadhyay, M. Cahay, www.arXiv.org/cond-mat/0604532.

26 M. E. Flatte, K. C. Hall, www.arXiv.org/cond-mat/0607432.

27 P. A. Dowben, R. Skomski, *J. Appl. Phys.* 2004, **95**, 7453.

28 T. Koga, J. Nitta, H. Takayanagi, S. Datta, *Phys. Rev. Lett.* 2002, **88**, 126601.

29 S. Bandyopadhyay, M. Cahay, *Appl. Phys. Lett.* 2005, **86**, 133502.

30 S. Bandyopadhyay, B. Das, A. E. Miller, *Nanotechnology* 1994, **5**, 113.

31 D. A. Hodges, H. G. Jackson, *Analysis and Design of Digital Integrated Circuits*, 2nd edition, McGraw-Hill, New York, 1988, p. 2.

32 M. Ikezawa, B. Pal, Y. Masumoto, I. V. Ignatiev, S. Yu. Verbin, I. Ya. Gerlovin, *Phys. Rev. B* 2005, **72**, 153302.

33 (a) R. Hanson, L. H. Willems van Beveren, J. M. Elzerman, W. J. M. Naber, G. H. L. Koppens, L. P. Kouwenhoven, L. M. K. Vandersypen, *Phys. Rev. Lett.* 2005, **94**, 196802; (b) M. Kroutvar, Y. Ducommun, D. Heiss, M. Bichler, D. Schuh, G. Abstreiter, J. J. Finley, *Nature* 2004, **432**, 81; (c) S. Amasha, K. MacLean, I. Radu, D. M. Zumbuhl, M. A. Kastner, M. P. Hanson, A. C. Gossard, www.arXiv.org/cond-mat/0607110.

34 S. Pramanik, C.-G. Stefanita, S. Patibandla, S. Bandyopadhyay, K. Garre, N. Harth, M. Cahay, *Nature Nanotech.* 2007, **2**, 216.

35 E. Knill, *Nature* 2005, **434**, 39.

36 D. Rugar, R. Budakian, H. J. Mamin, B. H. Chui, *Nature* 2004, **430**, 329.

37 J. M. Elzerman, R. Hanson, L. H. Willems van Beveren, B. Witkamp, L. M. K. Vandersypen, L. P. Kouwenhoven, *Nature* 2004, **430**, 431.

38 M. Xiao, I. Martin, E. Yablonovitch, H. W. Jiang, *Nature* 2004, **430**, 435.

39 (a) H. Agarwal, S. Pramanik, S. Bandyopadhyay, *New J. Phys.* 2008, **10**, 015001. (b) S. Bandyopadhyay, M. Cahay, *An Introduction to Spintronics*, CRC Press, Boca Raton, 2007, Chapter 13.

40 (a) S. N. Molotkov, S. S. Nazin, *JETP Lett.* 1995, **62**, 273; (b) S. N. Molotkov, S. S. Nazin, *Phys. Low Dim. Struct.* 1997, **10**, 85; (c) S. N. Molotkov, S. S. Nazin, *Zh. Eksp. Teor. Fiz.* 1996, **110**, 1439.

41 A. M. Bychkov, L. A. Openov, I. A. Semenihin, *JETP Lett.* 1997, **66**, 298.

42 (a) S. Bandyopadhyay, V. P. Roychowdhury, Proceedings International Conference on Superlattice Microstructures, Liege, Belgium 1996; (b) S. Bandyopadhyay, V. P. Roychowdhury, *Superlatt. Microstruct.* 1997, **22**, 411.

43 S. K. Sarkar, T. Bose, S. Bandyopadhyay, *Phys. Low Dim. Struct.* 2006, **2**, 69.

44 T. Bose, S. K. Sarkar, S. Bandyopadhyay, *IEE Proc. - Circuits, Dev. Syst.* 2007, **1**, 194.

45 (a) D. V. Melnikov, J.-P. Leburton, *Phys. Rev. B* 2006, **73**, 155301; (b) J.-P. Leburton, private communication.

46 (a) P. Bakshi, D. Broido, K. Kempa, *J. Appl. Phys.* 1991, **70**, 5150; (b) K. Kempa, D. A. Broido, P. Bakshi, *Phys. Rev. B* 1991, **43**, 9343; The logic implementations were made available to the author in early 1992 by P. Bakshi in a private communication.

47 C. S. Lent, P. D. Tougaw, W. Porod, G. H. Bernstein, *Nanotechnology* 1993, **4**, 49.

48 S. Bandyopadhyay, V. P. Roychowdhury, D. B. Janes, in: M. A. Stroscio, M. Dutta (Eds.), *Quantum-Based Electronic Devices and Systems*, World Scientific, Singapore, 1998, Chapter 1.

49 R. P. Cowburn, M. E. Welland, *Science* 2000, **287**, 1466.

50 M. Anantram, V. P. Roychowdhury, *J. Appl. Phys.* 1999, **85**, 1622.

51 S. Bandyopadhyay, V. P. Roychowdhury, *Jpn. J. Appl. Phys.* 1996, **35** (Part 1), 3350.

52 S. Bandyopadhyay, *Superlatt. Microstruct.* 2005, **37**, 77.

53 L. B. Kish, *Phys. Lett. A* 2002, **305**, 144.

54 The International Technology Roadmap for Semiconductors published by the Semiconductor Industry Association. http://public.itrs.net/.

55 C. F. Hirjibehedin, C. P. Lutz, A. J. Heinrich, *Science* 2006, **312**, 1021.

56 C. S. Lent, M. Liu, Y. Lu, *Nanotechnology* 2006, **17**, 4240.

57 P. P. Gelsinger, Proceedings, IEEE International Solid State Circuits Conference, IEEE Press, 2001.

58 D. Deutsch, *Proc. Royal Soc. London, Ser. A* 1985, **400**, 97.

59 P. W. Shor, in: Proceedings 37th Annual Symposium Foundations of Computer Science, IEEE Computer Society Press, p. 56, 1996.

60 (a) L. K. Grover, *Phys. Rev. Lett.* 1997, **79**, 325; (b) L. K. Grover, *Phys. Rev. Lett.* 1997, **79**, 4709.

61 R. Josza, *Proc. Royal Soc. London Ser. A* 1991, **435**, 563.

62 P. Mohanty, E. M. Q. Jariwalla, R. A. Webb, *Phys. Rev. Lett.* 1997, **78**, 3366.

63 S. Pramanik, B. Kanchibotla, S. Bandyopadhyay, Proceedings IEEE NANO 2006 Conference, Cincinnati, 2006.

64 M. Ciorga, A. S. Sachrajda, P. Hawrylak, C. Gould, P. Zawadzki, S. Jullian, Y. Feng, Z. Wasilewski, *Phys. Rev. B* 2000, **61**, R16315.

65 M. Piero-Ladriere, M. Ciorga, J. Lapointe, P. Zawadzki, M. Korukusinski, P. Hawrylak, A. S. Sachrajda, *Phys. Rev. Lett.* 2003, **91**, 026803.

66 C. Livermore, C. H. Crouch, R. M. Westervelt, K. L. Campman, A. C. Gossard, *Science* 1996, **274**, 1332.

67 T. H. Oosterkamp, T. Fujisawa, W. G. van der Wiel, K. Ishibashi, R. V. Hijman, S. Tarucha, L. P. Kouwenhoven, *Nature* 1998, **395**, 873.

68 A. W. Holleitner, R. H. Blick, A. K. Huttel, K. Eberl, J. P. Kotthaus, *Science* 2001, **297**, 70.

69 N. J. Craig, J. M. Taylor, E. A. Lester, C. M. Marcus, M. P. Hanson, A. C. Gossard, *Science* 2004, **304**, 565.

70 R. Hanson, B. Witkamp, L. M. K. Vandersypen, L. H. W. van Beveren, J. M. Elzerman, L. P. Kouwenhoven, *Phys. Rev. Lett.* 2003, **91**, 196802.

71 J. R. Petta, A. C. Johnson, J. M. Taylor, E. A. Laird, A. Yacoby, M. D. Lukin, C. M. Marcus, M. P. Hanson, A. C. Gossard, *Science* 2005, **309**, 2180.

6
Organic Transistors

Hagen Klauk

6.1
Introduction

Organic transistors are metal-insulator-semiconductor (MIS) field-effect transistors (FETs) in which the semiconductor is not an inorganic crystal, but a conjugated organic material. The fact that organic materials can be semiconductors may initially surprise, as most organic materials encountered today are excellent insulators. The fundamental property that leads to electrical conduction in carbon-based solids is conjugation – that is, the presence of alternating single and double bonds between neighboring carbon atoms (see Figure 6.1). Conjugation results in the delocalization of the π-electrons over the entire molecule – or at least over the conjugated portion of the molecule – and this allows electronic charge to be transported along the molecule.

Electrical conductivity in conjugated organic materials has been studied for a century, yet a complete picture of the charge transport physics in organics is still evolving. A central observation is that the intermolecular bonds in organic solids are not covalent bonds, but much weaker van der Waals interactions. As a consequence, electronic states are not delocalized over the entire solid, but are localized to a single molecule (or the conjugated portion of the molecule). Charge transport through organic solids is therefore limited by trapping in localized states, and likely involves some form of hopping between molecules. This means that the carrier mobilities in organic semiconductors are expected to be much smaller than the mobilities in inorganic semiconductor crystals.

In fact, carrier mobilities observed in organic solids vary greatly depending on the choice of material, its chemical purity, and the degree of molecular order in the solid (which determines the orbital overlap between neighboring molecules). Semiconducting polymers that arrange in amorphous films when prepared from solution usually have room-temperature mobilities in the range of 10^{-6} to 10^{-3} $cm^2 V^{-1} s^{-1}$. (For comparison, the carrier mobilities in single-crystalline silicon are near 10^3 $cm^2 V^{-1} s^{-1}$.) Through molecular engineering, improved purification, and better control of the film deposition (so that charges can be transferred more easily

Nanotechnology. Volume 4: Information Technology II. Edited by Rainer Waser

ISBN: 978-3-527-31737-0

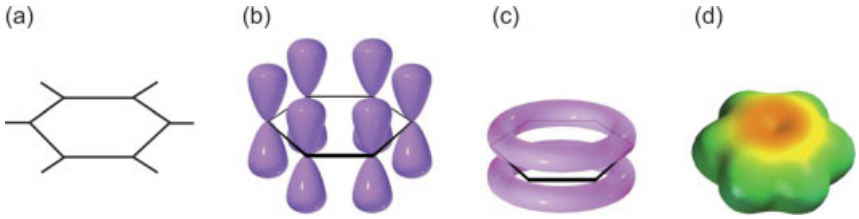


Figure 6.1 The concept of delocalized electrons in a conjugated molecule. (a) The carbon–carbon and carbon–hydrogen σ bonds in benzene. (b) The p orbital on each carbon can overlap with two adjacent p orbitals. (c) The clouds of electrons above and below the molecular plane. (d) The electrostatic potential map of benzene, showing that all the carbon–carbon bonds have the same electron density. (Figure adapted from: P. Y. Bruice, *Organic Chemistry*, Pearson Education, Upper Saddle River, NJ, USA.)

between molecules), the mobilities of conjugated polymers can be increased to about $0.5\,cm^2\,V^{-1}\,s^{-1}$. Small-molecule materials, on the other hand, often spontaneously arrange themselves into semicrystalline films when deposited by vacuum sublimation, and this results in room-temperature mobilities as large as $7\,cm^2\,V^{-1}s^{-1}$. Reports on carefully prepared single crystals of ultrapurified naphthalene suggest that mobilities at cryogenic temperatures can be as large as $400\,cm^2\,V^{-1}\,s^{-1}$ [1].

It is generally agreed that no single transport model can account for this wide a range of observed mobilities. Also, the temperature dependence of the mobility can be quite different for different organic materials, and in some cases different temperature-dependent mobility behavior has been observed even for the same organic material. Consequently, several different models for charge transport in organics have been proposed.

The model of *variable-range hopping* (VRH) assumes that carriers hop between localized electronic states by tunneling through energy barriers, and that the probability of a hopping event is determined by the hopping distance and by the energy distribution of the states. Specifically, carriers either hop over short distances with large activation energies, or over long distances with small activation energies. Since the tunneling is thermally activated, the mobility increases with increasing temperature. With increasing gate voltage, carriers accumulated in the channel fill the lower-energy states, thus reducing the activation energy and increasing the mobility. As Vissenberg and Matters have shown [2], the tunneling probability depends heavily on the overlap of the electronic wave functions of the hopping sites, which is consistent with the observation that the carrier mobility is significantly greater in materials with a larger degree of molecular ordering. Thus, the mobility is dependent on temperature, gate voltage, and molecular ordering (see Figure 6.2). The variable-range hopping model is usually discussed in the context of low-mobility amorphous semiconductor films (with room-temperature mobility less than about $10^{-2}\,cm^2\,V^{-1}\,s^{-1}$).

The *multiple trapping and release* (MTR) model adapted for organic transistors by Gilles Horowitz and coworkers [3] is based on the assumption that most of the charge carriers in the channel are trapped in localized states, and that carriers cannot move directly from one state to another. Instead, carriers are temporarily promoted to an

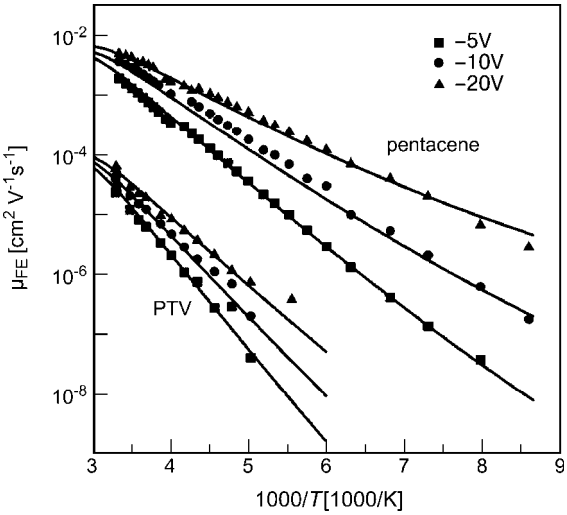


Figure 6.2 Temperature-dependent and gate voltage-dependent carrier mobility in a disordered polythienylenevinylene (PTV) film and in a solution-processed pentacene film. (Reproduced with permission from Ref. [2].)

extended-state band in which charge transport occurs. The number of carriers available for transport then depends on the difference in energy between the trap level and the extended-state band, as well as on the temperature and on the gate voltage (see Figure 6.3). The MTR model is generally considered to apply to materials with a significant degree of molecular ordering, such as polycrystalline films of small-molecule and certain polymeric semiconductors – that is, to materials which show room-temperature mobilities approaching or exceeding $0.1\,\mathrm{cm^2\,V^{-1}\,s^{-1}}$. Carrier mobilities this large are not easily explained in the framework of the VRH model. On the other hand, the existence of an extended-state transport band in organic semiconductors as postulated by the MTR model is still the subject of debate.

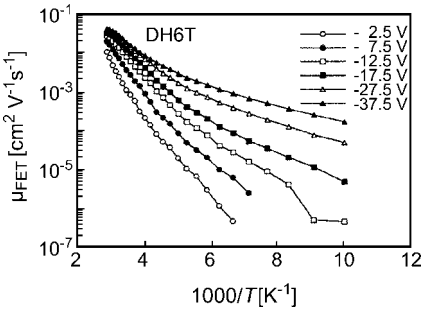


Figure 6.3 Temperature-dependent and gate voltage-dependent carrier mobility in a vacuum-deposited polycrystalline dihexylsexithiophene (DH6T) film. (Reproduced with permission from Ref. [3].)

6.2 Materials

Organic semiconductors essentially are available as two types: *conjugated polymers* and *conjugated small-molecule materials*. The prototypical semiconducting polymer is polythiophene (Figure 6.4a). While genuine polythiophene is insoluble and thus difficult to deposit in the form of thin films, alkyl-substituted polythiophenes, such as poly(3-hexylthiophene) (P3HT) (Figure 6.4b) have excellent solubility in a variety of organic solvents, and thin films are readily prepared by spin-coating, dip-coating, drop-coating, screen printing, or inkjet printing. Polythiophene was one of the first polymers to be used for organic transistors [4, 5], and remains one of the most popular semiconductors in organic electronics.

Generic polythiophenes usually form amorphous films with virtually no long-range structural order, very short π-conjugation length, and consequently poor mobilities, typically below $10^{-3}\,cm^2\,V^{-1}\,s^{-1}$. Obtaining usefully large mobilities in the polythiophene system requires highly purified derivatives which have been specifically synthesized to allow the molecules to self-organize into crystalline structures with a high degree of molecular order. An early example of such an engineered polythiophene is regioregular (RR) head-to-tail (HT) P3HT, initially synthesized by McCullough and coworkers in 1993 [6] and first employed for transistor fabrication by Bao and colleagues in 1996 [7]. In RR-HT-P3HT, the strong interactions between the regularly oriented alkyl side chains lead to a three-dimensional (3-D) lamellar structure in which the thienylene moieties along the polymer backbone are held in coplanarity (see Figure 6.5). The coplanarity of the thienylene moieties greatly increases the extent of π-conjugation, one consequence of which is a substantially increased carrier mobility (0.05 to $0.1\,cm^2\,V^{-1}\,s^{-1}$) compared with regiorandom P3HT (less than $10^{-3}\,cm^2\,V^{-1}\,s^{-1}$).

The microstructure of regioregular P3HT, its dependence on the degree of regioregularity, molecular weight, and deposition conditions, and the relationship between microstructure and carrier mobility, has been studied in detail by Sirringhaus and coworkers [8]. These authors found that the orientation of the microcrystalline lamellar domains with respect to the substrate surface is influenced by the molecular weight (i.e. the average polymer chain length), by the degree of

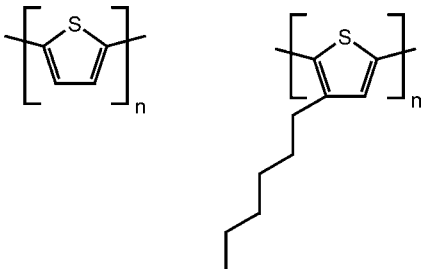


Figure 6.4 The chemical structures of polythiophene (left) and poly(3-hexylthiophene) (P3HT) (right).

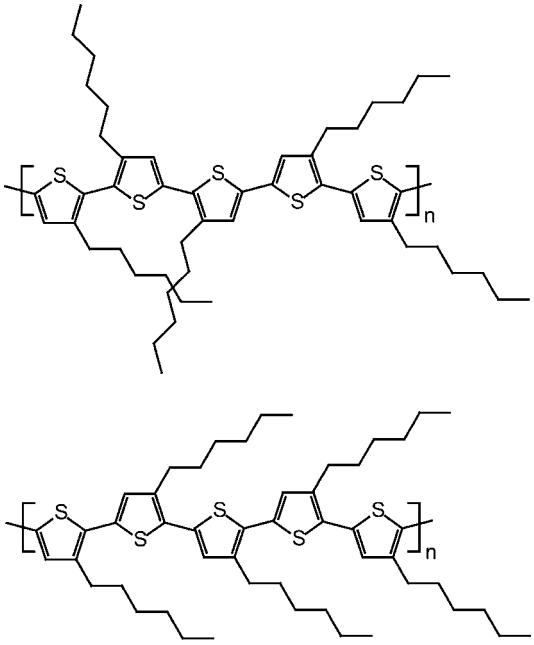


Figure 6.5 Schematic representation of regiorandom P3HT (left) and regioregular P3HT (right).

regioregularity, and by the deposition conditions (i.e. whether the film formation occurred quickly or slowly). The formation of ordered lamellae leads to a substantial overlap of the π-orbitals of neighboring molecules, but only in the direction perpendicular to the lamella plane. As a result, charge carrier transport and mobility in ordered P3HT films is highly anisotropic. In field-effect transistors (FETs), current flows parallel to the substrate, so the orientation of the lamellae with respect to the substrate surface is critical for the electrical performance of the transistors. Sirringhaus et al. were able to show that the transistor-friendly orientation of the lamellae (shown in Figure 6.6a) can be induced by selecting a polymer with a high degree of regioregularity (see Figure 6.6b) and, to a lesser extent, by choosing deposition conditions that favor a slow crystallization of the film.

Unfortunately, the large extent of π-conjugation in regioregular P3HT also leads to a significantly reduced ionization potential that makes the material very susceptible to photoinduced oxidation. This explains the commonly observed instability of P3HT transistors when operated in ambient air without encapsulation. A successful route to environmentally more stable self-organizing, high-mobility polythiophene derivatives was devised by Ong and coworkers [9]. This group recognized that the strategic placement of unsubstituted moieties along the polymer backbone and the resulting torsional deviations from coplanarity would reduce the effective π-conjugation length sufficiently to increase the ionization potential (and thus greatly improve oxidation resistance and environmental stability) while compromising the mobility only slightly, if at all. One particularly successful material which has emerged from this line of study is poly(3,3′′′′-didodecylquaterthiophene), better

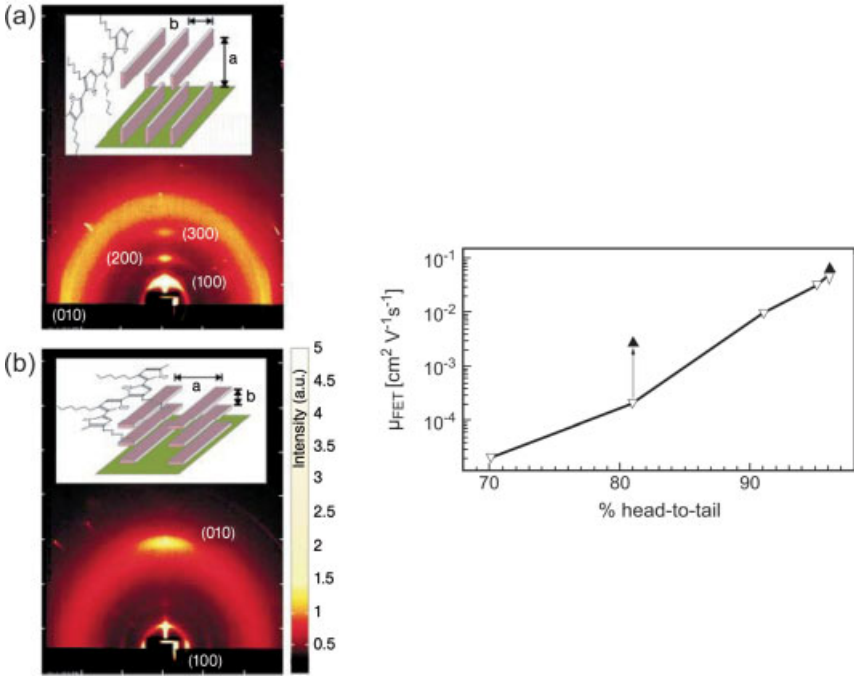


Figure 6.6 Left: Upon deposition on a flat substrate, regioregular poly(3-hexylthiophene) (P3HT) forms ordered lamellar domains, the orientation of which depends on the degree of regioregularity, molecular weight, and deposition conditions. Right: Relationship between the degree of regioregularity (quantified as the head-to-tail ratio) and the carrier mobility of P3HT transistors. (Reproduced with permission from Ref. [8].)

known as PQT-12 (see Figure 6.7). PQT-12 has shown air-stable carrier mobilities as large as $0.2\,cm^2\,V^{-1}\,s^{-1}$, and has been employed successfully in the fabrication of functional organic circuits and displays.

To further improve the performance and stability of alkyl-substituted polythiophenes, researchers at Merck Chemicals incorporated thieno[3,2-*b*]thiophene moieties into the polymer backbone [10] (see Figure 6.8). The effect of this is two-fold:

- The delocalization of carriers from the fused aromatic unit is less favorable than from a single thiophene unit, so the effective π-conjugation length is further reduced and the ionization potential becomes even larger than for polyquaterthiophene.
- The rotational invariance of the thieno[3,2-*b*]thiophene in the backbone promotes the formation of highly ordered crystalline domains with an extent not previously seen in semiconducting polymers. The molecular ordering is induced by annealing the spun-cast films in their liquid-crystalline phase and subsequent crystallization upon cooling. Carrier mobilities of $0.5\,cm^2\,V^{-1}\,s^{-1}$ have been reported.

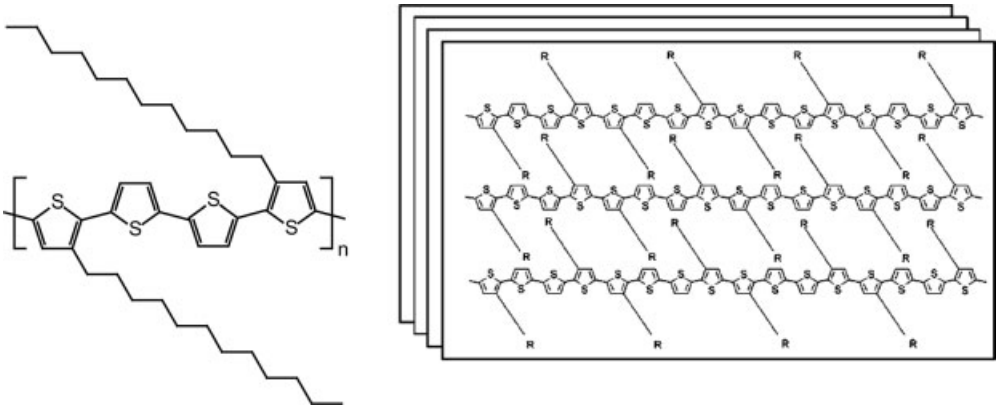


Figure 6.7 Left: Chemical structure of poly(3,3′′′′-didodecylquaterthiophene) (PQT-12). Right: A schematic representation of the lamellar π-stacking arrangement. (Reproduced with permission from Ref. [9].)

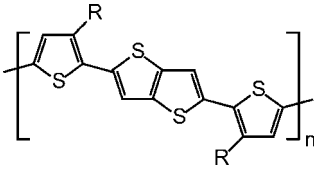


Figure 6.8 The chemical structure of poly(2,5-bis(3-alkylthiophen-2-yl)thieno[3,2-*b*]thiophene (PBTTT).

Among the small-molecule organic semiconductors, the most widely studied systems include pentacene, sexithiophene, copper phthalocyanine, and the fullerene C_{60} (see Figure 6.9). Many small-molecule materials are insoluble in common organic solvents, but they often can be conveniently deposited using vacuum-deposition methods, such as thermal evaporation or organic vapor-phase deposition. In most cases, small-molecule organic semiconductors readily self-organize into semicrystalline films with a significant degree of molecular order (see Figure 6.10).

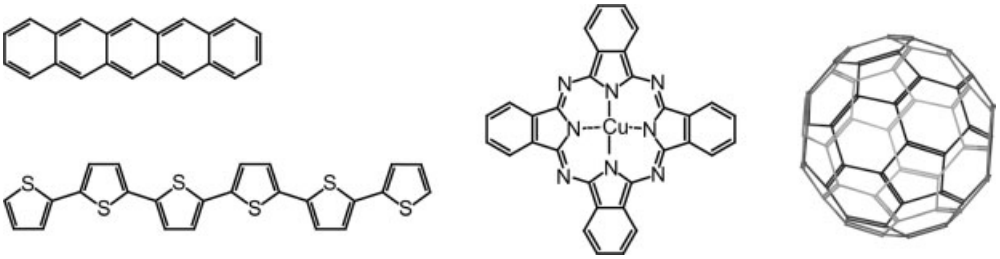


Figure 6.9 Chemical structures of pentacene (top left), sexithiophene (bottom left), copper phthalocyanine (center), and the fullerene C_{60} (right).

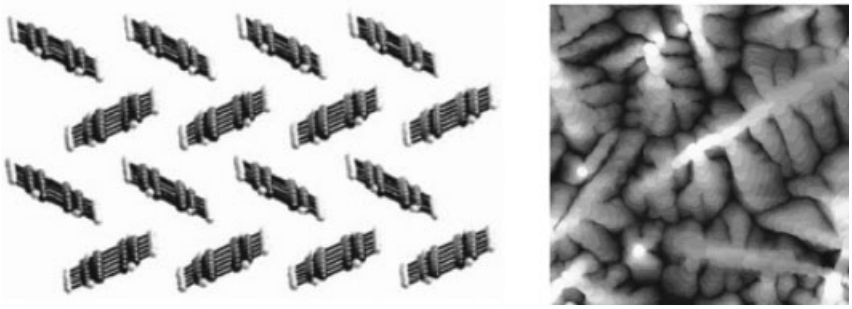


Figure 6.10 Pentacene self-organizes into an edge-to-face, or herringbone structure, forming semicrystalline films when deposited by evaporation onto amorphous substrates (left part of figure from: J. E. Anthony et al., Engineered pentacenes, in: *Organic Electronics*, Wiley-VCH, 2006.)

The use of vacuum-grown films of conjugated small-molecule materials for organic transistors was pioneered during the late 1980s by Madru and Clarisse (using metal phthalocyanines [11, 12]) and by Horowitz and Garnier (using oligothiophenes [13, 14]). Initial carrier mobilities were around $10^{-3}\,cm^2\,V^{-1}\,s^{-1}$, but these quickly improved to about $0.1\,cm^2\,V^{-1}\,s^{-1}$. In 1996, Jackson predicted and demonstrated that the carrier mobility of many organic semiconductors could be substantially improved by growing the films on low-energy surfaces [15, 16]. Inorganic dielectrics, such as silicon dioxide, are usually characterized by large surface energies favoring two-dimensional (2-D) growth of the first organic layer. Two-dimensional film growth typically results in large crystalline grains, and it was long believed that this was desirable to achieve good transistor performance. The surface energy of inorganic dielectrics is readily reduced by covering the surface with a self-assembled monolayer (SAM) of a methyl-terminated alkylsilane, such as octadecyltrichlorosilane (OTS). Organic film growth on low-energy SAM surfaces is distinctly three-dimensional, with much smaller grains and significantly more grain boundaries, yet the transistor mobilities were found to be significantly larger (by as much as an order of magnitude) compared with the large-grain films on high-energy surfaces. One explanation for this apparent discrepancy is that 2-D growth results in voids between disconnected grains, reducing the effective channel width of the transistor, and that such voids are efficiently filled when 3-D growth is favored [17, 18]. Carrier mobilities on high-energy surfaces (such as bare oxides) peak around $0.5\,cm^2\,V^{-1}\,s^{-1}$, while mobilities on low-energy surfaces (SAM-treated oxides or polymer dielectrics) have reached $1\,cm^2\,V^{-1}\,s^{-1}$ for alkyl-substituted oligothiophenes [19] and $5\,cm^2\,V^{-1}\,s^{-1}$ for pentacene [20–22].

An interesting alternative to solution-processed polymers and vacuum-grown small-molecules was developed by Herwig and Müllen during the early 1990s in the form of solution-deposited pentacene [23]. The initial rationale was to combine the best of two worlds – that is, the simplicity of solution-processing with the large carrier mobility of pentacene. Herwig and Müllen (and later Afzali and coworkers [24]; see Figure 6.11) synthesized a soluble pentacene precursor that was

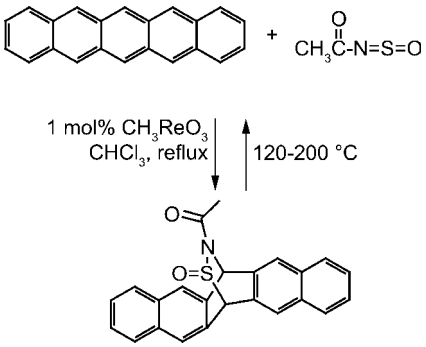


Figure 6.11 Synthesis of a soluble pentacene precursor and thermally induced conversion of the precursor to pentacene. (Reproduced with permission from Ref. [24].)

spin-coated and subsequently converted to pentacene at elevated temperature. Carrier mobilities for thermally converted pentacene are between 0.1 and 1 $cm^2 V^{-1} s^{-1}$, depending on the conversion temperature (130 to 200 °C).

The concept of solution-processable, high-mobility, small-molecule organic semiconductors was further developed by Anthony and colleagues, who designed and synthesized a number of soluble pentacene and anthradithiophene derivatives that do not require chemical conversion after deposition. Two examples – triisopropylsilyl (TIPS) pentacene and triethylsilyl (TES) anthradithiophene – are shown in Figure 6.12 [25]. In addition to providing solubility in common organic solvents, the functionalization of pentacene and anthradithiophene at the center rings can be utilized to strategically tune the molecular packing in the solid state in order to induce π-stacking with reduced intermolecular distances. With optimized solution-deposition, carrier mobilities as large as 3 $cm^2 V^{-1} s^{-1}$ have been achieved with these materials [26–28].

Organic transistors prepared with any of the semiconductors discussed so far operate efficiently only as p-channel transistors; that is, the drain currents in these

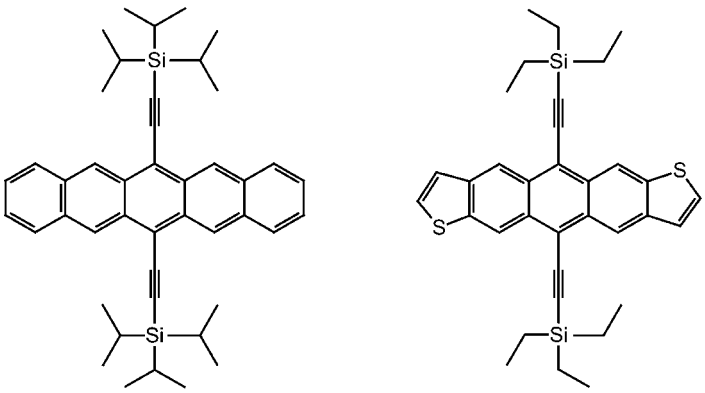


Figure 6.12 Left: Triisopropylsilyl (TIPS) pentacene. Right: Triethylsilyl (TES) anthradithiophene.

transistors are almost exclusively due to positively charged carriers. Currents due to negatively charged carriers are almost always extremely small in these materials, and often too small to be measurable, even when a large positive gate potential is applied to induce a channel of negatively charged carriers. Several explanations for the highly unbalanced currents have been suggested. One explanation is that the difference in mobility between the two carrier types is very large, due either to different scattering probabilities or perhaps to different probabilities for charge trapping, either at grain boundaries within the film or at defects at the dielectric interface [29, 30]. Another explanation is that charge injection from the contacts is highly unbalanced due to different energy barriers for positively and negatively charged carriers.

Interestingly, there are a number of organic semiconductors that show usefully large mobilities for negatively charged carriers. These materials include perfluorinated copper phthalocyanine (F_{16}CuPc), a variety of naphthalene and perylene tetracarboxylic diimide derivates, fluoroalkylated oligothiophenes, and the fullerene C_{60}. The carrier mobilities measured in n-channel FETs based on these materials are in the range of 0.01 to 1 $cm^2 V^{-1} s^{-1}$. Some of these materials are very susceptible to redox reactions and thus have poor environmental stability. For example, C_{60} shows mobilities as large as 5 $cm^2 V^{-1} s^{-1}$ when measured in ultra-high vacuum [31, 32], but when exposed to air the mobility drops rapidly by as many as four or five orders of magnitude. Similar degradation has been reported for some naphthalene and perylene tetracarboxylic diimide derivatives [33, 34]. Other materials, in particular F_{16}CuPc [35], have been found to be very stable in air, although the exact mechanisms that determine the degree of air stability are still unclear. Air-stable n-channel organic FETs with carrier mobilities as large as 0.6 $cm^2 V^{-1} s^{-1}$ [36] have been reported.

6.3 Device Structures and Manufacturing

From a technological perspective, the most useful organic transistor implementation is the thin-film transistor (TFT). The TFT concept was initially proposed and developed by Weimer during the 1960s for transistors based on polycrystalline inorganic semiconductors, such as evaporated cadmium sulfide [37]. The idea was later extended to TFTs based on plasma-enhanced chemical-vapor deposited (PECVD) hydrogenated amorphous silicon (a-Si:H) TFTs [38]. Today, a-Si:H TFTs are widely employed as the pixel drive devices in active-matrix liquid-crystal displays (AMLCDs) which accounted for $50 billion in global sales in 2005. Organic TFTs were first reported during the 1980s [4, 39]. To produce an organic TFT, the organic semiconductor and other materials required (gate electrode, gate dielectric, source and drain contacts) are deposited as thin layers on the surface of an electrically insulating substrate, such as glass or plastic foil. Depending on the sequence in which the materials are deposited, three organic TFT architectures can be distinguished, namely inverted staggered, inverted coplanar, and top-gate (see Figure 6.13).

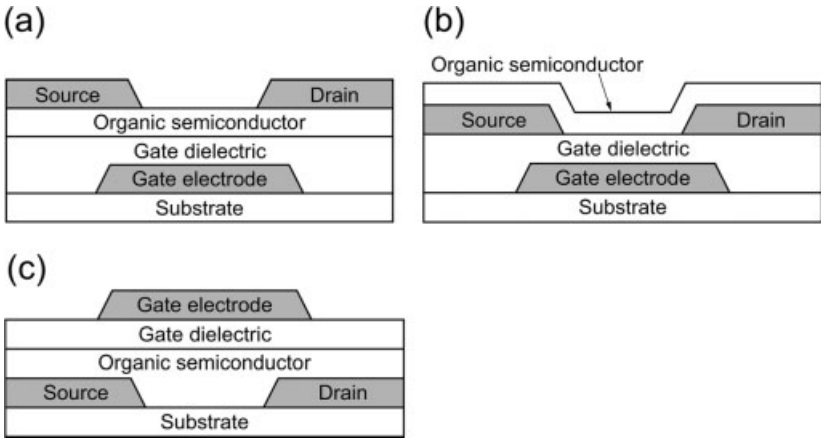


Figure 6.13 Schematic cross-sections of the (a) inverted staggered, (b) inverted coplanar, and (c) top-gate organic TFT structures.

Each of these structures has certain advantages and disadvantages:

- The inverted coplanar architecture allows for the use of photolithography or solution-based printing to pattern source and drain contacts with high resolution and short contact spacing, without exposing the organic semiconductor to potentially harmful process chemicals such as organic solvents, photoresists, and etchants [40]. However, as the contacts in organic transistors are not easily doped, the coplanar structure is often associated with larger contact resistance compared with the staggered architecture [41].
- The inverted staggered TFT structure is usually implemented by evaporating the source/drain metal through a shadow mask (also called an aperture or stencil mask). In this way, the organic semiconductor is not exposed to process chemicals, and the contact resistance can be quite low, as the entire area underneath the contacts is available for charge injection, though with shadow masks it is more difficult to reduce the channel length reliably below about 10 μm.
- For the top-gate structure the source and drain contacts can be patterned with high resolution prior to the deposition of the semiconductor, and the contact resistance can be as low as for the inverted staggered architecture. However, the deposition of a gate dielectric and a gate electrode on top of the semiconductor layer means that great care must be exercised to avoid process-induced degradation of the organic semiconductor, and the possibility of material mixing at the semiconductor/dielectric interface must be taken into account.

A variety of methods exists for the deposition and patterning of the individual layers of the TFT. For example, gate electrodes and source and drain contacts are often made using vacuum-deposited inorganic metals. Non-noble metals, such as aluminum or chromium, are suitable for the gate electrodes in the inverted device

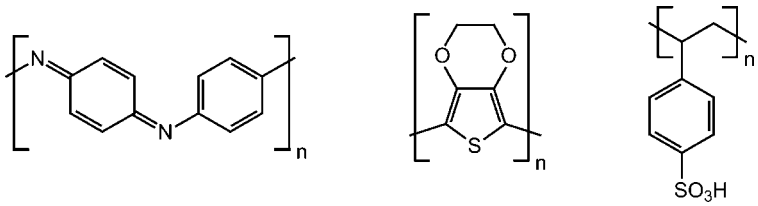


Figure 6.14 Left: Polyaniline. Right: Poly(3,4-ethylenedioxythiophene) (PEDOT) and poly (styrene sulfonic acid) PSS.

structures, as these metals have excellent adhesion on glass substrates. Noble metals (most notably gold) are a popular choice for the source and drain contacts, as they tend to give lower contact resistance than other metals. The metals are conveniently deposited by evaporation in vacuum and can be patterned either by photolithography and etching or lift-off, by lithography using an inkjet-printed wax-based etch resist [42], or simply by deposition through a shadow mask [43].

An alternative to inorganic metals are conducting polymers, such as polyaniline and poly(3,4-ethylenedioxythiophene):poly(styrene sulfonic acid) (PEDOT:PSS; see Figure 6.14). These are chemically doped conjugated polymers that have electrical conductance in the range between 0.1 and 1000 S cm^{-1}. The way in which continuous advances in synthesis and material processing have improved the conductivity of Baytron® PEDOT:PSS over the past decade is shown in Figure 6.15. Unlike inorganic metals, conducting polymers can be processed either from organic solutions or from aqueous dispersions, so gate electrodes and source and drain contacts for organic TFTs can be prepared by spin-coating and photolithography [44], or by inkjet-printing [45].

One important aspect in organic TFT manufacturing is the choice of the gate dielectric. Depending on the device architecture (inverted or top-gate), the dielectric material and the processing conditions (temperature, plasma, organic solvents, etc.)

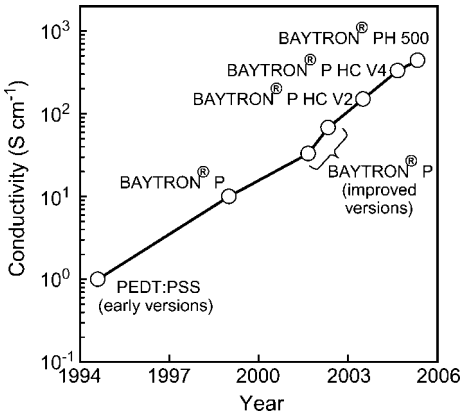


Figure 6.15 Improvements in the electrical conductivity of the conducting polymer PEDOT:PSS. (Adapted from a graph kindly provided by Stephan Kirchmeyer, H. C. Starck, Leverkusen, Germany.)

must be compatible with the previously deposited device layers and with the substrate. For example, chemical-vapor-deposited (CVD) silicon oxide and silicon nitride – which are popular gate dielectric materials for inorganic (amorphous or polycrystalline silicon) TFTs – may not be suitable for use on flexible polymeric substrates, as the high-quality growth of these films often requires temperatures that exceed the glass transition temperature of many polymeric substrate materials.

The thickness of the gate dielectric layer is usually a compromise between the requirements for large gate coupling, low operating voltages, and small leakage currents. Large gate coupling (i.e. a large dielectric capacitance) means that the transistors can be operated with low voltages, which is important when the TFTs are used in portable or handheld devices that are powered by small batteries or by near-field radio-frequency coupling, for example. Also, a large dielectric capacitance ensures that the carrier density in the channel is controlled by the gate and not by the drain potential, which is especially critical for short-channel TFTs. One way to increase the gate dielectric capacitance is to employ a dielectric with larger permittivity ε ($C = \varepsilon/t$). However, as Veres [46], Stassen [47], and Hulea [48] have pointed out, the carrier mobility in organic field effect transistors is systematically reduced as the permittivity of the gate dielectric is increased, presumably due to enhanced localization of carriers by local polarization effects (see Figure 6.16).

Alternatively, low-permittivity dielectrics with reduced thickness or thin multilayer dielectrics with specifically tailored properties may be employed. The greatest concern with thin dielectrics is the inevitable increase in gate leakage due to defects and quantum-mechanical tunneling as the dielectric thickness is reduced. A number of promising paths towards high-quality thin dielectrics with low gate leakage for low-voltage organic TFTs have recently emerged. One such approach is the anodization of aluminum, which has resulted in high-quality aluminum oxide films thinner than 10 nm providing a capacitance around $0.4\,\mu F\,cm^{-2}$. Combined with an ultra-thin molecular SAM, such dielectrics can provide sufficiently low leakage currents to allow the fabrication of functional low-voltage organic TFTs with large carrier

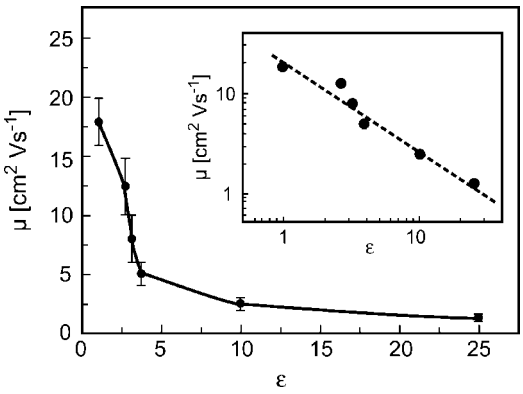


Figure 6.16 Relationship between the permittivity of the gate dielectric and the carrier mobility in the channel. (Reproduced with permission from Ref. [47].)

mobility [49]. Another path is the use of very thin crosslinked polymer films prepared by spin-coating [50, 51]. With a thickness of about 10 nm, these dielectrics provide capacitances as large as $0.3\,\mu F\,cm^{-2}$ and excellent low-voltage TFT characteristics. Finally, the use of high-quality insulating organic SAMs or multilayers provides a promising alternative [52–54]. Such molecular dielectrics typically have a thickness of 2 to 5 nm and a capacitance between 0.3 and $0.7\,\mu F\,cm^{-2}$, depending on the number and structure of the molecular layers employed, and they allow organic TFTs to operate with voltages between 1 and 3 V.

6.4 Electrical Characteristics

Like silicon metal-oxide-semiconductor-field-effect-transistors (MOSFETs), organic TFTs are metal-insulator-semiconductor FETs in which a sheet of mobile charge carriers is induced in the semiconductor by applying an electric field across the gate dielectric. Silicon MOSFETs normally operate in inversion – that is, the drain current is due to minority carriers generated by inverting the conductivity at the semiconductor/dielectric interface from p-type to n-type (for n-channel MOSFETs) or from n-type to p-type (for p-channel MOSFETs). The contact regions of silicon MOSFETs are heavily doped so that minority carriers are easily injected at the source and extracted at the drain, while the undesirable flow of majority carriers from drain to source is efficiently blocked by a space charge region.

Organic TFTs typically utilize intrinsic semiconductors. Positively charged carriers are accumulated near the dielectric interface when a negative gate-source voltage is applied, or negative charges are accumulated with a positive gate bias. Source and drain are usually implemented by directly contacting the intrinsic semiconductor with a metal. Depending on the choice of materials for the semiconductor and the metal, organic TFTs may operate as p-channel, n-channel, or ambipolar transistors. In p-channel and n-channel TFTs, the transport of one type of carrier is far more efficient than that of the other carrier type, either because the semiconductor/metal contacts greatly favor the injection or extraction of one carrier type over the other, or because the mobilities in the semiconductor are very different (perhaps due to different trapping rates), or because the electronic properties of the semiconductor/dielectric interface allow the accumulation of only one type of carrier, but not the other. In ambipolar organic TFTs the injection and transport of both positive and negative carriers is possible and both carrier types contribute to the drain current.

Despite the fact that the transport physics in organic transistors is different from that in silicon MOSFETs, the current–voltage characteristics can often be described with the same formalism:

$$I_D = \frac{\mu C_{diel} W}{L}\left((V_{GS} - V_{th})V_{DS} - \frac{V_{DS}^2}{2}\right) \quad \text{for } |V_{GS} - V_{th}| > |V_{DS}| \tag{6.1}$$

$$I_D = \frac{\mu C_{diel} W}{2L}(V_{GS} - V_{th})^2 \quad \text{for } |V_{DS}| > |V_{GS} - V_{th}| > 0 \tag{6.2}$$

Equation (6.1) describes the relationship between the drain current I_D, the gate-source voltage V_{GS} and the drain-source voltage V_{DS} in the linear regime, while Eq. (6.2) relates I_D, V_{GS} and V_{DS} in the saturation regime. C_{diel} is the gate dielectric capacitance per unit area, μ is the carrier mobility, W is the channel width, and L is the channel length of the transistor. For silicon MOSFETs, the threshold voltage V_{th} is defined as the minimum gate-source voltage required to induce strong inversion. Although this definition cannot strictly be applied to organic TFTs, the concept is nonetheless useful, as the threshold voltage conveniently marks the transition between the different regions of operation.

Figure 6.17 shows the current–voltage characteristics of an organic TFT that was manufactured on a glass substrate using the inverted staggered device structure (see Figure 6.13a) with a thin layer of vacuum-evaporated pentacene as the semiconductor, a self-assembled monolayer gate dielectric, and source/drain contacts prepared by evaporating gold through a shadow mask [54]. The device operates as a p-channel transistor with a threshold voltage of -1.2 V. By fitting the current–voltage characteristics to Eqs. (6.1) or (6.2), the carrier mobility μ can be estimated; for this particular device it is about $0.6\,cm^2\,V^{-1}\,s^{-1}$ ($C_{diel} = 0.7\,\mu F\,cm^{-2}$, $W = 100\,\mu m$, $L = 30\,\mu m$).

Equations (6.1) and (6.2) describe the drain current for gate-source voltages above the threshold voltage. Below the threshold voltage there is a region in which the drain current depends exponentially on the gate-source voltage. This is the *subthreshold region*; for the TFT in Figure 6.17 it extends between about -0.5 V and about -1 V. Within this voltage range the drain current is due to carriers that have sufficient thermal energy to overcome the gate-controlled barrier near the source and mainly diffuse, rather than drift, to the drain:

$$I_D = I_0 \exp\left(\frac{q|V_{GS} - V_{th}|}{nkT}\right) \qquad \text{for } V_{GS} \text{ between } V_{th} \text{ and } V_{SO} \tag{6.3}$$

The slope of the $\log(I_D)$ versus V_{GS} curve in the subthreshold region is determined by the ideality factor n and the temperature T (q is the electronic charge, k is

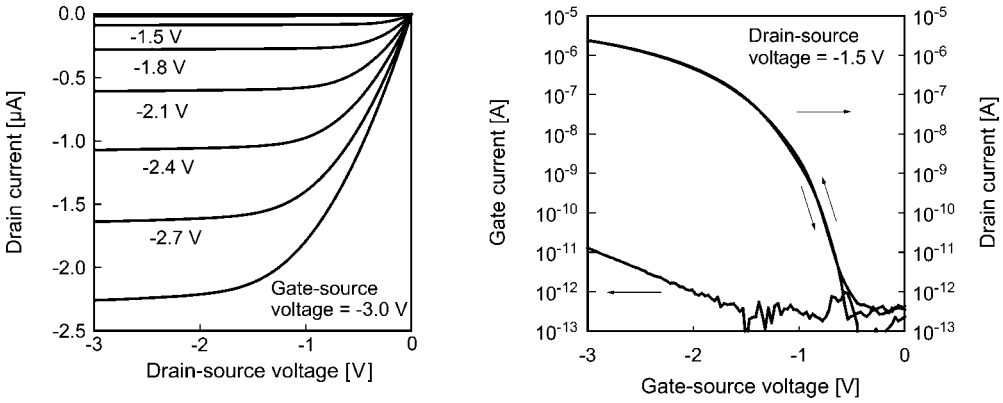


Figure 6.17 The electrical characteristics of a p-channel pentacene TFT.

Boltzmann's constant, and V_{SO} is the switch-on voltage which marks the gate-source voltage at which the drain current reaches a minimum [55]). It is usually quantified as the inverse subthreshold slope S (also called subthreshold swing):

$$S = \frac{\partial V_{GS}}{\partial(\log_{10} I_D)} = \frac{nkT}{q}\ln 10 \tag{6.4}$$

The ideality factor n is determined by the density of trap states at the semiconductor/dielectric interface, N_{it}, and the gate dielectric capacitance, C_{diel}:

$$n = 1 + \frac{qN_{it}}{C_{diel}} \tag{6.5}$$

$$S = \frac{kT}{q}\ln 10\left(1 + \frac{qN_{it}}{C_{diel}}\right) \tag{6.6}$$

When N_{it}/C_{diel} is small, the ideality factor n approaches unity. Silicon MOSFETs often come close to the ideal room-temperature subthreshold swing of 60 mV per decade, as the quality of the Si/SiO_2 interface is very high. In organic TFTs the semiconductor/dielectric interface is typically of lower quality, and thus the subthreshold swing is usually larger. The TFT in Figure 6.17 has a subthreshold swing of 100 mV dec^{-1}, from which an interface trap density of 3×10^{12} cm^{-2} V^{-1} is calculated.

The subthreshold region extends between the threshold voltage V_{th} and the switch-on voltage V_{SO}. Below the switch-on voltage (-0.5 V for the TFT in Figure 6.17) the drain current is limited by leakage through the semiconductor, through the gate dielectric, or across the substrate surface. This off-state current should be as small as possible. The TFT in Figure 6.17 has an off-state current of 0.5 pA, which corresponds to a on off-state resistance of 3 TΩ.

To predict an upper limit for the dynamic performance of the transistor, it is useful to calculate the cut-off frequency [56]. This is the frequency at which the current gain is unity, and is determined by the transconductance and the gate capacitance:

$$f_T = \frac{g_m}{2\pi C_{gate}} \tag{6.7}$$

The transconductance g_m is defined as the change in drain current with respect to the corresponding change in gate-source voltage:

$$g_m = \frac{\partial I_D}{\partial V_{GS}} \tag{6.8}$$

Thus, the transconductance can be extracted from the current–voltage characteristics; for the pentacene TFT in Figure 6.17 the transconductance is about 2 µS at $V_{GS} = -2.5$ V. The transconductance is related to the other transistor parameters as follows:

$$g_m = \frac{\mu C_{diel} W}{L} V_{DS} \quad \text{in the linear regime} \tag{6.9}$$

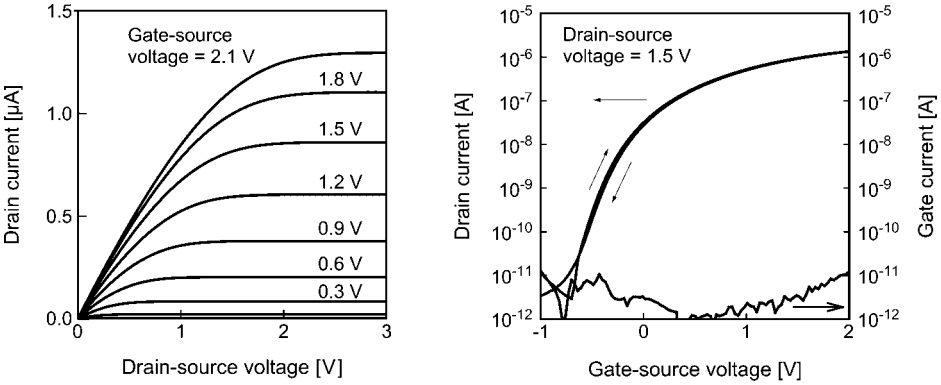


Figure 6.18 The electrical characteristics of an n-channel F_{16}CuPc TFT.

$$g_m = \frac{\mu C_{diel} W}{L}(V_{GS} - V_{th}) \quad \text{in the saturation regime} \tag{6.10}$$

The gate capacitance C_{gate} is the sum of the intrinsic gate capacitance (representing the interaction between the gate and the channel charge) and the parasitic gate capacitances (including the gate/source overlap capacitances and any fringing capacitances). For the transistor in Figure 6.17 the total gate capacitance is estimated to be about 30 pF, so the cut-off frequency will be on the order of 10 kHz.

The current–voltage characteristics of an n-channel TFT based on perfluorinated copper phthalocyanine (F_{16}CuPc) are shown in Figure 6.18 [54]. It has a threshold voltage of -0.2 V, a switch-on voltage of -0.8 V, an off-state current of about 5 pA, and a carrier mobility of about $0.02\,cm^2\,V^{-1}\,s^{-1}$ ($C_{diel} = 0.7\,\mu F\,cm^{-2}$, $W = 1000$ μm, $L = 30\,\mu m$). The transconductance is about 0.8 μS at $V_{GS} = 1.5$ V and the gate capacitance is about 300 pF, so the TFT will have a cut-off frequency of about 500 Hz.

The availability of both p-channel and n-channel organic TFTs makes the implementation of organic complementary circuits possible. From a circuit design perspective, complementary circuits are more desirable than circuits based only on one type of transistor, as complementary circuits have smaller static power dissipation and greater noise margin [54]. The schematic and electrical characteristics of an organic complementary inverter with a p-channel pentacene TFT and an n-channel F_{16}CuPc TFT are shown in Figure 6.19.

In a complementary inverter the p-channel FET is conducting only when the input is low ($V_{in} = 0$ V), while the n-channel FET is conducting only when the input is high ($V_{in} = V_{DD}$). Consequently, the static current in a complementary circuit is essentially determined by the leakage currents of the transistors and can be very small (less than 100 pA for the inverter in Figure 6.19). As a result, the output signals in the steady states are essentially equal to the rail voltages V_{DD} and ground. During switching there is a brief period when both transistors are simultaneously in the low-resistance on-state and a significant current flows between the V_{DD} and ground rails. Thus, most of the power consumption of a complementary circuit is due to switching, while the static power dissipation is very small.

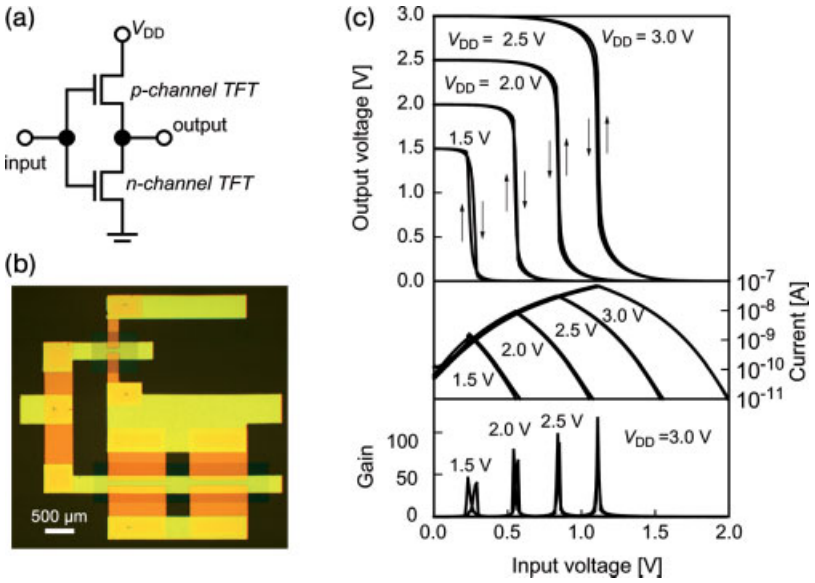


Figure 6.19 (a) Schematic, (b) actual photographic image and (c) transfer characteristics of an organic complementary inverter based on a p-channel pentacene TFT and an n-channel F_{16}CuPc TFT.

The dynamic performance of the inverter is limited by the slower of the two transistors, in this case the n-channel F_{16}CuPc TFT. Figure 6.20 shows, in graphical from, the inverter's response to a square-wave input signal with an amplitude of 2 V and a frequency of 500 Hz – that is, the cut-off frequency of the F_{16}CuPc TFT.

To allow organic circuits to operate at higher frequencies, it is necessary to increase the transconductance and reduce the parasitic capacitances. From a materials point of view, this can be done by developing new organic semiconductors that provide larger carrier mobilities [36]. Ideally, the carrier mobilities of the p- and n-channel TFTs should be similar. From a manufacturing point of view, the critical dimensions of the devices must be reduced – that is, the channel length and overlap capacitances must be made smaller. However, as the channel length of organic TFTs is reduced, the

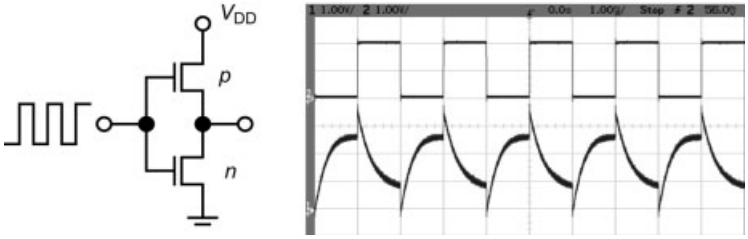


Figure 6.20 Response of an organic complementary inverter to a square-wave input signal with an amplitude of 2 V and a frequency of 500 Hz. Both TFTs have a channel length of 30 µm.

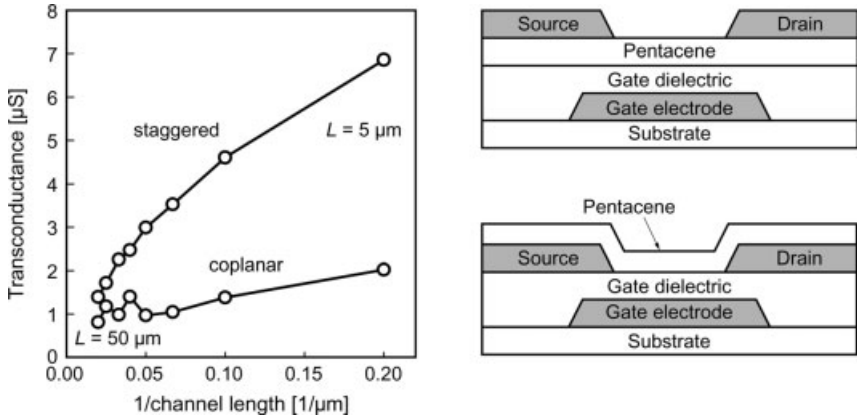


Figure 6.21 Transconductance as a function of channel length for two series of pentacene TFTs (top: inverted staggered configuration; bottom: inverted coplanar configuration). The channel width is 100 μm.

transconductance does not necessarily scale as predicted by Eqs. (6.9) and (6.10). The main reason for this is that the contact resistance in organic TFTs can be very large, as the contacts are typically not doped. Consequently, as the channel length is reduced, the drain current becomes increasingly limited by the contact resistance (which is independent of channel length), rather than by the channel resistance. This is shown in Figure 6.21 for pentacene TFTs in the inverted staggered configuration and in the inverted coplanar configuration, both with channel length ranging from 50 μm to 5 μm and all with a channel width of 100 μm.

For long channels ($L = 50\,\mu m$), where the effect of the contact resistance on the TFT characteristics is small, the transconductance is similar for both technologies ($g_m \sim 1\,\mu S$; $\mu \sim 0.5\,cm^2\,V^{-1}\,s^{-1}$). The difference between the two devices configurations becomes evident when the channel length is reduced. For the coplanar TFTs the potential benefit of channel length scaling is largely lost due to the significant contact resistance ($\sim 5 \times 10^4\,\Omega \cdot cm$). The staggered configuration offers significantly smaller contact resistance ($\sim 10^3\,\Omega \cdot cm$), as the area available for charge injection from the metal into the carrier channel is larger (given by the gate/contact overlap area), and as a result the transconductance for short channels (5 μm) is significantly larger in the case of the staggered TFTs (7 μS versus 2 μS). The staggered TFT with a channel length of 5 μm and a transconductance of 7 μS has a total gate capacitance of about 5 pF, so the cut-off frequency is estimated to be on the order of 200 kHz (at an operating voltage in the range of 2–3 V).

6.5 Applications

Unlike single-crystal silicon transistors, organic TFTs can be readily fabricated on glass or flexible plastic substrates, and this makes them useful for a variety of

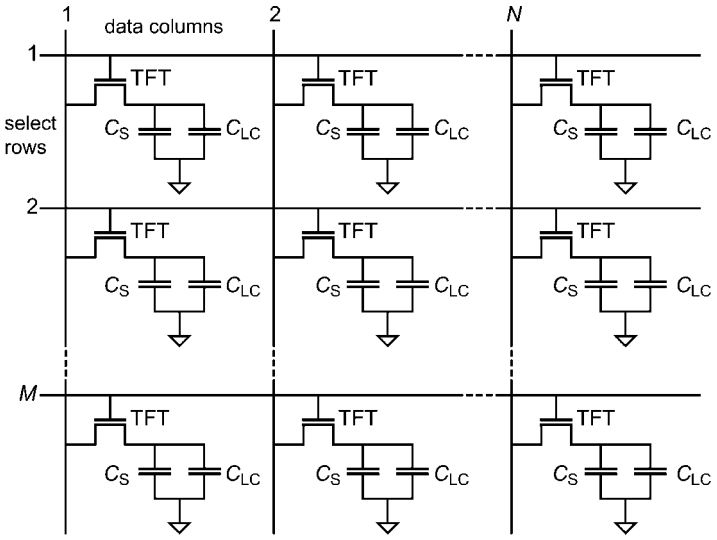


Figure 6.22 Schematic of an active-matrix liquid-crystal display.

large-area electronic applications, such as active-matrix flat-panel displays. In an active-matrix display, each of the pixels is individually controlled (and electrically isolated from the rest of the display) by a TFT circuit in order to reduce undesirable cross-talk and to increase fidelity and color depth. In active-matrix displays that utilize voltage-controlled display elements, such as a liquid crystal or an electrophoretic cell, each pixel circuit consists simply of a single TFT; in this case the display matrix has as many TFTs as it has pixels. If the display employs a current-controlled electro-optical device, such as a light-emitting diode, a more complex TFT circuit with two or more TFTs must be implemented in each pixel.

Figure 6.22 shows the circuit schematic of an active-matrix liquid-crystal display (AMLCD). The display is operated by applying a select voltage to one of the rows in order to switch all TFTs in that row to the low-resistance on-state. (All other rows are held at a lower potential that keeps the TFTs in these rows in the high-resistance off-state.) Data voltages that correspond to the desired brightness levels for each of the pixels in the selected row are then applied to each of the N columns. This charges the capacitors in the selected row to the applied data voltage. The time required to charge the capacitors (t_{select}) is determined by the on-state resistance of the TFTs, by the capacitances of the storage capacitor (C_S), the liquid crystal cell (C_{LC}) and the data lines, and by the maximum allowed deviation from the target voltage. Once the select voltage is removed, the capacitors are isolated and the charge is retained in the pixels. In this manner all rows are addressed one by one, and all pixel capacitors are charged to the desired voltage. The time required to sequentially address all M rows, and thus update the entire display, is the frame time, $t_{\text{frame}} = M \cdot t_{\text{select}}$.

In order to avoid visible flicker, the display information must be updated at least 50 times per second – that is, t_{frame} must be about 20 ms, or less. If a pixel capacitance ($C_S + C_{LC}$) of 1 pF is assumed, and if no more than 1% of the stored charge is allowed

to leak from the pixel during the t_{frame}, then the minimum required TFT off-state resistance can be estimated:

$$R_{\text{off}} \geq \frac{t_{\text{frame}}}{0.01 C_{\text{pixel}}} \tag{6.11}$$

Thus, for a t_{frame} of 20 ms and a pixel capacitance of 1 pF, the TFTs must have an off-state resistance of 2 TΩ, or greater.

For an extended graphics array (XGA) display with 768 rows and a t_{frame} of 20 ms the time available to charge the capacitors in one row is $t_{\text{select}} = t_{\text{frame}}/M = 26\,\mu\text{s}$. If a combined (pixel plus data line) capacitance of 2 pF is assumed, and it is specified that the capacitors be charged to within 1% of the target data voltage, then the maximum allowed TFT on-state resistance can be estimated:

$$R_{\text{on}} \leq \frac{t_{\text{select}}}{4.6 C_{\text{pixel}}} \tag{6.12}$$

For a t_{select} of 26 μs and a capacitance of 2 pF this sets an upper limit of about 2 MΩ for the on-state resistance of the TFTs. In order to create a small on-state resistance the TFTs are operated in the linear regime by applying a select voltage that is larger than the largest data voltage (plus the threshold voltage). In the linear regime (when the gate-source voltage is much larger than the drain-source voltage) the channel resistance is approximately given by:

$$R_{\text{on}} \sim \frac{L}{\mu C_{\text{diel}} W (V_{\text{GS}} - V_{\text{th}})} \tag{6.13}$$

Assuming a carrier mobility of $0.5\,\text{cm}^2\,\text{V}^{-1}\,\text{s}^{-1}$, a gate dielectric capacitance of $0.1\,\mu\text{F}\,\text{cm}^{-2}$, and an overdrive voltage $|V_{\text{GS}} - V_{\text{th}}|$ of 10 V, the on-state resistance requirement (2 MΩ) can be met with a TFT geometry of $W/L = 1$ (where W and L are the channel width and channel length of the transistor, respectively). A W/L ratio near or equal to unity is desirable, as this means that the transistor occupies a relatively small fraction of the total pixel area. Taking into account both the off-state and on-state resistance requirements ($R_{\text{off}} > 2$ TΩ, $R_{\text{on}} < 2$ MΩ), the TFTs must have an on/off ratio of at least 10^6.

These requirements can be met by state-of-the-art organic TFTs. An early demonstration of an active-matrix polymer-dispersed liquid-crystal (PDLC) display with solution-processed polythienylenevinylene (PTV) TFTs was developed by Philips Research in 2001 [57]. In 2005, Sony reported an active-matrix twisted-nematic liquid-crystal (TN-LC) display with vacuum-deposited pentacene TFTs [58]. Also in 2005, Polymer Vision demonstrated a flexible roll-up display based on electrophoretic microcapsules (electronic ink) and solution-processed pentacene TFTs (see Figure 6.23).

Unlike liquid-crystal valves and electrophoretic microcapsules, light-emitting diodes (OLEDs) are current-controlled display elements and thus require a more complex pixel circuit. The simplest implementation of an active-matrix organic light-emitting diode (AMOLED) pixel is shown in Figure 6.24. When a select voltage is

Figure 6.23 Left: A 64 × 64 pixel active-matrix polymer-dispersed liquid-crystal display with solution-processed polymer TFTs developed by Philips. (Reproduced with permission from Ref. [57].) Center: A 160 × 120 pixel active-matrix twisted-nematic liquid-crystal display with vacuum-deposited pentacene TFTs developed by Sony. (Reproduced with permission from Ref. [58].) Right: A 320 × 240 pixel active-matrix electronic-ink display with solution-processed pentacene TFTs developed by Polymer Vision. (Reproduced from: H. E. A. Huitema et al., Roll-up Active-matrix Displays, in: *Organic Electronics*, Wiley-VCH, 2006.)

applied, transistor T_1 switches to the low-resistance on-state so that capacitor C_S can be charged through the data line to a voltage corresponding to the desired luminous intensity. The voltage across C_S is the gate-source voltage of transistor T_2, and thus determines the drain current of T_2 and thereby the luminance of the OLED. When the select voltage is removed, T_1 switches off and the charge is retained on C_S, so T_2 remains active and drives a constant OLED current for the remainder of the frame time.

The on-state and off-state resistance requirements for the select transistor T_1 in an OLED pixel are similar to those for the TFT in a liquid-crystal or electrophoretic pixel – that is, they can be met by a TFT with $W/L \sim 1$. The drive transistor T_2 in an OLED pixel is usually operated in saturation, and must have a sufficiently large drain current to drive the OLED to the desired brightness. State-of-the-art small-molecule OLEDs have luminous efficiencies on the order of 2 to 50 cd A^{-1}, depending on emission color, material selection, and process technology [59]. For a typical display brightness of 100 cd m^{-2} and a pixel size of 6×10^{-4} cm^2 (which corresponds to a resolution of 100 dpi), this requires a maximum drive current up to about 3 μA. In the saturation regime the drain current of the transistor is given by Eq. (6.2). Assuming a carrier mobility of 0.5 $cm^2 V^{-1} s^{-1}$, a gate dielectric capacitance of 0.1 μF cm^{-2}, and an overdrive voltage $|V_{GS} - V_{th}|$ of 5 V, the drive current requirement (3 μA) can be met

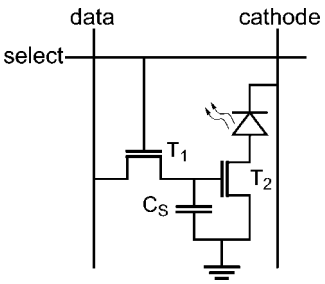


Figure 6.24 Schematic of a two-transistor active-matrix OLED pixel.

Figure 6.25 A flexible 48 × 48 pixel active-matrix organic light-emitting diode (OLED) display with pentacene organic TFTs developed at Penn State University. (Reproduced with permission from Ref. [60].)

with a TFT geometry of $W/L = 5$. Thus, the static transistor performance requirements for active-matrix OLED displays can be met by organic TFTs T_1 and T_2 occupying only a small fraction of the pixel area. A photograph of an active-matrix OLED display with two pentacene TFTs and a bottom-emitting OLED in each pixel [60] is shown in Figure 6.25.

Compared with liquid-crystal and electrophoretic displays, active-matrix OLED displays are far more demanding as far as the uniformity and stability of the TFT parameters are concerned. For example, if the TFT threshold voltage in a liquid-crystal display changes over time, or is not uniform across the display, the image quality is not immediately affected as the select voltage is usually large. In an OLED display, however, the threshold voltage of transistor T_2 directly determines the drive current and thus the OLED brightness. Consequently, even small differences in threshold voltage have a dramatic impact on image quality and color fidelity. In order to reduce or eliminate the effects of non-uniformities or time-dependent changes of the TFT parameters, more complex pixel circuit designs have been proposed [61]. In these designs, additional TFTs are implemented to make the OLED current independent of the threshold voltages of the TFTs. A pixel circuit with a larger number of TFTs is likely to occupy a greater portion of the total pixel area, but may significantly improve the performance of the display.

A second potential application for organic TFTs is in large-area sensors for the spatially resolved detection of physical or chemical quantities, such as temperature, pressure, radiation, or pH. As an example, Figure 6.26 shows the schematic of an active-matrix pressure sensor array. Mechanical pressure exerted on a sensor element leads to a reversible and reproducible change in the resistance of the sensor element. To allow external circuitry to access the resistance of each individual sensor it is necessary to integrate a transistor with each sensor element. During operation, the rows of the array are selected one by one to switch the TFTs in the selected row to the low-resistance state (similar to the row-select procedure in an active-matrix display) and the resistance of the each sensor element in the selected row is measured

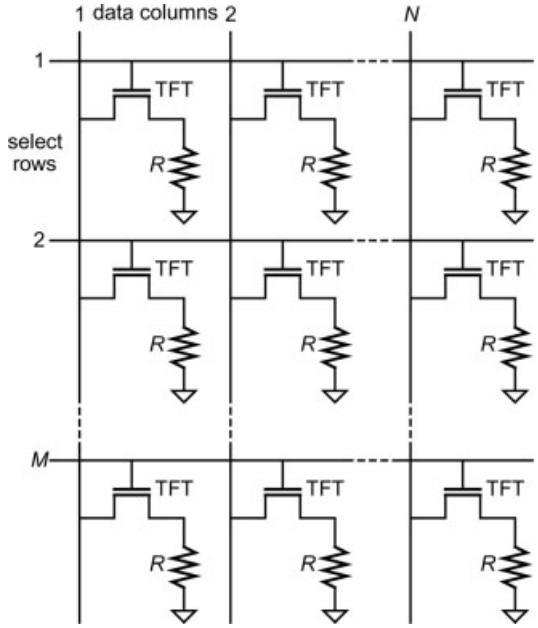


Figure 6.26 Left: Schematic of an active-matrix array with resistive sensor elements. Right: Demonstration of an artificial skin device with organic TFTs. (Reproduced from T. Someya et al., Large-area detectors and sensors, in: *Organic Electronics*, Wiley-VCH, 2006.)

through the data lines by external circuitry. This is repeated for each row until the entire array has been read out. The result is a map of the 2-D distribution of the desired physical quantity (in this case, the pressure) over the array. By reading the array continuously a dynamic image can be created (again, similar to an active-matrix display). One application of a 2-D pressure sensor array is a fingerprint sensor for personal identification purposes. Another interesting application is the combination of spatially resolved pressure and temperature sensing over large conformable surfaces to create the equivalent of sensitive skin for human-like robots capable of navigating in unstructured environments [62].

6.6 Outlook

Organic transistors are potentially useful for applications that require electronic functionality with low or medium complexity distributed over large areas on unconventional substrates, such as glass or flexible plastic film. Generally, these are applications in which the use of single-crystal silicon devices and circuits is either technically or economically not feasible. Examples include flexible displays and sensors. However, organic transistors are unlikely to replace silicon in applications

characterized by large transistor counts, small chip size, large integration densities, or high-frequency operation. The reason is that, in these applications, the use of silicon MOSFETs is very economical. For example, the manufacturing cost of a silicon MOSFET in a 1-Gbit memory chip is on the order of $\$10^{-9}$, which is less than the cost of printing a single letter in a newspaper.

The static and dynamic performance of state-of-the-art organic TFTs is already sufficient for certain applications, most notably small or medium-sized flexible displays in which the TFTs operate with critical frequencies in the range of a few tens of kilohertz. Strategies for increasing the performance of organic TFTs include further improvements in the carrier mobility of the organic semiconductor (either through the synthesis of new materials, through improved purification, or by enhancing the molecular order in the semiconductor layer) and more aggressive scaling of the lateral transistor dimensions (channel length and contact overlap). For example, an increase in cut-off frequency from 200 kHz to about 2 MHz can be achieved either by improving the mobility from $0.5\,cm^2\,V^{-1}\,s^{-1}$ to about $5\,cm^2\,V^{-1}\,s^{-1}$ (assuming critical dimensions of 5 μm and an operating voltage of 3 V), or by reducing the critical dimensions from 5 μm to about 1.6 μm (assuming a mobility of $0.5\,cm^2\,V^{-1}\,s^{-1}$ and an operating voltage of 3 V). A cut-off frequency of about 20 MHz is projected for TFTs with a mobility of $5\,cm^2\,V^{-1}\,s^{-1}$ and critical dimensions of 1.6 μm (again assuming an operating voltage of 3 V).

However, these improvements in performance must be implemented without sacrificing the general manufacturability of the devices, circuits, and systems. This important requirement has fueled the development of a whole range of large-area, high-resolution printing methods for organic electronics. Functional printed organic devices and circuits have indeed been demonstrated using various printing techniques, but further studies are required to address issues such as process yield and parameter uniformity.

One of the most critical problems that must be solved before organic electronics can begin to find use in commercial applications is the stability of the devices and circuits during continuous operation, and while exposed to ambient oxygen and humidity. Early product demonstrators have often suffered from short lifetimes due to a rapid degradation of the organic semiconductor layers. However, recent advances in synthesis, purification, processing, in addition to economically viable encapsulation techniques, have raised the hope that the degradation of organic semiconductors is not an insurmountable problem and that organic thin-film transistors may soon be commercially utilized.

References

1 W. Warta, N. Karl, Hot holes in naphthalene: High, electric-field dependent mobilities, *Phys. Rev. B* 1985, **32**, 1172.

2 M. C. J. M. Vissenberg, M. Matters, Theory of field-effect mobility in amorphous organic transistors, *Phys. Rev. B* 1998, **57**, 12964.

3 G. Horowitz, R. Hajlaoui, P. Delannoy, Temperature dependence of the field-effect mobility of sexithiophene. Determination of the density of

traps, *J. Phys. III France* 1995, **5**, 355.

4 A. Tsumura, H. Koezuka, T. Ando, Macromolecular electronic device: Field-effect transistor with a polythiophene thin film, *Appl. Phys. Lett.* 1986, **49**, 1210.

5 A. Assadi, C. Svensson, M. Willander, O. Inganas, Field-effect mobility of poly(3-hexylthiophene), *Appl. Phys. Lett.* 1988, **53**, 195.

6 R. D. McCullough, R. D. Lowe, M. Jayaraman, D. L. Anderson, Design, synthesis, and control of conducting polymer architectures: Structurally homogeneous poly(3-alkylthiophenes), *J. Org. Chem.* 1993, **58**, 904.

7 Z. Bao, A. Dodabalapur, A. Lovinger, Soluble and processable regioregular poly (3-hexylthiophene) for thin film field-effect transistor applications with high mobility, *Appl. Phys. Lett.* 1996, **69**, 4108.

8 H. Sirringhaus, P. J. Brown, R. H. Friend, M. M. Nielsen, K. Bechgaard, B. M. W. Langeveld-Voss, A. J. H. Spiering, R. A. J. Janssen, E. W. Meijer, P. Herwig, D. M. de Leeuw, Two-dimensional charge transport in self-organized, high-mobility conjugated polymers, *Nature* 1999, **401**, 685.

9 B. S. Ong, Y. Wu, P. Liu, S. Gardner, High-performance semiconducting polythiophenes for organic thin-film transistors, *J. Am. Chem. Soc.* 2004, **126**, 3378.

10 I. McCulloch, M. Heeney, C. Bailey, K. Genevicius, I. MacDonald, M. Shkunov, D. Sparrowe, S. Tierney, R. Wagner, W. Zhang, M. L. Chabinyc, R. J. Kline, M. D. McGehee, M. F. Toney, Liquid-crystalline semiconducting polymers with high charge-carrier mobility, *Nature Mater.* 2006, **5**, 328.

11 M. Madru, G. Guillaud, M. Al Sadoun, M. Maitrot, C. Clarisse, M. Le Contellec, J. J. Andre, J. Simon, The first field effect transistor based on an intrinsic molecular semiconductor, *Chem. Phys. Lett.* 1987, **142**, 103.

12 C. Clarisse, M. T. Riou, M. Gauneau, M. Le Contellec, Field-effect transistor with diphthalocyanine thin film, *Electronics Lett.* 1988, **24**, 674.

13 G. Horowitz, D. Fichou, X. Peng, Z. Xu, F. Garnier, A field-effect transistor based on conjugated alpha-sexithienyl, *Solid State Commun.* 1989, **72**, 381.

14 F. Garnier, G. Horowitz, X. Z. Peng, D. Fichou, An all-organic 'soft' thin film transistor with very high carrier mobility, *Adv. Mater.* 1990, **2**, 592.

15 Y. Y. Lin, D. J. Gundlach, S. F. Nelson, T. N. Jackson, Pentacene-based organic thin-film transistors, *IEEE Trans. Electron. Dev.* 1997, **44**, 1325.

16 Y. Y. Lin, D. J. Gundlach, S. F. Nelson, T. N. Jackson, Stacked pentacene layer organic thin film transistors with improved characteristics, *IEEE Electr. Dev. Lett.* 1997, **18**, 606.

17 W. Kalb, P. Lang, M. Mottaghi, H. Aubin, G. Horowitz, M. Wuttig, Structure-performance relationship in pentacene/Al_2O_3 thin-film transistors, *Synth. Metals* 2004, **146**, 279.

18 S. Y. Yang, K. Shin, C. E. Park, The effect of gate dielectric surface energy on pentacene morphology and organic field-effect transistor characteristics, *Adv. Funct. Mater.* 2005, **15**, 1806.

19 M. Halik, H. Klauk, U. Zschieschang, G. Schmid, S. Ponomarenko, S. Kirchmeyer, W. Weber, Relationship between molecular structure and electrical performance of oligothiophene organic thin film transistors, *Adv. Mater.* 2003, **15**, 917.

20 T. W. Kelley, L. D. Boardman, T. D. Dunbar, D. V. Muyres, M. J. Pellerite, T. P. Smith, High-performance OTFTs using surface-modified alumina dielectrics, *J. Phys. Chem. B* 2003, **107**, 5877.

21 S. Z. Weng, W. S. Hu, C. H. Kuo, Y. T. Tao, L. J. Fan, Y. W. Yang, Anisotropic field-effect mobility of pentacene thin-film transistor: Effect of rubbed self-assembled monolayer, *Appl. Phys. Lett.* 2006, **89**, 172103.

22 S. Lee, B. Koo, J. Shin, E. Lee, H. Park, H. Kim, Effects of hydroxyl groups in polymeric dielectrics on organic transistor performance all-organic active matrix flexible display, *Appl. Phys. Lett.* 2006, **88**, 162109.

23 P. Herwig, K. Müllen, A soluble pentacene precursor: Synthesis, solid-state conversion into pentacene and application in a field-effect transistor, *Adv. Mater.* 1999, **11**, 480.

24 A. Afzali, C. D. Dimitrakopoulos, T. L. Breen, High-performance, solution-processed organic thin film transistors from a novel pentacene precursor, *J. Am. Chem. Soc.* 2002, **124**, 8812.

25 M. M. Payne, S. R. Parkin, J. E. Anthony, C. C. Kuo, T. N. Jackson, Organic field-effect transistors from solution-deposited functionalized acenes with mobilities as high as 1 cm^2/Vs, *J. Am. Chem. Soc.* 2005, **127**, 4986.

26 C. C. Kuo, M. M. Payne, J. E. Anthony, T. N. Jackson, TES anthradithiophene solution-processed OTFTs with 1 cm^2/V-s mobility, 2004 International Electron Devices Meeting Technical Digest, 2004, p. 373.

27 S. K. Park, C. C. Kuo, J. E. Anthony, T. N. Jackson, High mobility solution-processed OTFTs, 2005 International Electron Devices Meeting Technical Digest, 2005, p. 113.

28 K. C. Dickey, J. E. Anthony, Y. L. Loo, Improving organic thin-film transistor performance through solvent-vapor annealing of solution-processable triethylsilylethynyl anthradithiophene, *Adv. Mater.* 2006, **18**, 1721.

29 L. L. Chua, J. Zaumseil, J. F. Chang, E. C. W. Ou, P. K. H. Ho, H. Sirringhaus, R. H. Friend, General observation of n-type field-effect behaviour in organic semiconductors, *Nature* 2005, **343**, 194.

30 R. Schmechel, M. Ahles, H. von Seggern, A pentacene ambipolar transistor: Experiment and theory, *J. Appl. Phys.* 2005, **98**, 084511.

31 J. Yamaguchi, S. Yaginuma, M. Haemori, K. Itaka, H. Koinuma, An in-situ fabrication and characterization system developed for high performance organic semiconductor devices, *Jpn. J. Appl. Phys.* 2005, **44**, 3757.

32 K. Itaka, M. Yamashiro, J. Yamaguchi, M. Haemori, S. Yaginuma, Y. Matsumoto, M. Kondo, H. Koinuma, High-mobility C_{60} field-effect transistors fabricated on molecular-wetting controlled substrates, *Adv. Mater.* 2006, **18**, 1713.

33 H. E. Katz, J. Johnson, A. J. Lovinger, W. Li, Naphthalenetetracarboxylic diimide-based n-channel transistor semiconductors: Structural variation and thiol-enhanced gold contacts, *J. Am. Chem. Soc.* 2000, **122**, 7787.

34 P. P. L. Malenfant, C. D. Dimitrakopoulos, J. D. Gelorme, L. L. Kosbar, T. O. Graham, A. Curioni, W. Andreoni, N-type organic thin-film transistor with high field-effect mobility based on a N,N′-dialkyl-3,4,9,10-perylene tetracarboxylic diimide derivative, *Appl. Phys. Lett.* 2002, **80**, 2517.

35 M. M. Ling, Z. Bao, Copper hexafluorophthalocyanine field-effect transistors with enhanced mobility by soft contact lamination, *Org. Electronics* 2006, **7**, 568.

36 B. A. Jones, M. J. Ahrens, M. H. Yoon, A. Facchetti, T. J. Marks, M. R. Wasielewski, High-mobility air-stable n-type semiconductors with processing versatility: Dicyanoperylene-3,4:9,10-bis (dicarboximides), *Angew. Chem. Int. Ed.* 2004, **43**, 6363.

37 P. K. Weimer, The TFT – A new thin film transistor, *Proc. IRE* 1962, **50**, 1462.

38 P. G. LeComber, W. E. Spear, A. Ghaith, Amorphous silicon field-effect device and possible application, *Electron. Lett.* 1979, **15**, 179.

39 F. Ebisawa, T. Kurokawa, S. Nara, Electrical properties of polyacetylene/polysiloxane interface, *J. Appl. Phys.* 1983, **54**, 3255.

40 D. J. Gundlach, T. N. Jackson, D. G. Schlom, S. F. Nelson, Solvent-induced phase transition in thermally evaporated

pentacene films, *Appl. Phys. Lett.* 1999, **74**, 3302.

41 D. J. Gundlach, L. Zhou, J. A. Nichols, T. N. Jackson, P. V. Necliudov, M. S. Shur, An experimental study of contact effects in organic thin film transistors, *J. Appl. Phys.* 2006, **100**, 024509.

42 A. C. Arias, S. E. Ready, R. Lujan, W. S. Wong, K. E. Paul, A. Salleo, M. L. Chabinyc, R. Apte, R. A. Street, Y. Wu, P. Liu, B. Ong, All jet-printed polymer thin-film transistor active-matrix backplanes, *Appl. Phys. Lett.* 2004, **85**, 3304.

43 D. V. Muyres, P. F. Baude, S. Theiss, M. Haase, T. W. Kelley, P. Fleming, Polymeric aperture masks for high performance organic integrated circuits, *J. Vac. Sci. Technol. A* 2004, **22**, 1892.

44 G. H. Gelinck, T. C. T. Geuns, D. M. de Leeuw, High-performance all-polymer integrated circuits, *Appl. Phys. Lett.* 2000, **77**, 1487.

45 C. W. Sele, T. von Werne, R. H. Friend, H. Sirringhaus, Lithography-free, self-aligned inkjet printing with sub-hundred-nanometer resolution, *Adv. Mater.* 2005, **17**, 997.

46 J. Veres, S. D. Ogier, S. W. Leeming, D. C. Cupertino, S. M. Khaffaf, Low-k insulators as the choice of dielectrics in organic field-effect transistors, *Adv. Funct. Mater.* 2003, **13**, 199.

47 A. F. Stassen, R. W. I. de Boer, N. N. Iosad, A. F. Morpurgo, Influence of the gate dielectric on the mobility of rubrene single-crystal field-effect transistors, *Appl. Phys. Lett.* 2004, **85**, 3899.

48 I. N. Hulea, S. Fratini, H. Xie, C. L. Mulder, N. N. Iossad, G. Rastelli, S. Ciuchi, A. F. Morpurgo, Tunable Fröhlich polarons in organic single-crystal transistors, *Nature Mater.* 2006, **5**, 982.

49 L. A. Majewski, R. Schroeder, M. Voigt, M. Grell, High performance organic transistors on cheap, commercial substrates, *J. Phys. D* 2004, **37**, 3367.

50 M. H. Yoon, H. Yan, A. Facchetti, T. J. Marks, Low-voltage organic field-effect transistors and inverters enabled by ultrathin cross-linked polymers as gate dielectrics, *J. Am. Chem. Soc.* 2005, **127**, 10388.

51 S. Y. Yang, S. H. Kim, K. Shin, H. Jeon, C. E. Park, Low-voltage pentacene field-effect transistors with ultrathin polymer gate dielectrics, *Appl. Phys. Lett.* 2006, **88**, 173507.

52 M. Halik, H. Klauk, U. Zschieschang, G. Schmid, C. Dehm, M. Schütz, S. Maisch, F. Effenberger, M. Brunnbauer, F. Stellacci, Low-voltage organic transistors with an amorphous molecular gate dielectric, *Nature* 2004, **431**, 963.

53 M. H. Yoon, A. Facchetti, T. J. Marks, σ-π molecular dielectric multilayers for low-voltage organic thin-film transistors, *Proc. Natl. Acad. Sci. USA* 2005, **102**, 4678.

54 H. Klauk, U. Zschieschang, J. Pflaum, M. Halik, Ultralow-power organic complementary circuits, *Nature* 2007, **445**, 745.

55 E. J. Meijer, C. Tanase, P. W. M. Blom, E. van Veenendaal, B. H. Huisman, D. M. de Leeuw, T. M. Klapwijk, Switch-on voltage in disordered organic field-effect transistors, *Appl. Phys. Lett.* 2002, **80**, 3838.

56 H. Klauk, U. Zschieschang, M. Halik, Low-voltage organic thin-film transistors with large transconductance, *J. Appl. Phys.* 2007, **102**, 074514.

57 H. E. A. Huitema, G. H. Gelinck, J. B. P. H. van der Putten, K. E. Kuijk, K. M. Hart, E. Cantatore, D. M. de Leeuw, Active-matrix displays driven by solution processed polymeric transistors, *Adv. Mater.* 2002, **14**, 1201.

58 K. Nomoto, N. Hirai, N. Yoneya, N. Kawashima, M. Noda, M. Wada, J. Kasahara, A high-performance short-channel bottom-contact OTFT and its application to AM-TN-LCD, *IEEE Trans. Electr. Dev.* 2005, **52**, 1519.

59 P. Wellmann, M. Hofmann, O. Zeika, A. Werner, J. Birnstock, R. Meerheim, G. He, K. Walzer, M. Pfeiffer, K. Leo, High-efficiency p-i-n organic light-emitting diodes with long lifetime, *J. Soc. Information Display* 2005, **13**, 393.

60 L. Zhou, A. Wanga, S. C. Wu, J. Sun, S. Park, T. N. Jackson, All-organic active matrix flexible display, *Appl. Phys. Lett.* 2006, **88**, 083502.

61 A. Kumar, A. Nathan, G. E. Jabbour, Does TFT mobility impact pixel size in AMOLED backplanes? *IEEE Trans. Electr. Dev.* 2005, **52**, 2386.

62 T. Someya, Y. Kato, T. Sekitani, S. Iba, Y. Noguchi, Y. Murase, H. Kawaguchi, T. Sakurai, Conformable, flexible, large-area networks of pressure and thermal sensors with organic transistor active matrixes, *Proc. Natl. Acad. Sci. USA* 2005, **102**, 12321.

7
Carbon Nanotubes in Electronics

M. Meyyappan

7.1
Introduction

Since the discovery of carbon nanotubes (CNTs) in 1991 [1] by Sumio Iijima of the NEC Corporation, research activities exploring their structure, properties and applications have exploded across the world. This interesting nanostructure exhibits unique electronic properties and extraordinary mechanical properties, and this has prompted the research community to investigate the potential of CNTs in numerous areas including, among others, nanoelectronics, sensors, actuators, field emission devices, and high-strength composites [2]. Although recent progress in all of these areas has been significant, the routine commercial production of CNT-based products is still years away. This chapter focuses on one specific application field of CNTs, namely electronics, and describes the current status of developments in this area. This description is complemented with a brief discussion of the properties and growth methods of CNTs, further details of which are available in Ref. [2].

7.2
Structure and Properties

A carbon nanotube is, configurationally, a graphene sheet rolled up into a tube (see Figure 7.1). If it is a single layer of a graphene sheet, the resultant structure is a single-walled carbon nanotube (SWNT), but with a stack of multiple layers a multi-walled carbon nanotube (MWNT) emerges. The SWNT is a tubular shell made from hexagonal rings of carbon atoms, with the ends of the shells capped by a dome-like, half-fullerene molecules [3]. They are classified using a nomenclature (n, m) where n and m are integer indices of two graphene unit lattice vectors ($\mathbf{a_1}, \mathbf{a_2}$) corresponding to the chiral vector of a nanotube, $c_a = n\mathbf{a_1} + m\mathbf{a_2}$. Based on the geometry, when $n = m$, the resulting structure is commonly known as an "arm chair" nanotube, as shown in Figure 7.1. The (n, 0) structure is called the "zig zag nanotube", while all other

Nanotechnology. Volume 4: Information Technology II. Edited by Rainer Waser

ISBN: 978-3-527-31737-0

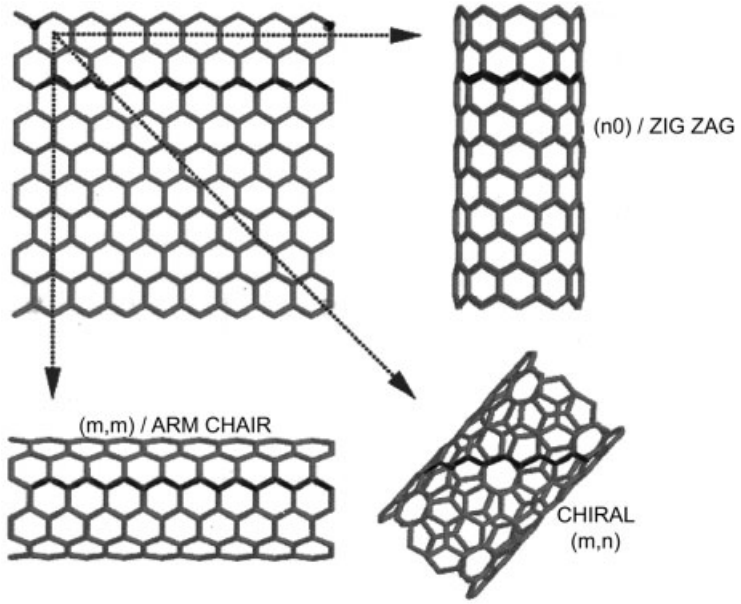


Figure 7.1 A strip of graphene sheet rolled into a carbon nanotube; m and n are chiral vectors.

structures are simply known as "chiral nanotubes". It is important to note that, at the time of this writing, exquisite control over the values of m and n is not possible. Transmission electron microscopy (TEM) images of a SWNT and a MWNT are shown in Figure 7.2, where the individual SWNTs are seen to have a diameter of about 1 nm. The MWNT has a central core with several walls, and a spacing close to 0.34 nm between the two neighboring walls (Figure 7.2b).

A SWNT can be either metallic or semiconducting, depending on its chirality – that is, the values of n and m. When $(n - m)/3$ is an integer, the nanotube is metallic, otherwise it is semiconducting. The diameter of the nanotube is given by $d = (a_g/\pi)(n^2 + mn + m^2)^{0.5}$, where a_g is the lattice constant of graphite. The strain energy

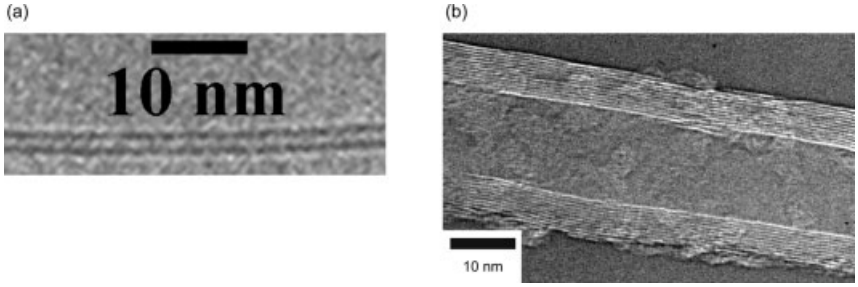


Figure 7.2 Transmission electron microscopy (TEM) images of (a) single-walled carbon nanotube and (b) a multi-walled carbon nanotube. (Image courtesy of Lance Delzeit.)

caused in the SWNT formation from the graphene sheet is inversely proportional to its diameter. There is a minimum diameter that can afford this strain energy, which is about 0.4 nm. On the other hand, the maximum diameter is about 3 nm, beyond which the SWNT may not retain its tubular structure and ultimately will collapse [3].

In the case of MWNTs, the smallest inner diameter found experimentally is about 0.4 nm, but typically is around 2 nm. The outer diameter of MWNT can be as large as 100 nm. Both, SWNTs and MWNTs, while preferentially being defect-free, have been observed experimentally in various defective forms such as bent, branched, helical, and even toroidal nanotubes.

The bandgap of a semiconducting nanotube is given by $E_g = 2d_{cc}\gamma/d$, where d_{cc} is the carbon–carbon bond length (0.142 nm), and γ is the nearest neighbor-hopping parameter (2.5 eV). Thus, the bandgap of semiconducting nanotubes of diameters between 0.5 and 1.5 nm may be in the range of 1.5 to 0.5 eV. The resistance of a metallic SWNT is $h/(4e^2) \approx 6.5\ \mathrm{K\Omega}$, where h is Planck's constant. However, experimental measurements typically show higher resistance due to the presence of defects, impurities, structural distortions and the effects of coupling to the substrate and/or contacts.

In addition to their interesting electronic properties, SWNTs exhibit extraordinary mechanical properties. For example, the Young's modulus of a (10,10) SWNT is over 1 TPa, with a tensile strength of 75 GPa. The corresponding values for graphite (in-plane) are 350 and 2.5 GPa, whereas the values for steel are 208 and 0.4 GPa [3]. Nanotubes can also sustain a tensile strain of 10% before fracturing, which is remarkably higher than other materials. The thermal conductivity of the nanotubes is substantially high [3, 4], with measured values being in the range of 1800 to 6000 $\mathrm{W\,mK^{-1}}$ [5].

7.3 Growth

The oldest process for preparing SWNTs and MWNTs is that of arc synthesis [6], with laser ablation subsequently being introduced during the 1990s to produce CNTs [7]. These bulk production techniques and large quantities are necessary when using CNTs in composites, gas storage. and similar applications. For electronics applications, it may be difficult to adopt "pick and place" strategies using bulk-produced material. Assuming that the need for *in-situ* growth approaches that currently used to produce devices for silicon-based electronics, it is important at this point to describe the techniques of chemical vapor deposition (CVD) and plasma-enhanced chemical vapor deposition (PECVD), both of which allow CNT growth on patterned substrates [8].

Chemical vapor deposition is a frequently used technique in silicon integrated circuit manufacture when depositing thin films of metals, semiconductors and dielectric materials. The CVD of CNTs typically involves a carbon-bearing feedstock such as CO or hydrocarbons including methane, ethylene, and acetylene. It is important to

maintain the growth temperature below that of the pyrolysis temperature of a particular hydrocarbon in order to avoid the production of amorphous carbon. The CNT growth is facilitated by the use of a transition metal catalyst, the choice comprising iron, nickel, palladium, or cobalt. These metals can be thermally evaporated as a thin film on the substrate, or sputtered using ion beam sputtering or magnetron sputtering. Alternatively, the catalyst metal can be applied to the substrate, starting from the metal-containing salt solution and passing through a number of steps such as precipitation, mixing, evaporation, drying, and annealing. It should be noted that solution-based techniques are more cumbersome and much slower than physical techniques such as sputtering. In addition, they may not be amenable for working with patterned substrates. Regardless of the approach used, the key here is to deposit the catalyst in the form of particles in order to facilitate nanotube growth. The characterization of as-deposited catalysts using TEM and atomic force microscopy [9] reveals that the particles are in the range of 1 to 10 nm in size. The catalyst deposition may be restricted to selected locations of the wafer through lithographic patterning. The type of lithography (optical, electron-beam, etc.) needed would be dictated primarily by the feature size of the patterns.

Typically, CNT-CVD is performed at atmospheric pressure and temperatures of 550 to 1000 °C. Low-pressure processes at several torr have also been reported. SWNTs require higher growth temperatures (above 800 °C) than MWNTs, whereas the latter can be grown at temperatures as low as 550 °C. Lower temperatures (<500 °C) may not be possible if the catalytic activation and realistic reaction/growth rates occur only at such elevated temperatures. At present, this restriction poses a serious problem for the adoption of CVD as an *in-situ* process in the device fabrication sequence, as common masking materials cannot withstand such high temperatures. Figure 7.3 shows bundles of SWNTs grown using methane with an iron catalyst prepared by ion beam sputtering. Typically, the SWNTs tend to bunch together to form bundles or ropes. Figure 7.4 shows a patterned MWNT growth on a silicon substrate using an iron catalyst, which appears to yield a vertical array of nanotubes. Although the ensemble appears vertical, a closer inspection would reveal – in all cases of thermal CVD – that the individual MWNT itself is actually not well aligned but is wavy.

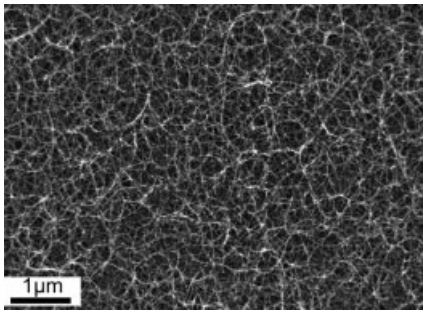


Figure 7.3 Bundles of SWNTs grown by chemical vapor deposition. (Image courtesy of Lance Delzeit.)

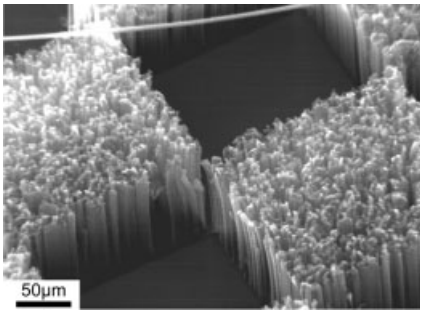


Figure 7.4 Patterned growth of multiwalled carbon nanotubes by CVD. (Image courtesy of H. T. Ng.)

In silicon integrated circuit manufacture, PECVD has emerged as a lower-temperature alternative to thermal CVD for the deposition of thin films of silicon, or its nitride or oxide. This strategy is not entirely successful in CNT growth, primarily because the growth temperature is tied to catalyst effectiveness as opposed to precursor dissociation [10]. Nevertheless, some reports have been made concerning nanotube growth at low temperature, or even at room temperature. However, these results are not reliable as they do not explicitly measure the growth temperature (i.e. the wafer temperature), but instead report only the temperatures on the bottom side of the substrate holder. Neither did any of these studies appreciate the fact that the plasma – and particularly the dc plasma used in most studies – heats the wafer substantially, particularly at the very high bias voltages commonly used. In such a case, even external heating via a heater may not be needed, and in most cases the temperature difference between the wafer and the bottom of the substrate holder may be several hundred degrees or more, depending on the input power [11]. Even if any degree of growth temperature reduction is achieved using PECVD, the material quality is relatively poor. Most of these structures are often conical in terms of configuration, with a continuously tapering diameter from the bottom to the top. Regardless of such issues, PECVD has one clear advantage over CVD, in that it enables the production of individual, freestanding, vertically aligned MWNT structures as opposed to individual, wavy nanotubes. These freestanding structures are invariably disordered with a bamboo-like inner core and, for that reason, are referred to as multi-walled carbon nanofibers (MWNFs) or simply carbon nanofibers (CNFs) [10]. PECVD is also capable of producing wavy MWNTs which are very similar to the thermally grown MWNTs.

To date, a variety of plasma sources have been used in CNT growth, including dc [12, 13], microwave [14], and inductive power sources [15]. The plasma efficiently breaks down the hydrocarbon feedstock, thus creating a variety of reactive radicals which are also the source for amorphous carbon. For this reason the feedstock is typically diluted with hydrogen, ammonia, argon or nitrogen to maintain the hydrocarbon fraction at less than about 20%. PECVD is performed at low pressures, typically in the range of 1 to 20 Torr. A scanning electron microscopy (SEM) image of PECVD-grown MWNFs is shown in Figure 7.5, wherein the individual structures are

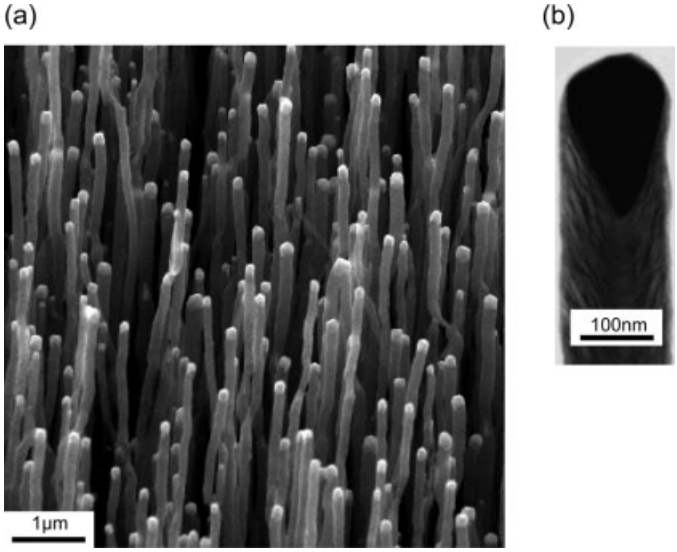


Figure 7.5 (a) SEM image showing vertical, freestanding carbon nanofibers grown by plasma-enhanced CVD. (Image courtesy of Alan Cassell.) (b) TEM image showing bamboo-like morphology and the catalyst particle at the head. (Image courtesy of Quoc Ngo and Alan Cassell.)

well separated and vertical. However, the TEM image reveals a disordered inner core and also the catalyst particle at the top. In contrast, in most cases of MWNT growth by thermal and plasma CVD, the catalyst particle is typically at the base of the nanotubes.

7.4 Nanoelectronics

Silicon complementary metal oxide semiconductor (CMOS) -based electronics has been moving forward impressively according to Moore's law, with 90-nm feature scale devices currently in production and 65-nm devices in the development stage. Research investigations are also well under way on the lower nodes and, as further miniaturization continues, a range of technological difficulties is anticipated, according to the Semiconductor Industry Association Roadmap [16]. These issues include lithography, novel dielectric materials, heat dissipation and efficient chip cooling to name a few. It was thought a few years ago that Si CMOS scaling may end at around 50 nm, beyond which alternatives such as CNT electronics or molecular electronics may be needed. However, this is no longer true as the current evidence suggests that scaling beyond 50 nm is possible, though with increased challenges. Regardless of

when the need for transition to alternatives emerges, there are a few expectations from a viable alternative:

- The new technology must be easier and cheaper to manufacture than Si CMOS.
- A high current drive is needed with the ability to drive capacitances of interconnects of any length.
- A reliability factor enjoyed to date must be available (i.e. operating time > 10 years).
- A high level of integration must be possible (>10^{10} transistors per circuit).
- A very high reproducibility is expected.
- The technology should not be handicapped with high heat dissipation problems currently forecast for the future-generation silicon devices, or attractive solutions must be available to tackle the anticipated heat loads.

Of course, the present status of CNT electronics has not yet reached the point where its performance in terms of the above goals can be evaluated. This is due to the fact that most efforts to date relate to the fabrication of single devices such as diodes and transistors, and little has been targeted at circuits and integration (as will be seen in the next section). In summary, the present status of CNT electronics evolution is similar to that of silicon technology between the invention of the transistor (during the late 1940s) and the development of integrated circuit in the 1960s. It would take at least a decade or more to demonstrate the technological progress required to meet the above-listed expectations.

7.4.1 Field Effect Transistors

The early attempts to investigate CNTs in electronics consisted of fabricating field effect transistors (FETs) with a SWNT as the conducting channel [17]. Tans *et al.* [18] reported first a CNT-FET where a semiconducting SWNT from a bulk-grown sample was transplanted to bridge the source and drain contacts separated by about a micron or more (see Figure 7.6). The contact electrodes were defined on a thick 300-nm SiO_2 layer grown on a silicon wafer acting as the back gate. The 1.4-nm tube with a corresponding bandgap of about 0.6 eV showed *I–V* characteristics indicating gate control of current flow through the nanotube. In the FET, the holes were the majority carriers and the conductance was shown to vary by at least six orders of magnitude. The device gain of this early device was below 1, due primarily to the thick gate oxide and high contact resistance.

At almost the same time, Martel *et al.* [19] presented their CNT-FET results using a similar back-gated structure. The oxide thickness was 140 nm, and 30 nm-thick

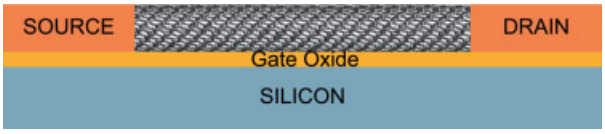


Figure 7.6 Schematic of an early CNT-FET with a back gate. A semiconducting nanotube bridges the source and drain, thus creating the conducting channel.

gold electrodes were defined using electron-beam lithography. The room-temperature I–V_{SD} characteristics showed that the drain current decreased strongly with increasing gate voltage, thus demonstrating CNT-FET operation through hole transport. The conductance modulation in this case spanned five orders of magnitude. The transconductance of this device was 1.7 nS at $V_{SD} = 10$ mV, with a corresponding hole mobility of 20 $cm^2 V^{-1} s^{-1}$. The authors concluded that the transport was diffusive rather than ballistic and, in addition, the high hole concentration was inherent to the nanotubes as a result of processing. This unipolar p-type device behavior suggested a Schottky barrier at the tube–contact metal interface. Later, the same IBM group [20] showed that n-type transistors could be produced simply by annealing the above p-type device in a vacuum, or by intentionally doping the nanotube with an alkali metal such as potassium.

All of the early CNT-FETs used the silicon substrate as the back gate. This unorthodox approach has several disadvantages. First, the resulting thick gate oxide required high gate voltages to turn the device on. Second, the use of the substrate for gating led to influencing all devices simultaneously. For integrated circuit applications, each CNT-FET needs its own gate control. Wind *et al.* [21] reported the first top gate CNT-FET which also featured embedding the SWNT within the insulator rather than exposing it to ambient, as had been done in the early devices. It was considered that such ambient exposure would lead to p-type characteristics and, as expected, the top gate device showed significantly better performance. A p-type CNT-FET with a gate length of 300 nm and gate oxide thickness of 15 nm showed a threshold voltage of –0.5 V and a transconductance of 2321 $\mu S\,\mu m^{-1}$. These results were better than those of a silicon p-MOSFET [22] with a much smaller gate length of 15 nm and an oxide thickness of 1.4 nm performing at a transconductance of 975 $\mu S\,\mu m^{-1}$. The CNT-FET also showed a three- to four-fold higher current drive per unit width compared to the above silicon device. Nihey *et al.* [23] also reported a top-gated device albeit with a thinner (6 nm) gate oxide TiO_2 with a higher dielectric constant. This device showed a 320 nS transconductance at a 100 mV drain voltage.

Most recently, Seidel *et al.* [24] reported CNT-FETs with short channels (18 nm), in contrast to all previous studies with micron-long channels. This group used nanotubes with diameters of 0.7 to 1.1 nm, and bandgaps in the range of 0.8 to 1.3 eV. The impressive performance of these devices included an on/off current ratio of 10^6 and a transconductance of 12 000 $\mu S\,\mu m^{-1}$. The current-carrying capacity was also very high, with a maximum current of 15 μA, corresponding to 10^9 $A\,cm^{-2}$. Another recent innovation involved a nanotube-on-insulator (NOI) approach [25], similar to the adoption of silicon-on-insulator (SOI) by the semiconductor industry, which minimizes parasitic capacitance.

As noted above, many of the CNT-FET studies conducted to date have used SWNTs, this being due to their superior properties compared to other types of nanotubes such as MWNTs and CNFs. As the bandgap is inversely proportional to the diameter, large-diameter MWNTs are invariably metallic. Martel *et al.* [19] fabricated the first MWNT FETs showing a significant gate effect. The real advantage of MWNTs is that they can be grown vertically up to reasonable lengths for a given diameter. Choi *et al.* [26, 27] took advantage of this point to fabricate vertical transistors using MWNTs grown using an anodic alumina template which essentially contained nanopores of various diameter

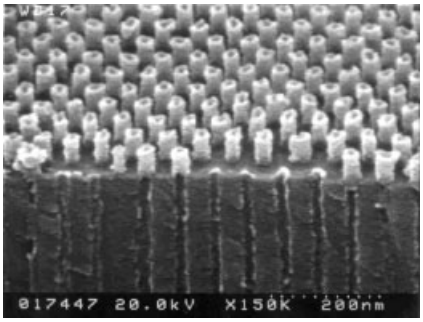
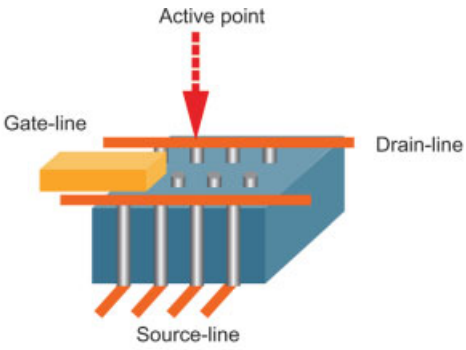
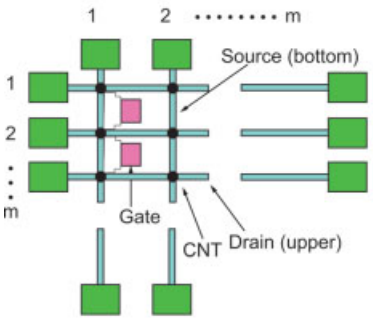


Figure 7.7 Vertically aligned MWNT grown using a nanoporous template. (Image courtesy of W. B. Choi.)

such that they were able to control both the diameter and pore density. The vertically aligned MWNTs grown using nanopores are shown in Figure 7.7. Following this, the device fabrication consisted of depositing SiO_2 on top of the aligned nanotubes. The electrode was then attached to the nanotubes through electron-beam-patterned holes, and finally the top metal electrode was attached. A schematic of the CNT-FET array and a SEM image of a 10×10 array is shown in Figure 7.8. For these devices, the authors claimed a tera-level transistor density of $2 \times 10^{11}\,\mathrm{cm}^{-2}$.





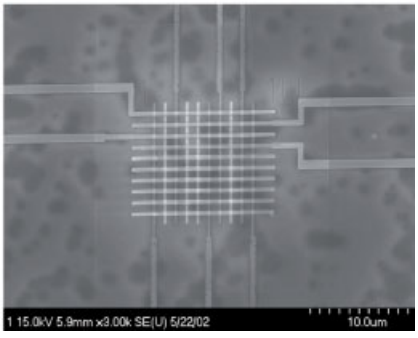


Figure 7.8 An array of CNT-FETs fabricated using the MWNTs in Figure 7.7. A schematic and an SEM image of an $n \times m$ array are shown. (Image courtesy of W. B. Choi.)

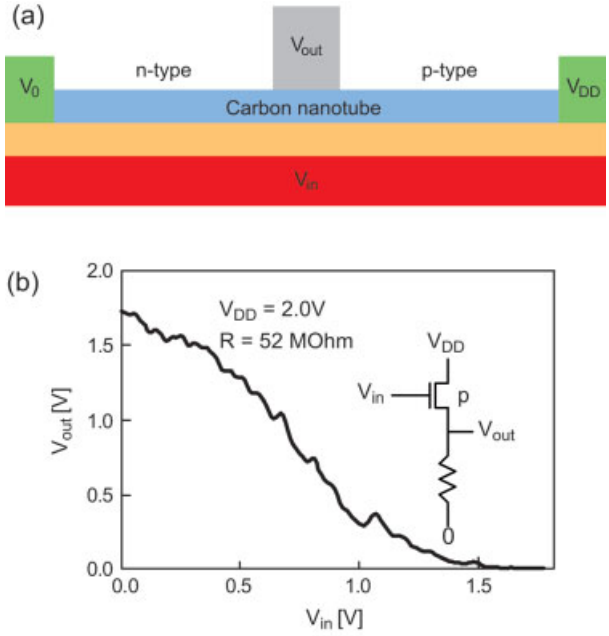


Figure 7.9 (a) n-type and p-type CNT-FETs. (b) Characteristics of an inverter circuit using the devices in (a). The inset shows the inverter circuit. (Reproduced from Ref. [28].)

Beyond single CNT-FETs, early attempts to fabricate circuit components have also been reported [28–30]. Liu *et al.* [28] fabricated both PMOS and CMOS inverters based on CNT-FETs. Their CMOS inverter connected two CNT-FETs in series using an electric lead of 2 mm length. One of the FETs was a p-type device, while the other – an n-type device – was obtained using potassium doping (see Figure 7.9). The devices used the silicon substrate as a back gate. The inverter was constructed and biased in the configuration shown in Figure 7.9b, after which a drain bias of 2.9 V was applied and the gate electrode was swept from 0 to 2.5 V, defining logics 0 and 1, respectively. As seen in the transfer curve in Figure 7.9b, when the input voltage is low (logic 0) the p-type and n-type devices were on and off, respectively (corresponding to their respective high- and low-conductance states). Then, the output is close to V_{DD}, producing a logic output of 1. When the input voltage is high (logic 1), the reverse is the case: the p-type transistor is off and the n-type device is on, with a combined output close to 0, producing a logic 0. The transfer curve in Figure 7.9b should not have a slope if this inverter were functioning ideally (which would correspond to a stepwise V_{out} versus V_{in} behavior). However, this first demonstration had a leaky p-device and so control of the threshold voltage of both devices was not perfect, thus leading to the slope seen in Figure 7.9b. Derycke *et al.* [29] also demonstrated an inverter using p- and n-type CNT-FETs. Beyond inverters, Bachtold *et al.* [30] fabricated circuits to perform logic operations such as NOR, and also constructed an ac ring oscillator.

As efforts continue in this direction, it is also important to consider issues such as $1/f$ noise, shot noise, and other similar concerns arising during operation. An analysis by Lin *et al.* [31] showed that the $1/f$ noise level in semiconducting SWNTs is correlated to the total number of charge carriers in the system. However, the noise level per carrier itself is not larger than that seen in silicon devices. Beyond the conventional binary logic approach, Raychowdry and Kaushik [32] discussed extensively implementation schemes for voltage-mode multiple-value logic (MVL) design. The MVL circuits reduce the number of operations per function and reduce the parasitics associated with routing and the overall power dissipation.

To date, the CNT-FET fabrication has essentially followed the silicon CMOS scheme by simply replacing the silicon conducting channel with a SWNT. This requires the presence of straight, aligned nanotubes controllably bridging a pair of electrodes laid out horizontally. As-grown SWNTs using any of the growth techniques, in contrast, exhibit a spaghetti-like morphology and occasionally consist of Y- and other types of junction. Menon and Srivastava [33, 34] postulated that such Y- and T-junctions are structurally stable and form the basis for three-terminal devices. However, such junctions in as-grown samples are, of course, neither controllable nor really amenable for further device processing. Satishkumer *et al.* [35] then set out to create these Y-junctions in a controllable manner using the anodic alumina template approach. Keeping two of the terminals at the same voltage, the two-terminal operation showed rectifying behavior when the voltage on the third terminal was varied. While their devices were constructed from MWNTs, SWNT Y-junctions have also been reported [36, 37], though FET operation using Y-junctions has not yet been demonstrated. Srivastava *et al.* [38] also proposed a radical neural tree architecture consisting of numerous Y-junctions (see Figure 7.10), wherein the concept is that the switching and processing of signals by these junctions in the tree would be similar to that of dentritic neurons in biological systems. In addition, acoustic, chemical or other signals may also be used instead of electrical signals.

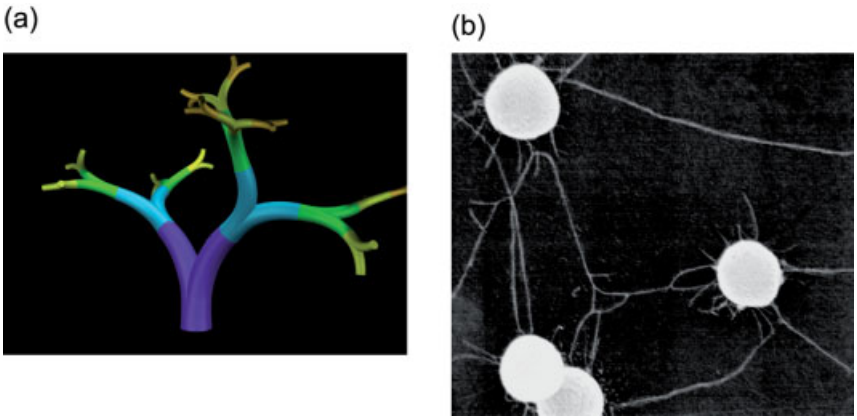


Figure 7.10 (a) A neural tree constructed using numerous Y-junctions of carbon nanotubes. (b) A network of interconnected SWNTs showing a few Y-junctions [36].

Figure 7.10b shows a very rudimentary attempt [36] to create such a CNT tree by utilizing self-assembled porous, collapsible polystrene/divinyl benzene microspheres to hold the catalyst. A controlled collapse of the spheres leads to the creation and release of the catalyst on the substrate for the CVD of SWNTs. Few Y-junctions showing an interconnected three-dimensional network of nanotubes are visible in Figure 7.10b.

7.4.2 Device Physics

The physics describing the operation of the CNT-FETs has been described in several theory papers [39–42] and summarized and reviewed elsewhere [17]. Here, the available information would be used to predict upper limits on CNT-FET performance. Yamada argues [17] that in nanoFETs the properties of the bulk material do not influence the device performance. In micro and macro FETs, the drain current is proportional to carrier mobility, which varies from material to material through its dependence on material-related properties such as effective mass and phonon scattering. In nanoFETs with ideal contacts, the drain current is determined by the transmission coefficient of an electron flux from the source to drain. When the carrier transport is ballistic, this coefficient is 1 and the material properties do not enter the picture directly. However, the material properties in practice enter indirectly as practical contacts (and hence the transmission coefficient at the contact) depend on the channel material and the interaction between the metal–channel semiconductor interface. The same applies to the preparation of the insulation material that determines the gate voltage characteristics. One-dimensional nanomaterials such as SWNTs inherently can suppress the short-channel effects arising from a deeper, broader distribution of carriers away from the gate, which occurs in reduced-size, two-dimensional silicon devices.

By using such an ideal device under ballistic transport and ideal contacts, Guo *et al.* [43] evaluated the performance of CNT-FETs. These authors considered a 1-nm SWNT with an insulator thickness of 1 nm and dielectric constant of 4. The geometry also was idealized to be a coaxial structure with contacts at either end of the nanotubes, and the gate wrapped around the nanotube. The computed on-off current ratio of 1120 was far higher than planar silicon CMOS devices with the same insulator parameters and power supply. The transconductance of this structure was also very high at 63 μS at 0.4 V, which was two orders of magnitude higher than any CNT-FET device discussed in Section 7.4.1. When a planar CNT-FET is compared with a planar Si-MOSFET with similar insulator parameters, the CNT-FET shows an on-current of $790\,\mu A\,\mu m^{-1}$ at $V_{DD} = 0.4$ V, in contrast to the $1100\,\mu A\,\mu m^{-1}$ value for silicon. In a following study, the same authors [44] computed the high-frequency performance of this ideal device and projected a unity gain cut-off frequency (f_T) of 1.8 THz. Their analysis also showed that the parasitic capacitance dominates the intrinsic gate capacitance by three orders of magnitude. In a similar investigation, Hasan *et al.* [45] computed f_T to be a maximum of $130\,L^{-1}$ GHz, where L is the channel length in microns. As there is a desire to increase the current drive and reduce parasitic capacitance per tube, parallel array of nanotubes as the channel warrants

consideration [44]. However, crosstalk becomes a serious issue in multiple tube systems. Leonard [46] analyzed this point, and established a length scale for tube separation below which inter-tube interaction becomes significant. For small channel lengths, this critical separation distance depends on the channel length; for long channel devices, the critical inter-tube separation distance is independent of channel length but depends on gate oxide thickness and dielectric constant.

7.4.3 Memory Devices

In relative terms, very few studies have been conducted on the use of nanotubes as memory devices. Rueckes *et al.* [47] proposed a crossbar architecture for constructing non-volatile random access memory with a density of 10^{12} element per cm^2 and an operation frequency of over 100 GHz. In this architecture, nanotubes suspended in a $n \times m$ array act like electromechanical switches with distinct on and off states. A carbon nanotube-based flash memory was fabricated by Choi *et al.* [48], in which the source-drain gap was bridged with a SWNT as a conducting channel and the structure had a floating gate and a control gate. By grounding the source and applying 5 V and 12 V at the drain and control gate, respectively, a writing of 1 was achieved. This corresponds to the charging of the floating gate. To write 0, the source was biased at 12 V, the control gate fixed at 0 V, and the drain allowed to float. Now, the electrons on the floating gate were tunneled to the source and the floating gate was discharged. In order to read, a voltage V_R was applied to the control gate and, depending on the state of the floating gate (1 or 0), the drain current was either negligible or finite, respectively. Choi *et al.* [48] reported an appreciable threshold modulation for their SWNT flash memory operation.

7.5 Carbon Nanotubes in Silicon CMOS Fabrication

Whilst the active role of CNTs in nanoelectronics (i.e. as a conducting channel in a transistor device) may be far away, it may play an important role in extending silicon nanoelectronics. Several areas exist in the Semiconductor Industry Association Roadmap [16] where CNTs may be useful, such as interconnects, heat dissipation and metrology.

7.5.1 Interconnects

One of the anticipated problems in the next few generations of silicon devices is that the copper interconnect would suffer from electromigration at current densities of 10^6 A cm^{-2} and above. The resistivity of copper increases significantly for wiring line widths lower than 0.5 μm. In addition, the etching of deep vias and trenches and void-free filling of copper in high-aspect ratio structures may pose technological

challenges as progress continues along the Moore's law curve. All of these issues together demand investigation of alternatives to the current copper damascene process. In this regard, it is important to note that CNTs do not break down – even at current densities of 10^8 to 10^9 A cm^{-2} [49] – and hence can be a viable alternative to copper. Kreupl *et al.* [50] were the first to explore CVD-produced MWNTs in vias and contact holes. They measured a resistance of about 1 Ω for a 150 μm^2 via which contained about 10 000 MWNTs, thus yielding a resistance of 10 KΩ per nanotube. Further studies by this group demonstrated current densities of 5×10^8 A cm^{-2}, which exceeds the best results for metals, although the individual resistance of the MWNTs was still high at 7.8 KΩ. Srivastava *et al.* [51] provided a systematic evaluation of CNT and Cu interconnects and showed that, for local interconnects, nanotubes may not offer any advantages, partly due to the fact that practical implementations of nanotube interconnects have an unacceptably high contact resistance. On the other hand, their studies showed an 80% performance improvement with CNTs for long global interconnects. It is important to note that, even in the case of local interconnects, very few studies have been conducted on contact and interface engineering; however, with further investigation the situation may well improve beyond these early expectations.

Given the potential of CNTs as interconnects, it is necessary to devise a processing scheme that is compatible with the silicon integrated circuit fabrication scheme. In the via and contact hole schemes, Kreupl *et al.* [50] followed a traditional approach by simply replacing the copper filling step with a MWNT CVD step. If this proves to be reliable – and specifically if the MWNTs do not become unraveled during the chemical mechanical polishing (CMP) step – then it would be a viable approach, provided that a dense filling of vertical nanotubes can be achieved. Even then, the conventional challenges in deep aspect ratio etching and void-free filling of features, which arise due to shrinking feature sizes, remain. The dry etching of high-aspect ratio vias with vertical sidewalls will increasingly become a problem, and further processing studies must be performed to establish the viability of this approach. In the meantime, Li *et al.* [52] described an alternative bottom-up scheme (see Figure 7.11) wherein the CNT interconnect is first deposited using PECVD at prespecified locations. This is followed by tetraethylorthosilicate (TEOS) CVD of SiO_2 in the space between CNTs, and then by CMP to yield a smooth top surface of SiO_2 with embedded CNT interconnects. Top metallization completes the fabrication. The interconnects grown using PECVD in Ref. [52] are CNFs with a bamboo-like morphology. Whilst they are really vertical and freestanding compared to MWNTs, thus allowing ease of fabrication, their resistance is higher. This, combined with high contact resistance, resulted in a value of about 6 KΩ for a single 50-nm CNF. Further annealing to obtain higher quality CNFs and, more importantly, interface engineering to reduce contact resistance, can prove this approach valuable. It would also be useful for future three-dimensional architectures. A detailed theoretical study conducted by Svizhenko *et al.* [53] showed that almost 90% of the voltage drop occurs at the metal–nanotube interface, while only 10% is due to transport in the nanotube, thus emphasizing the need for contact interface engineering.

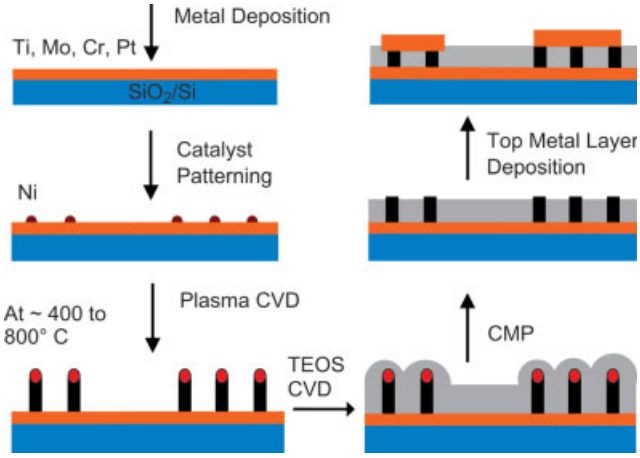


Figure 7.11 Carbon nanotube interconnect processing scheme for DRAM applications. TEOS = tetraethylorthosilicate.

7.5.2
Thermal Interface Material for Chip Cooling

The current trend in microprocessors is increasing operating frequency, decreasing dimensions, high packing density, and increasing power density. Together, these make thermal management in chip design a critical function to maintain the operating temperature at a prescribed, acceptable level. Otherwise, device reliability is severely compromised, and the speed of the microprocessor would also decrease with increasing operating temperatures [54]. The key figure-of-merit in thermal design and packaging is the thermal resistance, which is ΔT/input power. Here, ΔT is the temperature difference between the transistor junction and the ambient, which is fixed by the desirable operating junction temperature. As the power densities are on the rise, Shelling *et al.* [54] point out that the challenge is to develop high-conductivity structures which will accommodate the fixed ΔT, even with increased power densities.

Typically, the thermal packaging of a microprocessor consists of a heat spreader (primarily copper) and a heat sink. A variety of engineering designs is considered in all of the above to increase the heat-transfer efficiency [55] (which is beyond the scope of this chapter and not relevant at this point). However, one aspect which is relevant is a thermal interface material (TIM) commonly used to improve heat transfer between the chip and heat spreader, as well as between the heat spreader and the heat sink. Typically, a thermal grease has been used as TIM in the past, but more recent research on phase-change materials and polymers filled with high-conductivity particles has advanced knowledge of this subject. Carbon nanotubes exhibit very high axial thermal conductivity (see Section 7.2) which can be exploited in creating a TIM to address future thermal management needs.

Ngo *et al.* [56] reported a CNT-Cu composite for this purpose, wherein a PECVD-produced vertical CNF array is intercalated with copper using electrodeposition. As CNF surface coverage on the wafer from PECVD is only about 30–40% and air is a

poor conductor, it becomes necessary to undertake a "gap-filling" effort with copper to cover the space between the nanotubes. This structure maintained its structural integrity at the 60 psi pressure normally used in packaging. Ngo *et al.* [56] reported a thermal resistance of about 0.1 $cm^2\,K\,W^{-1}$ for this structure, which makes it desirable for laptop, desktop, and workstation processor chips, although further improvements and reliability testing are required in this area.

7.5.3 CNT Probes in Metrology

Atomic force microscopy (AFM) is a versatile technique for imaging a wide variety of materials with high resolution. In addition to imaging metallic, semiconducting and dielectric thin films in integrated circuit manufacture, AFM has been advocated for critical dimension metrology. Currently, the conventional probes of either silicon or silicon nitride which are sited at the end of an AFM cantilever have a tip radius of curvature about 20–30 nm, which is obtained by micromachining or reactive ion etching. These probes exhibit significant wear during continuous use, and the worn probes can also break during tapping mode or contact mode operation. Carbon nanotube probes can overcome the above limitations due to their small size, high aspect ratio and the ability to buckle reversibly. Their use in AFM was first demonstrated by Dai *et al.* [57], while a detailed discussion of CNT probes and their construction and applications is also available [58].

A SWNT tip, attached to the end of an AFM cantilever, is capable of functioning as an AFM probe and provides better resolution than conventional probes. This SWNT probe can be grown directly using thermal CVD at the end of a cantilever [59]. An image of an iridium thin film collected using a SWNT probe is shown in Figure 7.12.

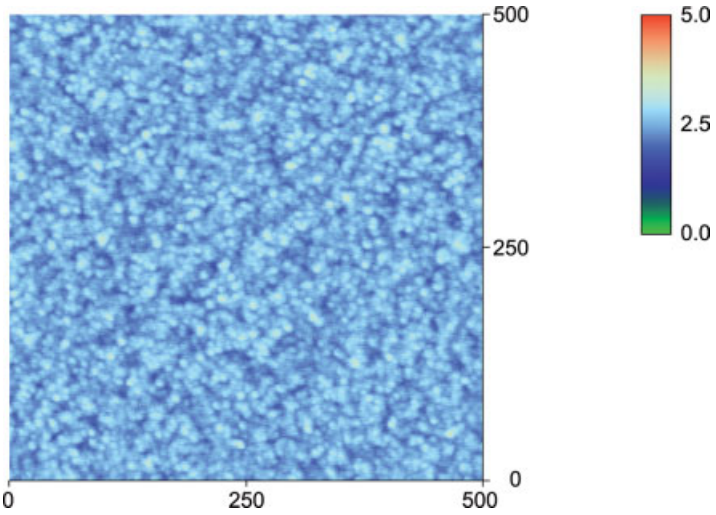


Figure 7.12 Atomic force microscopy image of an iridium thin film collected using a SWNT probe.

The nanoscale resolution is remarkable, but more importantly the tip has been shown to be very robust and significantly slow-wearing compared to conventional probes [59]. Due to thermal vibration problems, the SWNTs with a typical diameter of 1 to 1.5 nm cannot be longer than about 75 nm for probe construction. In contrast, however, the MWNTs – with their larger diameter – can form 2- to 3-μm-long probes. It is also possible to sharpen the tip of MWNTs to reach the same size as SWNTs, thus allowing the construction of long probes without thermal stability issues, but with the resolution of SWNTs [60]. Both, SWNT and sharpened MWNT probes have been used to image the semiconductor, metallic, and dielectric thin films commonly encountered in integrated circuit manufacture [58–60].

In addition to imaging, MWNT probes find another important application in the profilometry associated with integrated circuit manufacture. As via and other feature sizes continue to decrease, it will become increasingly difficult to use conventional profilometers to obtain sidewall profiles and monitor the depth of features. Although AFM is advocated as a replacement in this respect, the pyramidal nature of standard AFM probes would lead to artifacts when constructing the sidewall profiles of trenches. Hence, a 7- to 10-nm MWNT probe might be a natural choice for this task. An image of a MWNT probe for this purpose, and the results of profiling a photoresist pattern generated by interferometric lithography, are shown in Figure 7.13. While early attempts consisted of manually attaching a SWNT to a cantilever [57], followed by direct CVD of a nanoprobe on a cantilever [58, 59], Ye *et al*. [61] reported the first batch fabrication of CNT probes on a 100-mm wafer

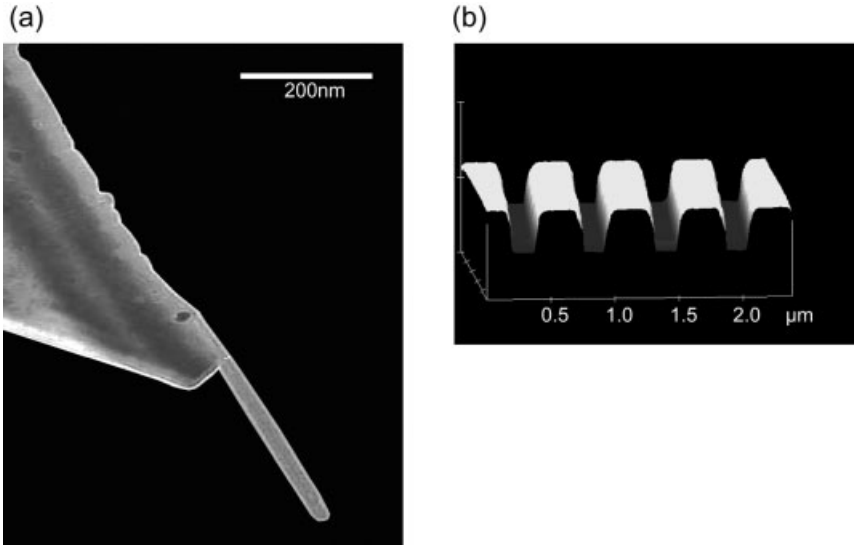


Figure 7.13 (a) Transmission electron microscopy image of a MWNT at the tip of an atomic force microscope cantilever. (b) Profile of a deep-UV photoresist pattern generated by interferometric lithography. The array has a pitch of 500 nm.

using PECVD. Unfortunately, the yield obtained was only modest, due mainly to difficulties encountered in controlling the angle of the nanotube to the plane.

7.6 Summary

In this chapter, the current status of CNT-based electronics for logic and memory devices has been discussed. Single-walled CNTs exhibit intriguing electronic properties that make them very attractive for future nanoelectronics devices, and early studies have confirmed this potential. Even with substantially longer channel lengths and thicker gate oxides, the performance of CNT-FETs is better than that of current silicon devices, although of course the design and performance of the former are far from being optimized. While all of this is impressive, the real challenge is in the integration of a large number of devices at reasonable cost to compete with and exceed the performance status quo of silicon technology at the end of the Moore's law paradigm. In addition, all of the studies conducted to date have been along the lines of following silicon processing schemes, with one-to-one replacement of a silicon channel with a CNT channel while maintaining the circuit and architectural schemes. Thus, aside from changing the channel material, there is no novelty in this approach. The structure and unique properties of SWNTs may be ideal for bold, novel architectures and processing schemes, for example in neural or biomimetic architecture, although very few investigations have been carried out in such non-traditional directions. Clearly, CNTs in active devices are a long-term prospect, at least a decade or more away. In the meantime, opportunities exist to include this extraordinary material into silicon CMOS fabrication not only as a high-current-carrying, robust interconnect but also as an effective heat-dissipating, thermal interface material.

Acknowledgments

The author is grateful to his colleagues at NASA Ames Center for Nanotechnology for providing much of the material described in this chapter.

References

1 S. Iijima, *Nature* 1991, **354**, 56.
2 M. Meyyappan (Ed.), *Carbon Nanotubes: Science and Applications*, CRC Press, Boca Raton, FL, 2004.
3 J. Han,in: M. Meyyappan (Ed.), *Carbon Nanotubes: Science and Applications*, CRC Press, Boca Raton, FL, 2004, Chapter 1.
4 M. A. Osman, D. Srivastava, *Nanotechnology* 2001, **12**, 21.
5 J. Hone, M. Whitney, C. Piskoti, A. Zetti, *Phys. Rev. B* 1999, **59**, R2514.
6 T. W. Ebbesen, P. M. Ajayan, *Nature* 1992, **358**, 220.
7 T. Guo, P. Nikolaev, A. Thess, D. T. Colbert, R. E. Smalley, *Chem. Phys. Lett*. 1995, **243**, 49.
8 M. Meyyappan,in: M. Meyyappan (Ed.), *Carbon Nanotubes: Science and Applications*,

CRC Press, Boca Raton, FL, 2004, Chapter 4 (and references therein).
9 L. Delzeit, B. Chen, A. M. Cassell, R. M. D. Stevens, C. Nguyen, M. Meyyappan, *Chem. Phys. Lett.* 2001, **348**, 368.
10 M. Meyyappan, L. Delzeit, A. Cassell, D. Hash, *Plasma Sources Sci. Technol.* 2003, **12**, 205.
11 K. Teo, D. Hash, R. Lacerda, N. L. Rupesinghe, M. B. Sell, S. H. Dalal, D. Bose, T. R. Govindan, B. A. Cruden, M. Chhowala, G. A. J. Amaratunga, M. Meyyappan, W. L. Milnes, *Nano Lett.* 2004, **4**, 921.
12 V. I. Merkulov, D. H. Lowndes, Y. Y. Wei, G. Eres, E. Voelkl, *Appl. Phys. Lett.* 2000, **76**, 3555.
13 M. Chhowalla, K. B. K. Teo, C. Ducati, N. L. Rupesinghe, G. A. J. Amaratunga, A. C. Ferrari, D. Roy, J. Robertson, W. I. Milne, *J. Appl. Phys.* 2001, **90**, 5308.
14 C. Bower, W. Zhu, S. Jin, O. Zhou, *Appl. Phys. Lett.* 2000, **77**, 830.
15 K. Matthews, B. A. Cruden, B. Chen, M. Meyyappan, L. Delzeit, *J. Nanosci. Nanotech.* 2002, **2**, 475.
16 International Technology Roadmap for Semiconductors (Semiconductor Industry Association, San Jose, CA 2001); http://public.itrs.net/.
17 T. Yamada,in: M. Meyyappan (Ed.), *Carbon Nanotubes: Science and Applications*, CRC Press, Boca Raton, FL, 2004, Chapter 7.
18 S. J. Tans, A. R. M. Verschueren, C. Dekker, *Nature* 1998, **393**, 49.
19 R. Martel, T. Schmidt, H. R. Shen, T. Hertel, Ph. Avouris, *Appl. Phys. Lett.* 1998, **76**, 2447.
20 V. Derycke, R. Martel, J. Appenzeller, Ph. Avouris, *Appl. Phys. Lett.* 2002, **80**, 2447.
21 S. J. Wind, J. Appenzeller, R. Martel, V. Derycke, Ph. Avouris, *Appl. Phys. Lett.* 2002, **80**, 3817.
22 B. Yu, *Proc. IEDM* 2001, 937.
23 F. Nihey, H. Hongo, M. Yudasaka, S. Iijima, *J. Appl. Phys.* 2002, **41**, L1049.
24 R. V. Seidel, A. P. Graham, J. Kretz, B. Rajasekharan, G. S. Duesberg, M. Liebau, E. Unger, F. Kreupl, W. Hoenlein, *Nano Lett.* 2005, **5**, 147.
25 X. Liu, S. Han, C. Zhou, *Nano Lett.* 2006, **6**, 34.
26 W. B. Choi, B. H. Cheong, J. J. Kim, J. Chu, E. Bae, *Adv. Funct. Mater.* 2003, **13**, 80.
27 W. B. Choi, E. Bae, D. Kang, S. Chae, B. H. Cheong, J. H. Ko, E. Lee, W. Park, *Nanotechnology* 2004, **15**, S512.
28 X. Liu, C. Lee, C. Zhou, J. Han, *Appl. Phys. Lett.* 2001, **79**, 3329.
29 V. Derycke, R. Martel, J. Appenzeller, Ph. Avouris, *Nano Lett.* 2001, **1**, 453.
30 A. Bachtold, P. Hadley, T. Nakanishi, C. Dekker, *Science* 2001, **294**, 1317.
31 Y. M. Lin, J. Appenzellar, J. Knoch, Z. Chen, Ph. Avouris, *Nano Lett.* 2006, **6**, 930.
32 A. Raychowdhury, K. Roy, *IEEE Trans. Nanotechnol.* 2005, **4**, 168.
33 M. Menon, D. Srivastava, *Phys. Rev. Lett.* 1997, **79**, 4453.
34 M. Menon, D. Srivastava, *J. Mater. Res.* 1998, **13**, 2357.
35 B. C. Satishkumar, P. J. Thomas, A. Govindaraj, C. N. R. Rao, *Appl. Phys. Lett.* 2000, **77**, 2530.
36 A. M. Cassell, G. C. McCool, H. T. Ng, J. E. Koehne, B. Chin, J. Li, J. Han, M. Meyyappan, *Appl. Phys. Lett.* 2003, **82**, 817.
37 W. B. Choi, unpublished results.
38 D. Srivastava, M. Menon, K. J. Cho, *Comput. Sci. Eng.* 2001, **3**, 42.
39 T. Yamada, *Appl. Phys. Lett.* 2000, **76**, 628.
40 T. Yamada, *Appl. Phys. Lett.* 2001, **78**, 1739.
41 T. Yamada, *Appl. Phys. Lett.* 2002, **80**, 4027.
42 T. Yamada, *Phys. Rev. B* 2004, **69**, 123408.
43 J. Guo, M. Lundstrom, S. Datta, *Appl. Phys. Lett.* 2002, **80**, 3192.
44 J. Guo, S. Hasan, A. Javey, G. Bosman, M. Lundstrom, *IEEE Trans. Nanotechnol.* 2005, **4**, 715.
45 S. Hasan, S. Salahuddin, M. Vaidyanathan, M. A. Alan, *IEEE Trans. Nanotechnol.* 2006, **5**, 14.
46 F. Leonard, *Nanotechnology* 2006, **17**, 2381.
47 T. Rueckes, K. Kim, E. Joselevich, G. Y. Tseng, C. L. Cheung, C. M. Lieber, *Science* 2000, **289**, 94.

48 W. B. Choi, S. Chae, E. Bae, J. W. Lee, B. Cheung, J. R. Kim, J. J. Kim, *Appl. Phys. Lett*. 2003, **82**, 275.

49 B. Q. Wei, R. Vajtai, P. M. Ajayan, *Appl. Phys. Lett*. 2001, **79**, 1172.

50 F. Kreupl, A. P. Graham, G. S. Duesberg, W. Steinhogl, M. Liebau, E. Unger, W. Honlein, *Microelec. Eng*. 2002, **64**, 399.

51 N. Srivastava, R. V. Joshi, K. Banerjee, *IEDM Proc*. 2005, 257.

52 J. Li, Q. Ye, A. Cassell, H. T. Ng, R. Stevens, J. Han, M. Meyyappan, *Appl. Phys. Lett*. 2003, **82**, 2491.

53 A. Svizhenko, M. P. Anantram, T. R. Govindan, *IEEE Trans. Nanotechnol*. 2005, **4**, 557.

54 P. Schelling, L. Shi, K. E. Goodson, *Mater. Today* 2005, 30.

55 R. Viswanatha, V. Wakharkar, A. Watwe, V. Lebonheur, *Intel Tech. J*. 2000, **Q3**, 1.

56 Q. Ngo, B. A. Cruden, A. M. Cassell, G. Sims, M. Meyyappan, J. Li, C. Yang, *Nano Lett*. 2004, **4**, 2403.

57 H. Dai, J. H. Hafner, A. G. Rinzler, D. T. Colbert, R. E. Smalley, *Nature* 1996, **384**, 147.

58 C. V. Nguyen,in: M. Meyyappan (Ed.), *Carbon Nanotubes: Science and Applications*, CRC Press, Boca Raton, FL, 2004, Chapter 6.

59 C. V. Nguyen, K. J. Chao, R. M. D. Stevens, L. Delzeit, A. M. Cassell, J. Han, M. Meyyappan, *Nanotechnology* 2001, **12**, 363.

60 C. V. Nguyen, C. So, R. M. D. Stevens, Y. Li, L. Delzeit, P. Sarrazin, M. Meyyappan, *J. Phys. Chem. B* 2004, **108**, 2816.

61 Q. Ye, A. M. Cassell, H. Liu, K. J. Chao, J. Han, M. Meyyappan, *Nano Lett*. 2004, **4**, 1301.

8
Concepts in Single-Molecule Electronics

Björn Lüssem and Thomas Bjørnholm

8.1
Introduction

Molecular electronics is a wide field of research, which consists of such diverging topics as organic light-emitting diodes (OLEDs), organic field effect transistors (OFETs; see Chapter 9) or, the topic of this chapter, single-molecule devices. Whereas, OLEDs and OFETs exploit the properties of a large number of molecules, in the field of single-molecule electronics an attempt is made to condense the entire functionality of an electronic device into a single molecule.

The field of (single) molecular electronics owes its significance to the tremendous downscaling that microelectronics has experienced during the past decades. In the ITRS roadmap [1], it is expected that, by the year 2013, the physical gate length of a transistor will scale down to 13 nm – that is, the transistor channel will consist of only a couple of atoms in a row. In order to obtain reliable devices, the composition of the devices must be controlled to only a few atoms – a demand that seems not to be feasible for conventional lithographic methods.

Chemistry – and especially organic chemistry – learned long ago how to control precisely the composition of a molecule to the last atom. Thus, the utilization of single organic molecules can be regarded as the ultimate miniaturization of electronic devices.

The concept of single-molecule electronics was first suggested in 1974 by Aviram and Ratner [2], who proposed that a single molecule consisting of a donor and an acceptor group could function as a diode. Unfortunately, however, at that time it was experimentally not feasible to test these predictions.

Molecular electronics gained impetus during the 1990s and early 2000s, when several molecular devices were proposed, including a single molecular switch [3] or a diode showing a negative differential resistance [4]. These early results raised great expectations, as evidenced by the election of molecular electronics as the "breakthrough of the year 2001" [5]. However, only two years later, the fledgling field

Nanotechnology. Volume 4: Information Technology II. Edited by Rainer Waser

ISBN: 978-3-527-31737-0

of molecular electronics experienced its first drawback when it was reported that some of the early results might be due to artifacts [6].

These reports on possible artifacts helped to settle the expectations laid on molecular electronics to a reasonable level, and in the current phase of development more emphasis has been placed on proving that the observed effects are in fact "molecular", and on identifying experimental set-ups that avoid the possible introduction of artifacts.

In this chapter, a brief overview is provided of the field of single-molecule electronics, beginning with a short theoretical introduction that aims to define the concepts and terminology used. (A more extensive explanation of the theory can be found in Part A of Volume III of this series.) In the following sections the text is more factual, and relates to how single molecules can actually be contacted and which functionalities they can provide. The means by which these molecules may be assembled to implement complex logical functions are then described, followed by a brief summary highlighting the main challenges of molecular electronics.

8.2 The General Set-Up of a Molecular Device

In this section, the basic concepts used in subsequent sections will be explained, and the presence of two domains of current transport – *strong coupling* and *weak coupling* – will be outlined.

Electrical transport across single molecules is remarkably different from conduction in macroscopic wires. In a large conductor, charge carriers move with a mean drift velocity v_d, which is proportional to the electric field, E. Together with the density of free charge carriers, this proportionality gives rise to Ohm's law.

For a single molecule this model is not applicable. Instead of considering drift velocities and resistances, which are only defined as average over a large number of charge carriers, concern is centered on the transmission of electrons across the molecule.

The general set-up of a molecular device is shown in Figure 8.1. The molecule is connected by two electrodes, labeled "source" and "drain", while the electrostatic potential of the molecule can also be varied by using a gate electrode.

If a voltage is applied between the source and drain (i.e. a negative voltage with respect to the drain), the electrochemical potential of the source, μ_S, shifts up and the potential of the drain, μ_D, moves down. An energy window is opened between these two potentials; in this energy window filled states in the source oppose empty states at the same energy in the drain.

However, as the two electrodes are isolated from each other, electrons cannot easily flow from the source to the drain. Only if a molecular level enters the energy window between μ_S and μ_D, can electron transport be mediated by these molecular levels (see Figure 8.1b). Therefore, each time the electrochemical potential of the source aligns with a molecular level the current rises sharply. In Figure 8.1b, for example, the source potential exceeds the lowest unoccupied molecular orbital (LUMO), and

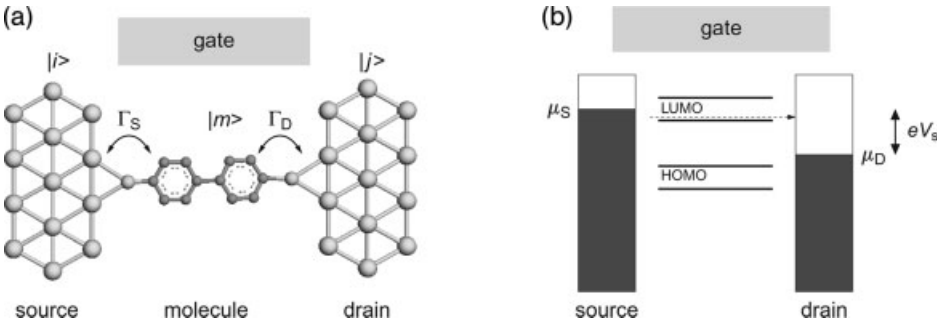


Figure 8.1 General set-up of a molecular device. In (a) a molecule is coupled to the source and drain (coupling strengths Γ_S and Γ_D). In (b) the molecule is replaced by its molecular levels. The electrodes are filled with electrons up to their electrochemical potential (indicated by the hatched area).

electrons can be transmitted across this level. Similarly, the drain potential can drop below the highest occupied molecular orbital (HOMO), which would also initiate electron flow.

8.2.1 The Strong Coupling Regime

Depending on the strength of the coupling of the electrodes with the molecule, there are two domains of the electron transport: the *weak coupling limit* and the *strong coupling limit*. To distinguish between these two limits, coupling strengths Γ_S and Γ_D can be defined that describe how strongly the electronic states of the source or drain $|i\rangle$, $|j\rangle$ interact with the molecular eigenstates $|m\rangle$. Γ_S and Γ_D have the dimension of energy; a high energy means that the electrode states can strongly couple with the molecular states.

High coupling energies therefore result in electronic wavefunctions of the electrodes that can extend into the molecule, so that charge can be easily transmitted from the source, across the molecule towards the drain. Thus, the current across the molecule can be expressed in terms of a transmission coefficient $T(E)$

$$I = \frac{2e}{h} \int_{\mu_D}^{\mu_S} T(E) dE \tag{8.1}$$

where e is the elementary charge and h is Planck's constant.

The transmission coefficient represents the transmission probability of electrons with a certain energy E to be transmitted across the molecule. This probability peaks at the molecular levels. It can be shown that the maximum conductance per molecular level $G_0 = \frac{\Delta I}{\Delta E \cdot e}$ becomes [7, 8]

$$G_0 = \frac{2e^2}{h} \tag{8.2}$$

This maximum conductance of a single electronic level is known as *quantum of conductance*, and corresponds to a resistance of 12.5 kΩ. Most interestingly, this conductance is not dependent on the length of the molecule as long as the ideal molecular level extends between the source and the drain electrodes.

8.2.2 The Weak Coupling Regime

In comparison to the strong coupling limit, current transport is remarkably different, if the coupling strengths Γ_S and Γ_D are weak. Here, the wavefunction of the electrodes cannot extend into the molecule and charge cannot be easily transmitted across the molecule. Rather, electrons must hop or tunnel *sequentially* from the source onto the molecule, and finally from the molecule to the drain. The lowest amount of charge that can be transferred onto the molecule is the elementary charge, e. This has an interesting consequence.

The molecule is electrostatically connected to the source, drain and gate electrode by the capacitances C_S, C_D, and C_G, respectively (see Figure 8.2a). Therefore, it is no longer sufficient that the electrochemical potential of the source aligns with the molecular level. Additionally, it must supply enough energy ($E_C(N)$, where N is the number of electrons) to charge the capacitances with an additional electron.

$$E_C(N+1) = \frac{1}{2}\frac{(N+1)^2e^2 - N^2e^2}{\underbrace{C_S + C_D + C_G}_{C_\Sigma}} = \frac{\left(N+\frac{1}{2}\right)e^2}{C_\Sigma} \tag{8.3}$$

To include this energy, the energy diagram shown in Figure 8.1b may be refined (see Figure 8.2b). Two levels are included in this diagram, which correspond to the HOMO (lower) and LUMO states shown in Figure 8.1b. The LUMO is floated upwards by $E_C(1) = \frac{e^2}{2C_\Sigma}$, while the HOMO is moved down by the same amount.

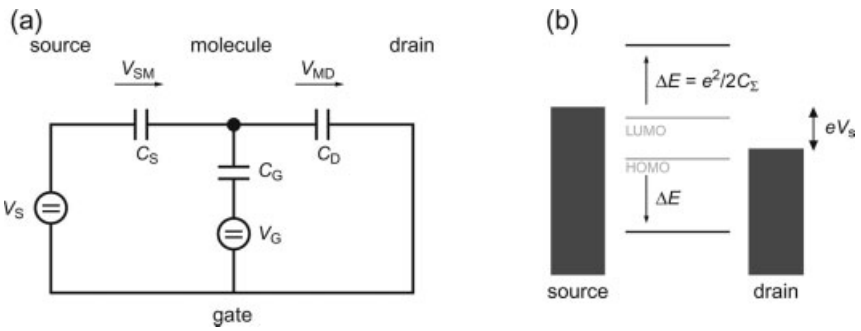


Figure 8.2 (a) The molecule is coupled to source, gate and drain by capacitances. (b) The energy level in the weak coupling limit. Additional charging energy must be provided by the source voltage.

Thereby, a large energy gap is opened within which no electrons can flow and hence the current is blocked; this effect is known as *coulomb blockade*.[1)]

8.3 Realizations of Molecular Devices

In the preceding section, the coupling of the molecule to the electrodes was described by the coupling strengths Γ_S and Γ_D. However, this theoretical description of contact between molecule and electrodes hides the complexity and difficulties that must be overcome in order to contact a single molecule. The main strategy that is followed to contact single molecules is to use specifically designed molecular anchoring groups that bind and self-organize on the contacts. In the following section, some key examples are provided of anchoring groups and self-organization strategies.

8.3.1 Molecular Contacts

Molecular end groups must provide a chemical bond to the contacting metal – that is, they must offer a self-organizing functionality. Furthermore, the nature of the contact determines the coupling strength Γ and, therefore, how strongly the molecular states couple with the electronic states of the electrodes. In the case of strong coupling, electrons can be easily transmitted across the molecule and the resistance of the molecule should be low; conversely, new effects such as coulomb blockade can occur for weak coupling, and this may be exploited for new devices. Thus, the suitable choice of molecular contact is one of the main issues in the design of a molecular device.

Various molecular contacts have been proposed. Besides the most common gold–sulfur bond, sulfur also binds to other metal such as silver [9] or palladium [10]. Sulfur may be replaced by selenium [11], which yields higher electronic coupling. A further increase in coupling strength is provided by dithiocarbamates [12, 13], which is explained by resonant coupling of the binding group to gold. Other binding groups include –CN [14], silanes [15], and molecules directly bound to either carbon [16] or silicon [17].

Using these binding groups, it is possible to contact single or at least a low number of molecules. In the past, several experimental set-ups have been developed which differ in the numbers of molecules contacted. Whereas some set-ups allow contacting single molecules (i.e. the method of *mechanically controlled break junctions, nanogaps* or *scanning probe methods*), in other arrangements the demand of a single molecule is relaxed and a small number of molecules is contacted (e.g. in the *crossed wire set-up* or in a *crossbar structure*). One further distinction between these set-ups is the number of electrodes that can contact each molecule – that is, if besides the source and drain a gate electrode is also present.

1) More details on single-electron effects can be found in Chapter 2 of Volume III in this series.

8.3.2
Mechanically Controlled Break Junctions

The concept of mechanically controlled break junctions dates back to 1985, when the method was used to obtain superconducting tunnel junctions [18]. In 1997, it was applied to contact a single molecule between two gold electrodes [19]. By comparing the current versus voltage characteristic of symmetric and asymmetric molecules, it was shown that only single molecules are contacted [20]. Since then, a variety of different molecules have been studied using this technique, and in particular the use of a low-temperature set-up was seen to provide a significant improvement in data quality [21–29].

The general set-up of a mechanically controlled break junction is shown in Figure 8.3. A metallic wire, which is thinned in the middle, is glued onto a flexible

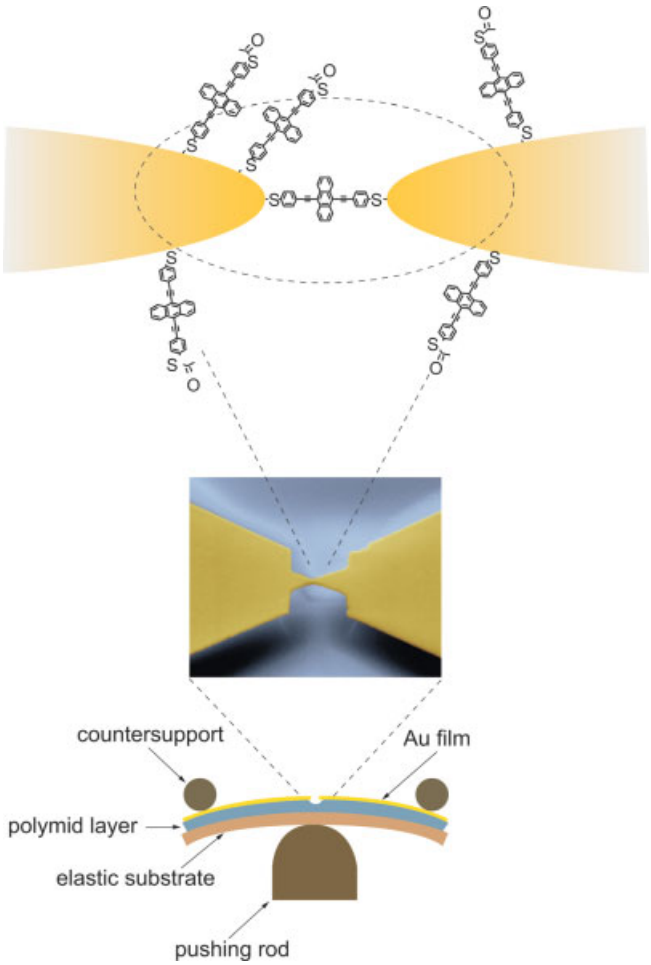


Figure 8.3 The mechanically controlled break junction. (From Ref. [30].)

substrate. Often, the wire is under etched so that a freestanding bridge is formed. Underneath the substrate, a piezo element can press the sample against two countersupports, which causes the substrate to bend upwards such that a strain is induced in the wire. If the strain becomes too large, the wire breaks and a small tunneling gap opens between the two parts of the wire. The length of the tunneling gap can be precisely controlled by the position of the piezo element.

To contact a single molecule, either a solution of the molecule is applied to the broken wire, or the molecules have already been preassembled onto the wire before it is broken. As described above, these molecules have chemical binding groups at both ends that easily bind to the material of the wire. As the molecule has binding groups at both ends it can bridge the tunneling gap if the length of the latter is properly adjusted. In this way a single-molecule device is formed.

Mechanically controlled break junctions represent a stable and reliable method for contacting single molecules. Most importantly, the correlation between molecular structure and current versus voltage characteristic can be studied, which will stimulate the understanding of the conduction through single molecules. At present, however, there is no way of integrating these devices – that is, it is not possible to contact a larger number of molecules in parallel and to combine these molecules into a logic device.

8.3.3
Scanning Probe Set-Ups

Due to its high spatial resolution, scanning probe methods (SPM) are well suited to contact single molecules, and several strategies have evolved during recent years.

One strategy is to contact a so-called self-assembled monolayer (SAM) of the molecule of interest with an atomic force microscope, using a conductive tip [31–34] (see Figure 8.4b). SAMs are formed by immersing a metallic bottom electrode into a solution of molecules (see Figure 8.4a) which must possess an end group that covalently binds to the metal layer. In the first layer, the molecules attach covalently to the metal; all following layers are then physisorbed onto this first chemisorbed layer. The physisorbed layers can easily be washed off using an additional rinsing step, such that only the first, chemisorbed, layer remains on the metal. A famous example of such a molecular end group/metal surface combination is sulfur on gold. Thiolates, and especially alkanethiols, are known to perfectly organize on a gold surface and to build a SAM that covers the gold electrode [35].[2)]

These SAMs can be contacted by a conductive atomic force microscopy (AFM) tip (see Figure 8.4b). Depending on the tip geometry, a low number of molecules (ca. 75) can be contacted [31]. Using this method, it has been shown that the current through alkanethiolates and oligophenylene thiolates decreases exponentially with the length of the molecule, and that the resistance of a molecule is dependent on the metal used to contact the molecule [36].

2) See also Chapter 9, which provides a broader introduction into self-organization and SAMs.

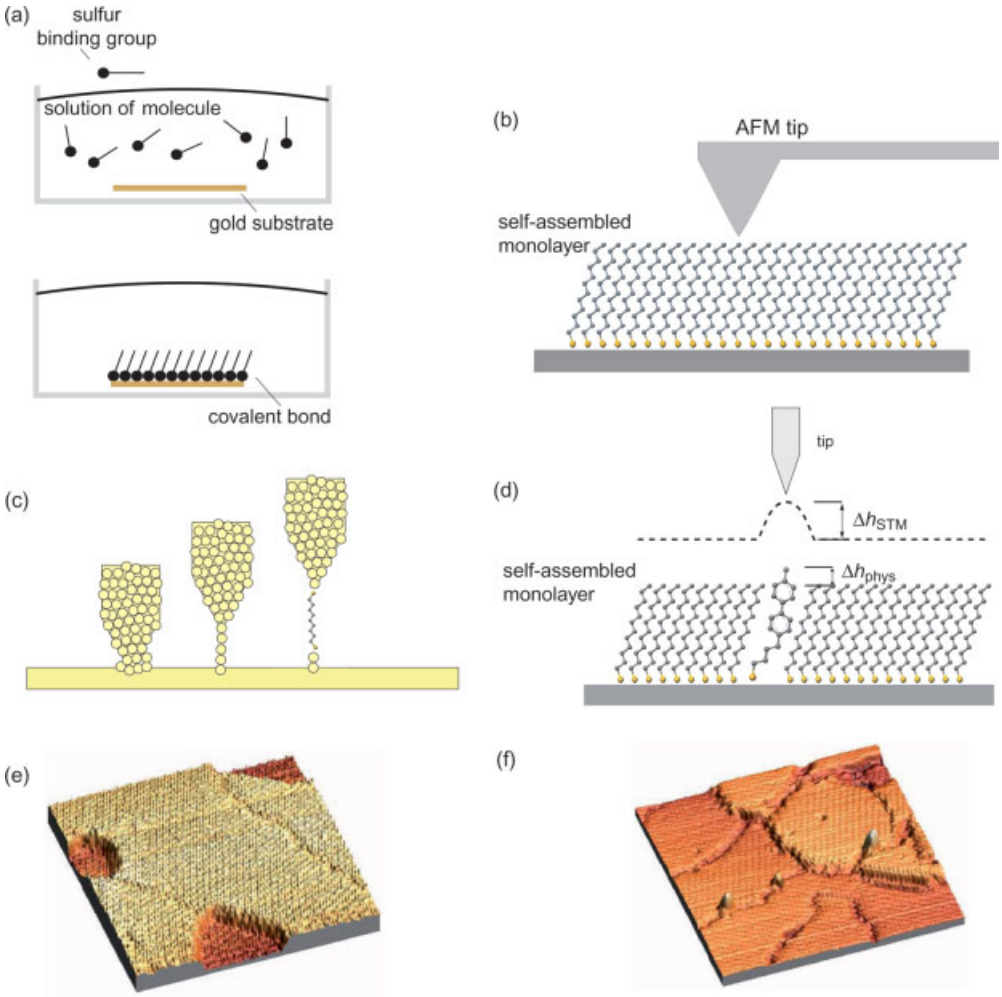


Figure 8.4 The different methods used to contact single molecules with SPM techniques. (a) The basic principle of self-assembled monolayer (SAM) formation. (b) A SAM of molecules is contacted by a conductive AFM tip. (c) The "tip crash" method, which forms a small tunneling gap. (d) Embedding conductive molecules in a SAM of insulating alkanethiols (molecules not drawn to scale). The height of the molecules can be used to deduce their conductivity. (e, f) Scanning tunneling microscopy (STM) image of a SAM of alkanethiols (e) and of oligo-phenylenevinylene molecules embedded into a SAM of alkanethiols. The molecules can be seen protruding from the SAM (f) [37].

An alternative method of contacting molecules using AFM or scanning tunneling microscopy (STM) is very similar to the break junction technique [38, 39]. A gold AFM or STM tip is moved into a gold substrate and subsequently slowly retracted (the "tip-crash" method shown in Figure 8.4c). Thereby, a thin gold filament is formed between the tip and the substrate. If the tip is moved too far away from the substrate,

the filament will break and a small tunneling gap is opened between the substrate and the tip. The whole set-up is immersed in a solution of molecules that have functional binding groups at both ends. If the tunneling gap is approximately the size of the molecule, there is a probability that one molecule will bind to the tip and the substrate and will therefore bridge the gap.

As a third strategy of contacting single molecules, the molecules of interest can be embedded into a SAM of insulating alkanethiols (see Figure 8.4d–f). In this way it is possible to obtain single, isolated molecules which "protrude" from the surrounding alkanethiol SAM (see Figure 8.4f). The conductance of the molecule can be measured either by placing the STM tip above the molecule [40, 41] or by measuring the height difference between the embedded molecule and the surrounding alkanethiol SAM [37, 42, 43]. This height difference is not only dependent on the differences in length of the alkanethiol and the molecule but also reflects differences in the conductivities of the molecules. Therefore, the conductivity of the embedded molecule can be calculated from the height difference.

8.3.4
Crossed Wire Set-Up

This set-up consists of two crossed wires, which almost touch at their crossing point. One of these wires is modified with a SAM of the molecule of interest. A magnetic field is applied perpendicular to one wire, and a dc current is passed through this wire. This causes the wire to be deflected due to the Lorentz force, and consequently the separation of the two wires can be adjusted by setting the dc current [44–47].

It has been shown that the number of contacted molecules is dependent on the wire separation, and that the current versus voltage characteristics measured at different separations are all integer multiplies of a fundamental characteristic. Thus, it is proposed that this fundamental curve represents the characteristic of only a single molecule [45].

These measurements can also be carried out at cryogenic temperatures [46]. In this case, the vibronic states of the molecule can be identified which provide a "molecular fingerprint" and prove that the molecule is actively involved in the conduction process (see also Section 8.4.6) [46].

8.3.5
Nanogaps

Similar to the break junction method, in the nanogap set-up a small gap is formed in a thin metal wire. However, this gap is not formed by bending a flexible substrate and mechanically breaking the wire, but it is prepared on a rigid substrate by using various methods.

One such method is *electromigration*. The preparation of the nanogap starts with the definition of a thin metallic wire on an insulating substrate. A SAM is then deposited on top of this wire by immersing the sample into a molecular solution.

Subsequently, a voltage ramp is applied to the wire. If the current that flows through the wire becomes too large, the wire breaks due to electromigration, which is reflected by a drop in current. Such gaps are approximately 1 nm in width [48] and, with a certain degree of probability, are bridged by a molecule that was deposited onto the wire in advance; thus, a single molecular device has been built.

Alternatively, nanogaps can be formed by preparing an electrode pair with a gap of ~30 nm [49, 50] using electron-beam lithography. This gap can be shrunk to molecular dimensions by electrochemically depositing metal atoms [51, 52]. Combined with electrochemical etching, this method allows a precise control over the wire separation. A combination of lithography and low-temperature evaporation has also been used to fabricate 1- to 2-nm gaps directly on a gate oxide [53].

Although all of these techniques define the gaps laterally, the precise control over the vertical thickness of thin films can be exploited to define a vertical nanogap [54]. The preparation of these gaps starts with the deposition of a thin SiO_2 layer on top of a highly p-doped silicon bottom electrode (see Figure 8.5). A gold electrode is then deposited on top of the SiO_2 layer, and subsequently the oxide can be etched in hydrofluoric acid, thus yielding a thin gap between the Si bottom and Au top electrode.

A rich variety of molecules has been measured using nanogaps, including a coordination complex containing a Co atom [55], a divanadium molecule [56], C60 [57, 58], C140 [59], and phenylenevinylene oligomers [53, 60].

8.3.6
Crossbar Structure

In terms of integration, the crossbar structure is a very interesting device set-up where the demand of single molecules is relaxed, and a rather low number of molecules are contacted.

In order to obtain a crossbar structure, parallel metallic wires are deposited onto an insulating substrate. A SAM of the molecule of interest is then deposited on top of these wires. Orthogonally to the bottom electrodes, metallic wires are deposited onto the SAM. Thus, a single crossbar structure consists of many possible devices (e.g. see Figure 8.21 in Section 8.5.1).

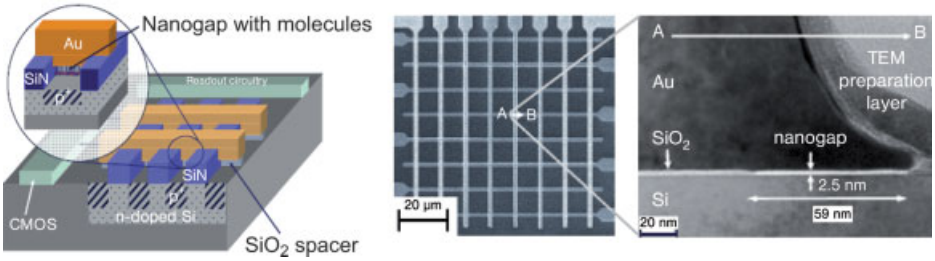


Figure 8.5 Vertical nanogaps integrated in a crossbar structure. (a) A schematic of the set-up. (b) Left: Scanning electron microscopy image of the crossbar, and (right) a transmission electron image of the nanogap. (From Ref. [54].)

The major technological problem in the crossbar set-up lies in the deposition of the top electrode. The metal/molecule interface may be unstable and metal ions can migrate through the molecular layer [61–63], thus shorting the device. The probability of metal ions penetrating the molecular layer is dependent on the molecular top group. A group which binds the metal at the top is more resistant, and many metal/molecular end groups have been examined, including Al on CO_2H [64], OH and OCH_3 [65], Cu, Au, Ag, Ca and Ti on OCH_3 [66, 67] or Au, Al and Ti on disulfides [68]. Ti is shown to be critical for metallization, because it reacts strongly with the molecule and partially destroys the SAM [69].

An alternative to these molecular end groups is to use aromatic end groups and to crosslink them with electron irradiation [70, 71]. This method yields stable Ni films on top of a molecular layer. Similarly, the molecular layer can be protected by a spun-on film of a highly conducting polymer film (e.g. PEDOT) [72].

In most devices the top electrode is deposited by evaporation techniques, so that the metal atoms arrive at the molecular layer with a high energy, and the probability of the atoms punching through the layer is high. Attempts have been made to reduce the energy of the atoms by indirect evaporation and cooling the substrate [73], or by so-called "printing methods" in which the metal film is gently deposited on the molecular layer from a polymeric stamp [74, 75].

8.3.7
Three-Terminal Devices

The incorporation of a third electrode (the gate) opens up new possibilities. First, the molecular levels can be shifted upwards and downwards by the gate relative to the levels of the electrodes, which can be used to analyze the electronic structure of the molecule. The gate is also necessary for building a molecular transistor. As will be shown later, these transistors can be used to build logic circuits.

The basic working principle of a molecular single-electron transistor is illustrated in Figure 8.6. Without a gate voltage applied, no molecular states lie in the energy window between the source and drain potential and thus, the current is blocked. However, the molecular level can be moved down into the energy window, if the gate voltage is increased. Therefore, by applying a gate voltage it is possible to switch the transistor on.

In order to understand, how the gate can shift the molecular levels, the capacitive network shown in Figure 8.2a should be considered. The voltage between the source and the molecule, V_{SM}, and between the molecule and the drain, V_{MD}, is related to the source V_S and source and gate voltage V_G as follows:

$$V_{SM} = \frac{C_D + C_G}{C_\Sigma} V_S - \frac{C_G}{C_\Sigma} V_G \tag{8.4}$$

$$V_{MD} = \frac{C_S}{C_\Sigma} V_S + \frac{C_G}{C_\Sigma} V_G \tag{8.5}$$

with $C_\Sigma = C_S + C_D + C_G$

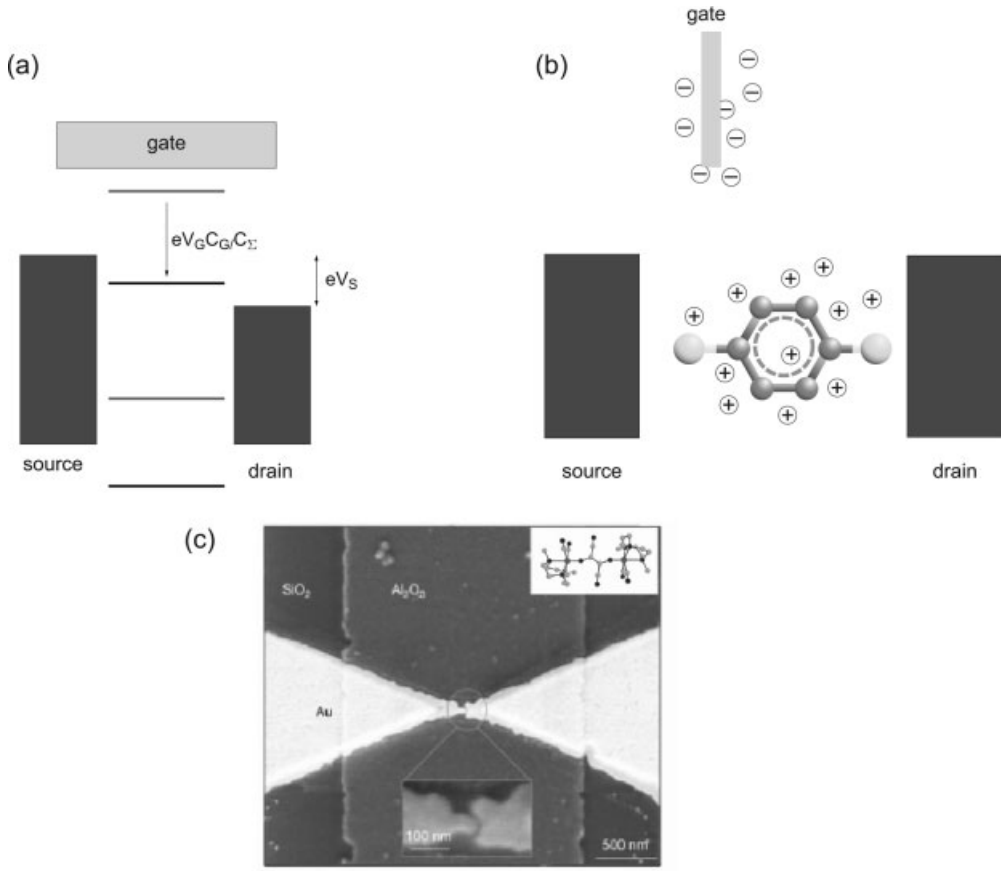


Figure 8.6 The working principle of a molecular single-electron transistor (a) and an electrochemical gate (b). In (c) a scanning electron image of a nanogap fabricated by the electromigration method is shown. The nanogap is prepared on an aluminum strip, which is covered by a thin Al_2O_3 layer and forms the gate. (From Ref. [56].)

Therefore, the molecular level shifts up or down relative to the source and drain energies. The amount of the shift is proportional to the term $\frac{C_G}{C_\Sigma} V_G$. In order to obtain good gate control, $\frac{C_G}{C_\Sigma}$ (the "gate-coupling parameter") must be large and, ideally, close to unity; hence, the gate must be placed very close to the molecule.

One elegant method of obtaining a high gate control is to use an electrochemical gate (c.f. Figure 8.6b). The molecular device (e.g. the nanogap or the STM set-up) is immersed in an electrolyte, and the source and drain voltages are varied relative to a reference electrode which is also immersed in the solution [38, 76, 78] and takes on the function of the gate. The effective gate distance is given by the thickness of the double layer of ions at the electrodes [38], which allows the application of

high electric fields. Several molecules, including peralene tetracarboxylic diimide [76], a molecule containing a viologen group [79], oligo(phenylene ethynylene)s [80] or different transition metal complexes [77], have been studied using this type of gate.

Champagne *et al.* succeeded in including a gate in a break junction set-up [81] which consists of a freestanding, under-etched gold bridge deposited onto a silicon wafer. Underneath the bridge, the silicon is degenerately doped and serves as the gate electrode. The bridge is broken by the electromigration technique, and the size of the so-formed gap is adjusted by bending the silicon substrate. A C_{60} molecule is immobilized in this gap, so that a molecular transistor with a gate-molecule spacing of about 40 nm is realized.

The most straightforward method of including a gate is provided by the nanogap set-up. Here, the source and drain are formed on an insulating substrate; however, the insulating layer (e.g. SiO_2 or Al_2O_3) can be very thin, and the underlying (conductive) substrate may be used as gate [55–58, 82] (cf. Figure 8.6c). Compared to an electrochemical gate, this set-up has the advantage that the measurements can be conducted at cryogenic temperatures, which makes the observance of coulomb blockade effects easier and also allows the use of inelastic electron tunneling spectroscopy (see Section 8.4.6) to study the molecules.

8.3.8
Nanogaps Prepared by Chemical "Bottom-Up" Methods

Several strategies for nanogap preparation have been based on the pioneering studies of Brust *et al.* [83, 84], where the chemical preparation of metal nanoparticles was protected by a ligand shell. Two-dimensional arrays of such particles constitute a test bed that may be used to interconnect metal particles separated by a few nanometers by various organic molecules (see Figure 8.7) [85, 86].

By mixing hydrophobic nanoparticles with surfactants, more one-dimensional structures may be formed where molecules interconnect segments of gold nanowires separated by a few nanometers (see Figure 8.8a) [87, 88].

By using a single metal particle inserted in a metal gap prepared "top-down", two well-defined nanogaps may also be realized at the gap–particle interface (cf. Figure 8.8b) [90, 91].

Although all of these systems are easily prepared and are stable at room temperature, as yet it has not been possible to control the gap formation accurately enough to prepare a single gap bridged by a single molecule. Neither have individual gates been reported.

8.3.9
Conclusion

At present, the list of measurement set-ups used in molecular electronics is by far incomplete, and an ever-growing number of techniques is available, including the mercury drop method [92], nanogap preparation by the deposition of gold electrodes

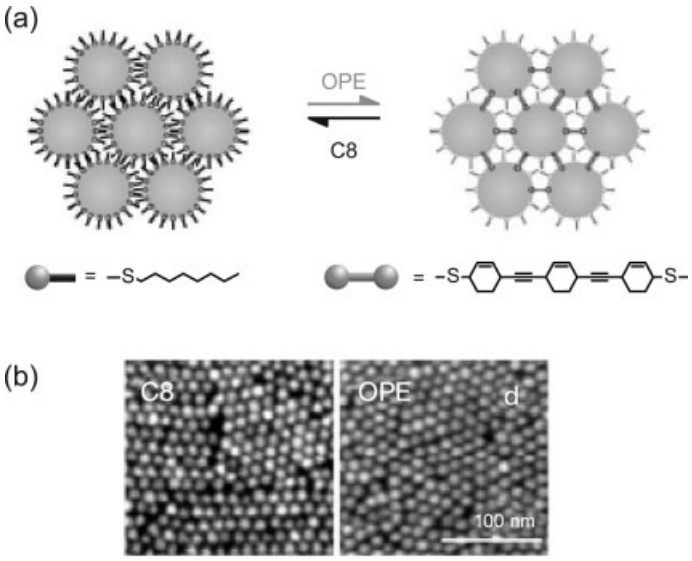


Figure 8.7 Two-dimensional gold nanoparticle arrays interlinked with octanethiol (left) and thiolated oligo(phenylene ethynylene) (right). (a) Schematic representation. (b) SEM image of the nanoparticle arrays. (From Ref. [85].)

through a shadow mask at shallow angles [53], the magnetic bead junction [90, 93], or the nanopore concept [94–97]. Each set-up has its own strengths and weaknesses: some allow the characterization of single molecules (e.g. break junction experiments and SPM set-ups), whereas others are interesting in terms of later applications (e.g. crossbar set-ups) or allow the inclusion of a gate electrode (nanogaps). It appears, however, that there is no ideal set-up, and often the intrinsic molecular behavior may be determined only by a combination of different experimental methods.

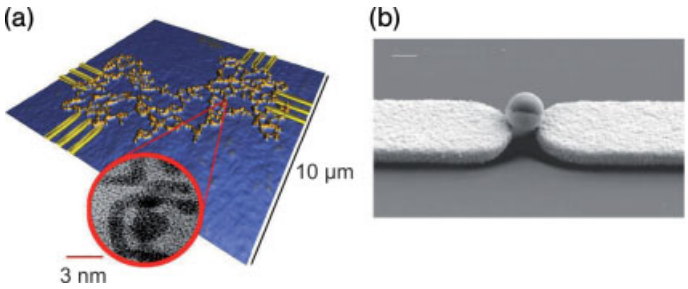


Figure 8.8 (a) Network of gold nanowires. The gaps can be bridged by molecules. (From ref. [89].) (b) A single nanoparticle immobilized between SAM-functionalized electrodes. (From Ref. [90].)

8.4 Molecular Functions

In this section, it is shown which molecules have been measured and which functionalities these molecules can provide. As the ultimate goal of molecular electronics is to provide a universal logic, each technology which aims to achieve this must fulfill several basic requirements [30].

The most basic requirement is that *a complete set of Boolean operators* can be built out of the molecular devices, such that every Boolean function can be obtained. One complete set of operators is for example a disjunction (OR) and an inversion (NOT), or alternatively, a conjunction (AND) and inversion. However, all complete sets have to include inverting gates – that is, an inversion.

Disjunction and conjunction can be relatively easily built out of a resistor and a diode (for a description, see Section 8.4.2.4), and so it is vital to identify molecules that conduct current only in one direction.

Similarly, complete fields of disjunctions and conjunctions can be implemented in crossbar structures in the form of so-called *programmable logic arrays* (PLA) (see Section 8.5.1). At the crossing points of these PLAs, molecules that can be switched on or off are needed; that is, the molecules must possess two conduction states, one isolating and one conducting. Molecules which demonstrate this behaviour include *hysteretic switches*, examples of which are described in Section 8.4.4.

These set-ups do not provide inverting logic, as negation is still missing. One set-up which provides inversion is the *molecular single electron transistor* (see Section 8.4.5). Even with only two terminal devices it is pfossible to construct an inversion using a so-called *crossbar latch* (see Section 8.4.4.1). Again, hysteretic switches are required for this. Inverting logic (e.g. an exclusive disjunction, XOR)[3] can also be built from a variant of a simple diode which displays a *negative differential resistance* (NDR) region – that is, a region in which the current drops with increasing voltage. These gates are described in Section 8.4.3.1.

A second requirement for the implementation of logic is that the gates must provide a means of *signal restoration*. At each stage of the circuit the signal voltage, which represents a logical 0 or 1, will be degraded. To be able to concatenate several logic gates, a means of restoring the original levels must be found, and this requirement can only be relaxed for small circuits, as long as the degradation of the signal voltage is tolerable.

In conventional CMOS logic, such restoration is provided by a non-linear transfer characteristic of the gates [30]. This strategy can also be followed with molecular single-electron transistors. Similarly, signal restoration can also be obtained by two terminal devices using hysteretic switches in the form of the crossbar latch or using NDR diodes in the form of the *molecular latch*, as proposed by Goldstein *et al.* [98].

Another requirement for the technology is that there must be elements that transmit signals across longer distances – that is, a type of *molecular wire* (see Section 8.4.1) must

3) The XOR gate can be converted into a negation if one input is fixed to "1".

be found. A molecular wire alone is insufficient, however, and a defined flow of information must be established, with feedback signals being prevented. This, again, can be achieved by using molecular diodes.

8.4.1 Molecular Wires

The most basic electronic device is a simple wire. However, in molecular electronics it is less easy than might be thought to construct a suitable molecule which transmits current with a low conductance across longer distances. So, what makes a molecule a good conductor? Starting from the short theoretical instruction provided in Section 8.2, certain conclusions can be drawn regarding the properties of an ideal molecular wire.

First, in order to obtain a low resistance a strong electronic coupling of the molecule and the electrodes is preferable. As discussed in Section 8.3.1, such a coupling can be obtained by choosing a suitable molecular binding group, for example a group that provides resonant coupling to the electrodes, such as dithiocarbamates.

Due to this choice of molecular binding group, the limit of strong coupling is valid and electrons are transmitted across the molecular levels. However, a prerequisite for such transmission is that this molecular level, extending from source to drain, actually exists. Extended molecular levels can be formed by delocalized π-systems, where in the tight binding approximation p_z orbitals of isolated carbon atoms add and form an extended, delocalized orbital. Therefore, aromatic groups (e.g. polyphenylene) are often used as the building blocks for molecular wires.

Another important property of a molecular wire is its ability to conduct current at a low bias, and therefore the molecular level used for transport should be close to the Fermi level of the contacts. Often, this requirement is expressed in terms of a low HOMO–LUMO gap.

Many molecules have been proposed as molecular wires, including polyene, polythiophene, polyphenylenevinylene, polyphenylene-ethynylene [99] (see Figure 8.9), oligomeric linear porphyrin arrays [100] or carbon nanotubes (CNTs) [[101] and references therein]. CNTs constitute a special category among molecular wire candidates as they may be either metallic or semiconducting, depending on their chirality. However, it is difficult to selectively prepare or isolate only one type of CNT, namely the metallic form. Furthermore, it is still challenging to organize and orient CNTs, although some techniques are available to arrange CNTs in a crossbar structure [e.g. see [102]]. A more detailed discussion on CNTs is provided in Chapter 10.

8.4.2 Molecular Diodes

Molecular diodes are the next step towards a higher complexity. Indeed, when combined with resistors, diodes are already sufficient to build AND and OR gates.

The first molecular electronic device to be proposed by Aviram and Ratner was just such a molecular diode [103], and consisted of a donor and an acceptor group

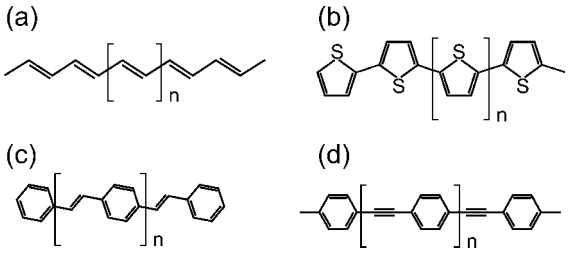


Figure 8.9 Building blocks for molecular wires: (a) polyene; (b) polythiophene; (c) polyphenylenevinylene; and (d) polyphenylene-ethynylene.

separated by a tunneling barrier. This set-up is often compared to the p- and n-layers in a conventional diode. As alternative to the Aviram–Ratner approach, a molecular diode can also be formed by asymmetric tunnel barriers at the source and drain electrodes [104]. This concept is based on the different electrostatic coupling of the electrodes. However, as will be seen later, it is very difficult to couple the molecule *symmetrically* to both electrodes, which in turn makes it difficult to distinguish between rectification due to the Aviram–Ratner mechanism and rectification due to asymmetric coupling.

8.4.2.1 **The Aviram–Ratner Concept**

The Aviram–Ratner concept is illustrated schematically in Figure 8.10. A molecule, which consists of an acceptor and a donor group, is connected to the source and drain. The acceptor and donor are isolated by a tunneling barrier, which ensures that the

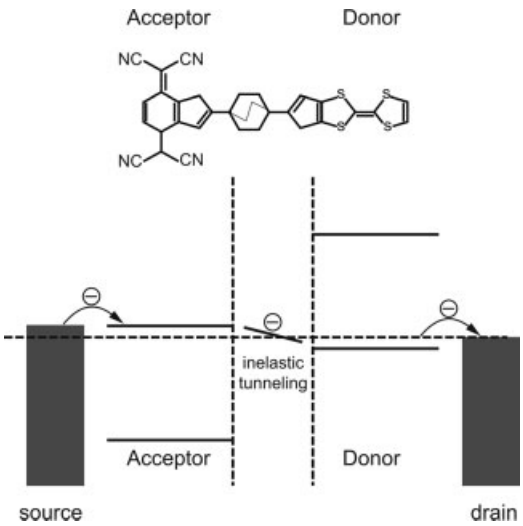


Figure 8.10 The diode proposed by Aviram and Ratner, and the suggested rectifying mechanism. For further details, refer to Section 8.4.2.1.

molecular levels of the two parts do not couple. The HOMO of the donor lies close to the Fermi level, in contrast to the acceptor, where the LUMO is adjacent to the Fermi level.

If a negative voltage V_S is applied to the diode (see Figure 8.2a for polarity), the potential of the source is raised with respect to the drain. Electrons can flow relatively easily from the source, across the acceptor and donor, towards the drain. However, at the opposite polarity a much higher voltage is needed to allow electrons to flow from drain to source. Thus, the molecule is considered to rectify the current.

8.4.2.2 Rectification Due to Asymmetric Tunneling Barriers

In contrast to the Aviram–Ratner mechanism, rectification due to asymmetric tunneling barriers is based on a difference in the source and drain capacitances. This difference can be obtained by attaching two insulating alkane chains to a conjugated part (see Figure 8.11). The alkane chains are functionalized with an end group, which provides binding functionality (e.g. sulfur for gold electrodes). The capacitance between the conjugated part of the molecule and the electrode is inversely proportional to the length of the alkane chains; varying these lengths is therefore a suitable way of adjusting the source and drain capacitances.

The rectifying mechanism can be explained by the energy diagram shown in Figure 8.11. The HOMO and the LUMO levels of the conjugated part of the molecule are included in the figure. These energy levels correspond to the molecular level of the (unbound) molecule plus the charging energy, as explained in Section 8.2. As can be seen in Figure 8.11, the levels lie asymmetrical with respect to the Fermi level of the electrodes.

Current can only flow when the electrochemical potential of the source or the drain aligns with, or even exceeds, the LUMO – that is, when $-eV_{SM} = \Delta$ for electrons flowing from source to drain, or $eV_{MD} = \Delta$ for the reverse bias. Here, Δ is the difference between the Fermi level of source and drain at zero bias and the LUMO.

V_{SM} and V_{MD} are given by Eqs. (8.4) and (8.5) (the gate capacitance must be set to zero). It follows for the voltage $V_{D \to S}$, at which electrons start to flow from drain to

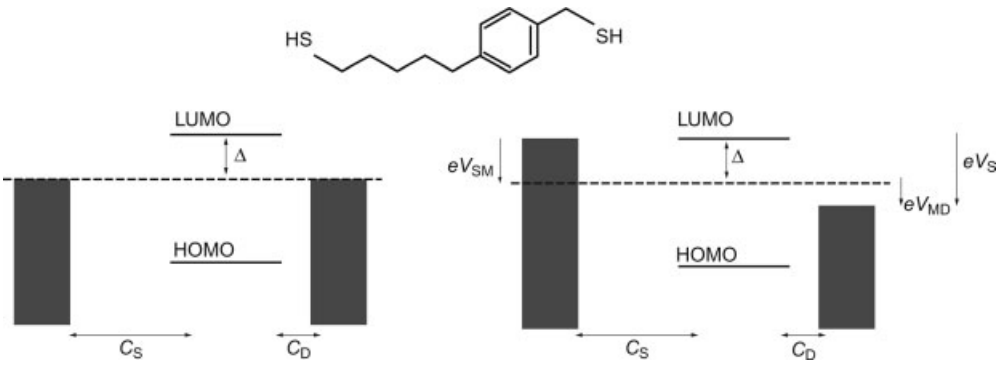


Figure 8.11 The concept of rectification due to asymmetric tunneling barriers.

source (which corresponds to a positive current flow from source to drain), and $V_{S\to D}$, at which electrons flow from source to drain [104]

$$V_{D\to S} = \frac{1 + C_S/C_D}{\underbrace{C_S/C_D}_{\equiv\eta}} \frac{\Delta}{e} = \frac{1+\eta}{\eta} \frac{\Delta}{e} \tag{8.6}$$

$$V_{S\to D} = -(1+\eta)\frac{\Delta}{e} \tag{8.7}$$

If the source capacitance is smaller than the drain capacitance ($\eta < 1$), $|V_{S\to D}|$ is smaller than $|V_{D\to S}|$ and electrons can flow from source to drain at a lower absolute bias than in the opposite direction; that is, the molecule shows a rectification behavior.

8.4.2.3 Examples

Starting from the molecule proposed by Aviram and Ratner (see Figure 8.10) [105], many other molecules containing acceptor and donor groups have been proposed [106]. One of the most extensively studied is γ-hexadecylquinolinium tricyannoquinodimethanide ($C_{16}H_{33}$ Q-3CNQ; c.f. Figure 8.12) [107–110]. Although this molecule consists of a donor and an acceptor group, it deviates from the normal Avriam–Ratner diode in that the two parts are not coupled by an insulating σ-group but rather by a delocalized π-group, which makes an analysis of the rectification behavior more difficult [109]. Furthermore, due to the alkane chain at one side of the molecule, it is often coupled asymmetrically to the electrodes. Thus, it is difficult to distinguish between the Aviram–Ratner mechanism and rectification due to asymmetric tunneling barriers. To circumvent this problem, decanethiol-coated gold STM tips are used as a second contact to a SAM of $C_{10}H_{21}$ Q-3CNQ, which places the same length of alkane groups at both ends of the donor and acceptor groups [111].

8.4.2.4 Diode–Diode Logic

Already with these rather simple diodes it is possible to build AND and OR gates in the form of so-called diode–diode logic [112]. In the AND gate (see Figure 8.13a) both

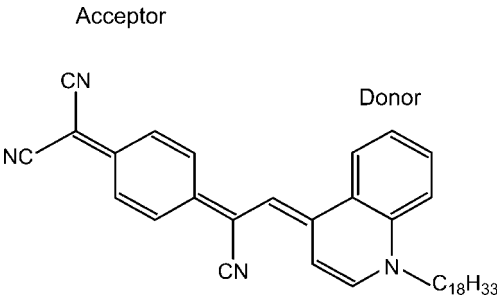


Figure 8.12 The molecular diode $C_{16}H_{33}$ Q-3CNQ in its neutral state.

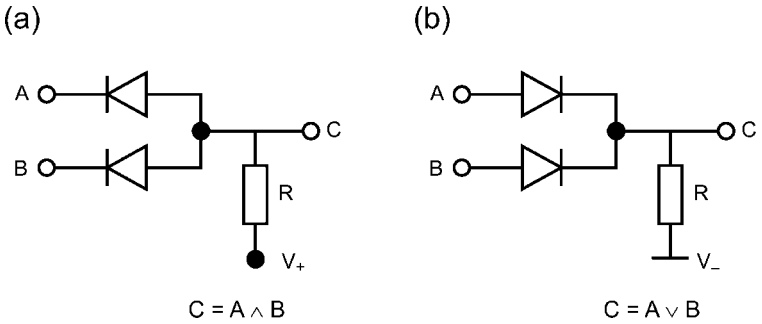


Figure 8.13 AND and OR gate using diode–diode logic.

inputs, A and B, are connected via reversely biased diodes and a resistor to the operating voltage. Only if both inputs are high (i.e. 1) is the output C high, and this results in an AND function.

The OR gate is shown in Figure 8.13b. In contrast to the AND gate, one input is already sufficient to push the output to a high voltage, and thus this gate implements a disjunction.

8.4.3
Negative Differential Resistance Diodes

Diode–diode logic yields AND and OR gates and, in order to obtain a complete set of Boolean variables, these gates can be combined with diodes that show a negative differential resistance (NDR). By using these modified gates it is possible to obtain inversion.

The NDR effect is illustrated schematically in Figure 8.14a. The current of the diode rises with increasing voltage up to a certain threshold voltage, above which the current drops. This rather odd behavior is already known from conventional semiconductors in the form of resonant tunneling diodes.

The concept of resonant tunneling can also be used for molecular NDR devices, as shown in Figure 8.14b and c [113]. The molecule consists of two conjugated molecular leads and an isolated benzene ring in the middle. In the absence of any inelastic processes, current can only flow if the molecular levels of the left lead of the isolated benzene ring and of the right lead, align. Such alignment occurs only at certain voltages, at which the levels are said to be in "resonance". Such resonance is illustrated graphically in Figure 8.14c. If the voltage is detuned from this resonant value – for example, if it is further increased – then the current will drop and the device will show an NDR effect.

Another molecule, which shows a prominent NDR effect, is shown in Figure 8.15b [4, 94]. It exhibits peak-to-valley ratios as high as 1030:1, although the exact nature of the NDR effect observed in this molecule is currently a matter of intense research [114].

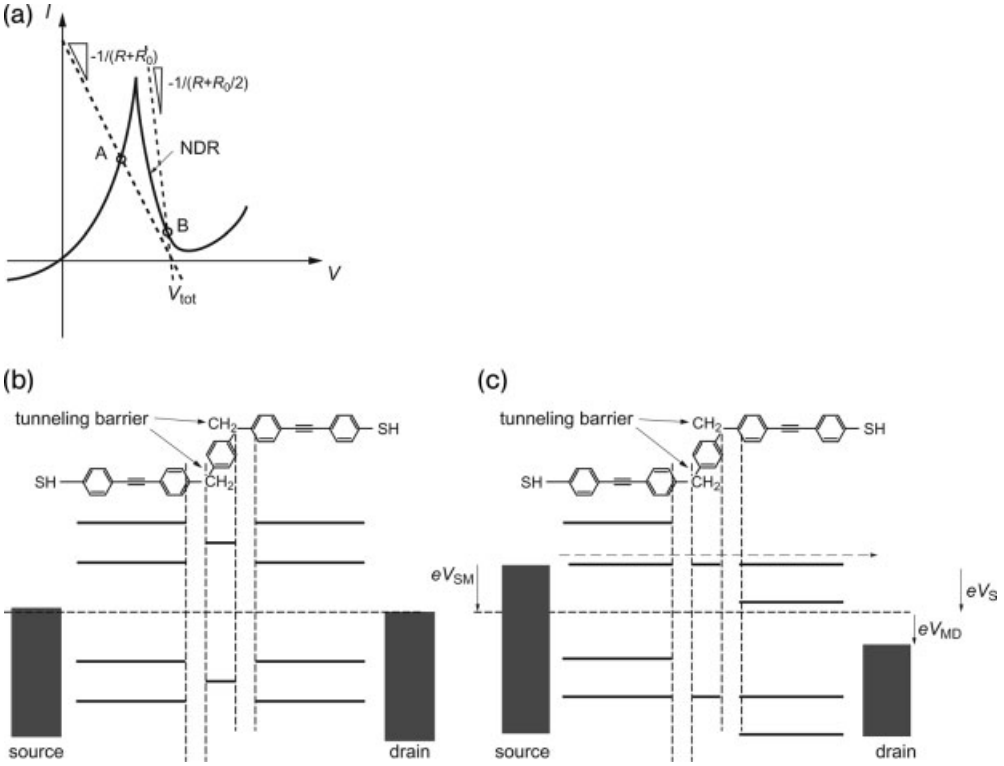


Figure 8.14 (a) The NDR effect. (b, c) The concept of a resonant tunneling diode consisting of a single molecule.

8.4.3.1 Inverting Logic Using NDR Devices

NDR devices can be used to obtain inverting logic. One example is shown in Figure 8.15a [112]; this gate implements an exclusive OR functionality – that is, the output C is high if either A or B is high, and low if both inputs are high or both inputs are low.

The XOR gate shown in Figure 8.15a resembles the normal OR gate shown in Figure 8.13b, the only difference being a NDR diode which is connected to the two diodes of inputs A and B. The voltage drop across the entire XOR circuit (V_{tot}, measured from input A and B to V–) is divided between a voltage drop across the NDR diode (V_{NDR}) and a voltage drop across the resistances R_0 and R (V_R). Assuming ohmic behavior for the resistances, it follows for the current flowing through the resistances:

$$I = \frac{V_R}{R + R_0} = \frac{V_{tot} - V_{NDR}}{R + R_0} \quad \text{if either A or B is high} \tag{8.8}$$

$$I = \frac{V_R}{R + R_0/2} = \frac{V_{tot} - V_{NDR}}{R + R_0/2} \quad \text{if both, A and B, are high} \tag{8.9}$$

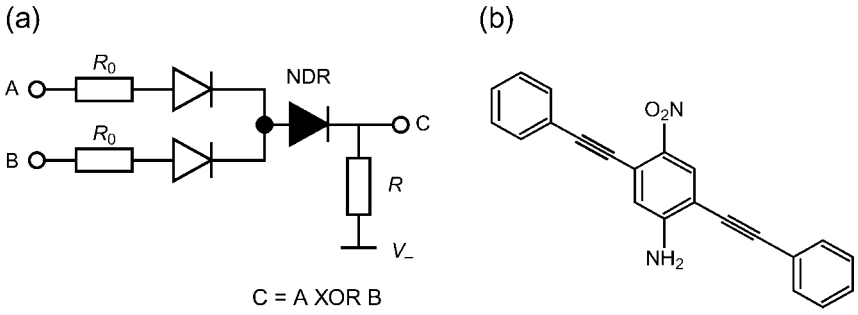


Figure 8.15 (a) XOR gate using NDR-diodes. (b) A well-known molecule that shows NDR.

These two characteristics, called "load lines", are included in Figure 8.14a. The crossing points of the NDR characteristic with these load lines are the operating points for only one input high (point A), or for both inputs high (point B). It transpires that, if both inputs are high, the NDR diode is forced into its valley region, and therefore only a low current flows and only a low voltage drops across *R*. Thereby, the output signal C goes low. From this XOR gate it is easy to obtain inversion; the only change to be made is to set one input (e.g. A) fixed to "1". The output C is then simply the negation of B.

8.4.4
Hysteretic switches

As noted above, hysteretic switches can be used to build PLAs and to yield signal restoration and inversion. The general current versus voltage characteristic of a hysteretic switch is shown in Figure 8.16a.

A hysteretic switch displays two conduction states: one insulating, and one conducting. It is possible to toggle between the two by applying a voltage which

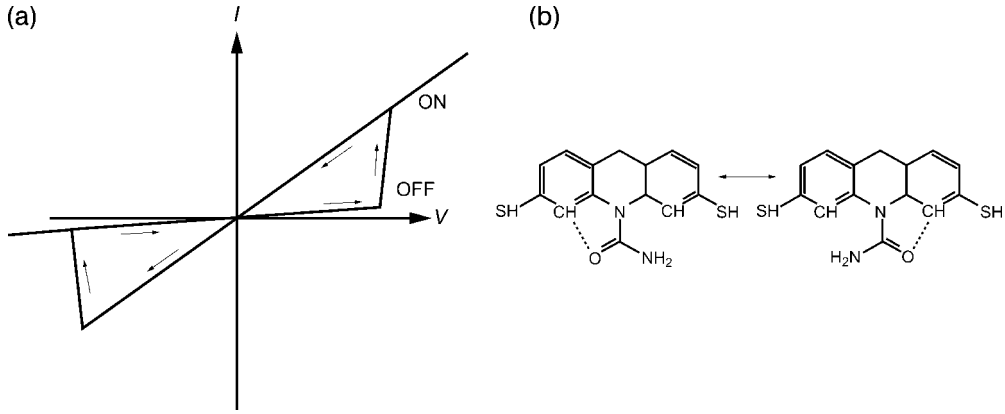


Figure 8.16 (a) General current versus voltage characteristic of a hysteretic switch. (b) A proposed molecule that would be expected to show switching effects.

exceeds a certain threshold value. For example, in Figure 8.16 a positive voltage is needed to switch the device on (i.e. from the insulating to the conducting state), and a negative voltage to switch it off.

Such bistability can be obtained when the molecule possesses two different states that are almost equal in energy, that are separated by an energy barrier, and that show different conduction behaviors. Different origins of these two states are conceivable [99]; for example, they may result from redox processes, from a change in configuration of the molecule, a change in conformation, or a change in electronic excitation.

The molecule shown in Figure 8.16b is an example of a molecule which has been proposed, in theory, to show hysteretic switching [115]. It consists of a fixed molecular backbone (the *stator*) and a side group with a high dipole moment (the *rotor*). By the application of an electric field, the rotor orients its dipole moment in the direction of the field. Bistability is obtained by the formation of hydrogen bonds between the stator and rotor that fix the latter in one of two stable positions relatively to the stator. The two conduction states are due to different conformations of the molecule (stabilized by hydrogen bonds). Switching is initiated by interaction of the dipole moment of the rotor with the electric field.

Bipyridyl-dinitro oligophenylene-ethynylene dithiols (BPDN-DT) are other examples of switching molecules. These are a variation of the molecule shown in Figure 8.15b, and their bistability has recently been confirmed by using various measurement techniques [28, 93].

Rotaxanes and catenanes are, even if controversially discussed, additional candidates for molecular switches. These molecules consist of two interlocked rings such that, by reducing or oxidizing the molecule, one ring rotates within the other. Two stable, neutral states which differ in the position of the inner ring are thus realized. The switching is therefore initiated by a redox process. The two different states are provided by the different conformations of the molecule, similar to the molecule shown in Figure 8.16b.

Although the preliminary results showing a switching effect of this molecule were questioned (see Section 8.4.6), recent results have confirmed the original proposed switching mechanism and indicates that, in the case of earlier results, this mechanism was occasionally hidden by artifacts [75, 116, 117].

8.4.4.1 **The Crossbar Latch: Signal Restoration and Inversion**

As noted in Section 8.4.2.4, AND and OR gates can be built using simple diodes. However, for a complete set of Boolean variables, negation is also required. Signal inversion can be obtained by NDR diodes (c.f. Section 8.4.3.1) or alternatively by the so-called crossbar latch, which also provides a means of signal restoration [118].

The crossbar latch (see Figure 8.17a) consists of two hysteretic switches which are connected to the signal line (at voltage V_L) and one control line (at voltage V_{CA} or V_{CB}). The two switches are oppositely oriented.

The idealized current versus voltage characteristic of a hysteretic switch shown in Figure 8.16 is assumed. By application of a positive voltage, the switch opens; an opposite voltage is then needed to close the switch. These voltages are always those

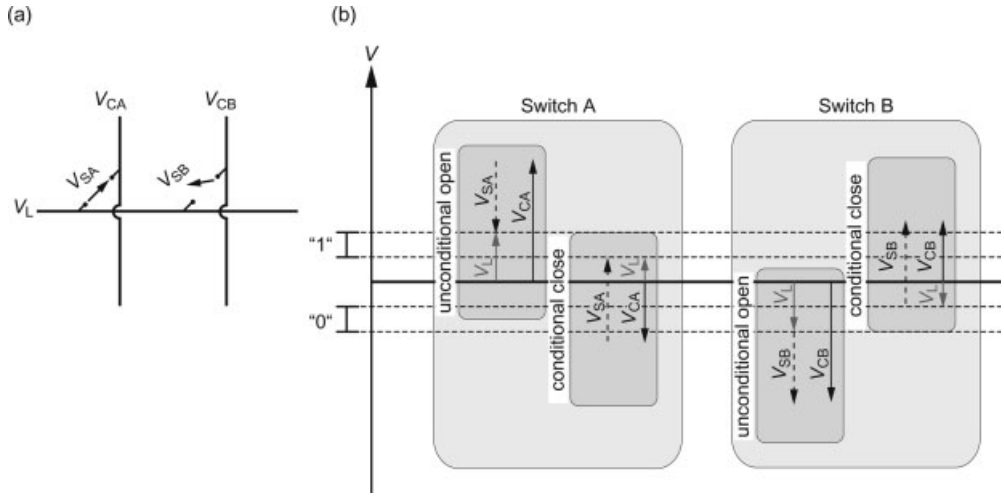


Figure 8.17 The crossbar latch as proposed by Kuekes *et al.* [118].

that are applied *across* the molecules, depicted in Figure 8.17a as V_{SA} and V_{SB}. However, in the circuit shown in Figure 8.17a, only the voltages of the control lines V_{CA} and V_{CB} are set. Therefore, the voltage that is applied *across* the molecule depends on the voltage on the signal line, V_L.

$$V_{SA} = V_L - V_{CA} \tag{8.10}$$

$$V_{SB} = -V_L + V_{CB} \quad \text{(note the opposite orientation of switch B)} \tag{8.11}$$

The voltage on the signal line in Figure 8.17 represents the logical state. Voltage intervals are defined that represent a logical "1" or "0". In general, the signal is degraded, so that the signal level will be at the lower end of the intervals.

To yield signal restoration – that is, to pull the signal level up to the upper end of the defined interval – the following procedure can be followed:

- First, a large positive voltage is applied to the control line A, and a large negative voltage to control line B. This voltage is large enough to always open switches A and B, regardless of the voltage level of the signal line. This is shown in Figure 8.17b (unconditional open).
- Second, a small negative voltage is applied to control line A, and a small positive voltage to control line B. These voltages are so small that they close switch A only if the signal line carries a "1", and they close switch B if there is a "0" on the signal line (see conditional close in Figure 8.17b).
- Therefore, depending on the voltage on the signal line, switch A or B is closed and the opposite switch is open. To yield signal restoration, V_{CA} is connected to a full "1" signal and V_{CB} to "0". Inversion can also be obtained; the only modification is that a logical "1" must be connected to V_{CB} and "0" to V_{CA}.

This scheme shows that it is possible to build a complete set of Boolean variables with only two terminal devices. Therefore, hysteretic switches represent a very valuable element, which explains the high activity in this field of research.

8.4.5
Single-Molecule Single-Electron Transistors

In Section 8.3.7 it was shown how a third electrode – the gate – could be included into the device set-up. The most convenient method to prepare such three-terminal devices is a nanogap that is deposited onto a gate, which is isolated from the device by a thin insulator. A three-terminal device essentially forms a transistor. If the molecule is only weakly coupled to the source and drain, the transistor is termed a *single-molecule single-electron transistor* [119–121]. Whilst the basic working principles of such a transistor were described in Section 8.3.7, and a more extensive explanation is provided in Figure 8.18.

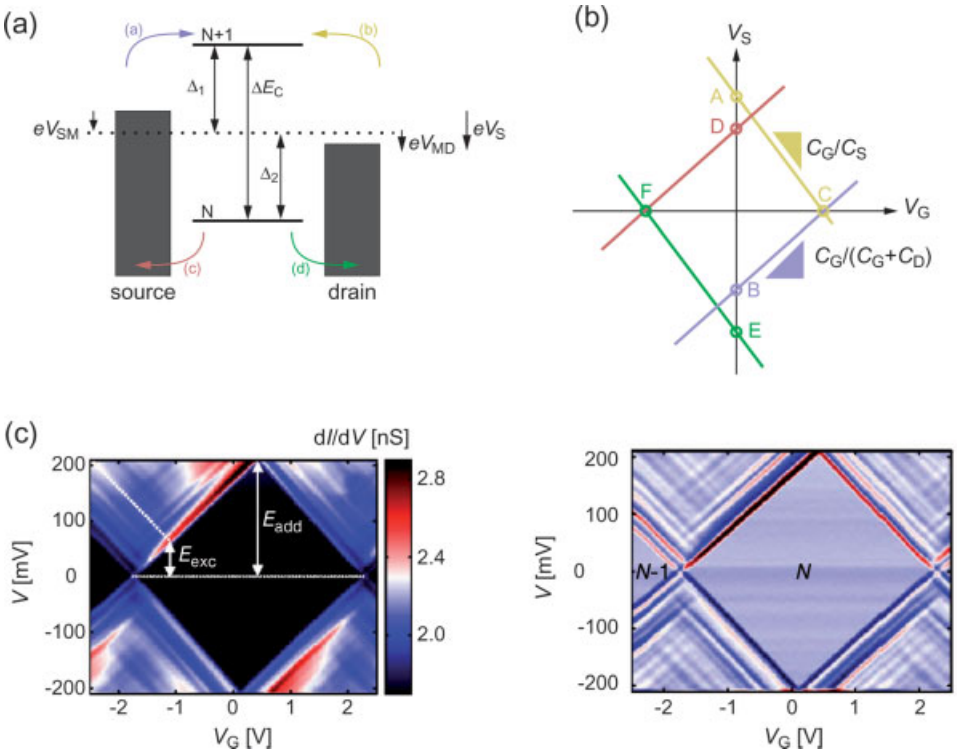


Figure 8.18 (a, b) The working principles of a single-molecule single-electron transistor. For details, see the text. (c) First (upper panel) and second (lower panel) derivatives of the current versus voltage characteristic of a single-molecule single-electron transistor containing an oligophenylenevinylene (OPV5) molecule. The black coulomb blockade diamond is clearly visible. (From Ref. [122].) The fine structure in the open state is due to vibrations of the molecule, and serves as a "fingerprint" of the molecular structure of the OPV5 molecule.

In Figure 8.18a, the N^{th} and $(N + 1)^{th}$ state of the molecule are shown. These two levels correspond, for example, to the HOMO and LUMO in Figure 8.2b. In Figure 8.18a, electrons can hop or tunnel onto or off the molecule by the processes (a) to (d) – that is, from the source or drain onto the molecule (processes a and b), or from the molecule to the source or drain (processes c and d).

Electrons can only flow from the source or the drain onto the molecule if the potential of the electrode aligns with (or exceeds) the molecular level. Below these potentials, no electrons can flow; rather, the current is blocked, which is known as coulomb blockade. This behavior is often visualized in a plot as in Figure 8.18b, in which the source voltage V_S is plotted against the gate voltage V_G.

In Figure 8.18b, four lines build a diamond, the color of these lines corresponding to processes (a) to (d) in Figure 8.18a. In the interior of the diamond, the source and gate voltages are too low to overcome the injection barriers, and the current is blocked. By increasing the source and gate voltage, the working point of the transistor can be moved to outside the diamond. As the working point crosses one line in the V_S/V_G plane, the current sets in (e.g. if it crosses the dark yellow line, electrons can hop from the drain onto the molecule).

Depending on the process, there are different barriers that must be surmounted by the electron. For electrons hopping from source onto the $(N + 1)^{th}$ level (process a), the voltage between source and molecule V_{SM} must exceed the barrier Δ_1 (see Figure 8.18 for a definition of Δ_1) – that is, $eV_{SM} \geq \Delta_1$.

V_{SM} and V_{MD} are governed by Eqs. (8.4) and (8.5), respectively. By combining Eqs. (8.4) and (8.5) with the conditions for current flow (e.g. $eV_{SM} \geq \Delta_1$ for process a), linear relationships are yielded between V_S and V_G that represent the equations of the four lines in Figure 8.18b.

Single-molecule single-electron transistors resemble in one important aspect conventional MOSFETs. The resistance between source and drain can be controlled by the gate voltage – that is, these transistors represent electronic switches and can be used to build logical circuits. As an example, an inverter based on a single-electron transistor is shown in Figure 8.19. In fact, logic circuits consisting of conventional single-electron transistors have already been presented [121, 123]. Single-electron transistors consisting of single molecules have also been realized

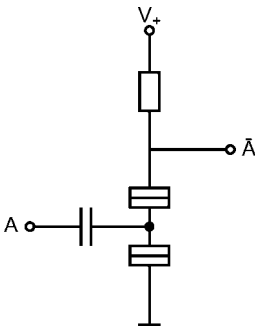


Figure 8.19 Inverting gate using a single-electron transistor.

[53, 55, 57, 59, 60, 82, 122], and an inverter consisting of a multiwall carbon nanotube has also been built [124]. For further information on single-electron devices, the reader is referred to Chapter 2 in Volume I of this Handbook, and to Chapter 6 in the present volume.

8.4.6 Artifacts in Molecular Electronic Devices

As noted above, molecules can provide a rich variety of functions. However, contacting single molecules reaches the limits (or even extends the limits) of current technology, and only recently has it been reported that some results in the field of molecular electronics were due to artifacts [6]. The most prominent example of this was the rotaxanes (see Section 8.4.4). In fact, it has been reported that the observed switching effect is independent of the molecule, and is thus an effect of the entire set-up, including the contacts and interfaces, and is not purely "molecular" [125]. However, this does not necessarily mean that the proposed switching mechanism is incorrect. It is rather covered by artifacts and may in effect exist in other experimental set-ups [75, 116, 126].

8.4.6.1 Sources of Artifacts

Several sources of artifacts are conceivable. One is concerned with the high electric fields that are applied across the molecular junction. Although only low voltages are applied (~1–2 V), the junctions are very thin, which generates high field strengths. These field strengths can cause metal atoms from the electrodes to migrate into the molecular layer and finally to shorten it by a metal filament [6, 128]. These filaments are thin, so that they can be easily broken by high currents flowing through them, for example due to resistive heating or to electromigration. Thus, the filaments can form and break and therefore, they can switch a device into a low- and high-impedance state, which can mitigate molecular switching.

One recent experiment concerning the formation of filaments was conducted by Lau *et al.* [127]. A Langmuir–Blodgett film of steric acid ($C_{18}H_{36}OH$) was sandwiched between a titanium and platinum electrode (see Figure 8.20). It was possible to switch this device between a high and a low conduction state, and a current versus voltage characteristic similar to that shown in Figure 8.16a was obtained. To examine the switching effect, Lau and colleagues scanned the device area using AFM, while simultaneously applying a bias voltage through external leads. In that way, the conductance of the whole device could be measured and correlated to the actual position of the AFM tip. Two-dimensional maps of conductance versus tip position were obtained for the on and off states (see Figure 8.20b).

The AFM tip exerts a certain pressure on the top electrode, which locally compresses the monolayer by ~0.2 Å. This compression did not alter the conductance in the off state. Independently of the position of the AFM tip, the conductance stayed almost constant (see bottom image in Figure 8.20b). However, in the on state, sharp peaks in conductance were observed, which appeared when the tip was scanning across a localized spot on the top electrode (see red spike in the middle

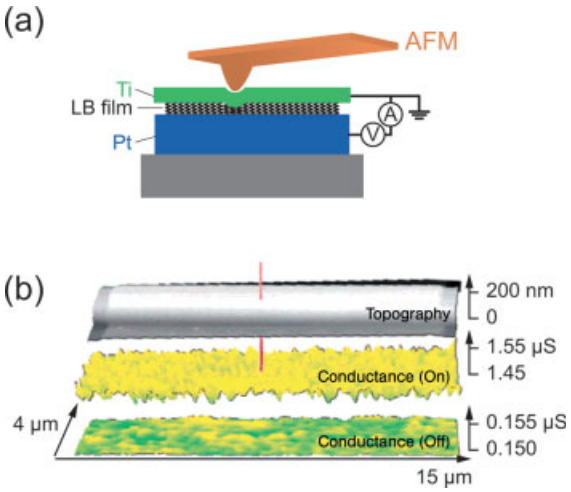


Figure 8.20 An experiment conducted by Lau *et al.* to examine the switching effect observed in their devices. Combining current versus voltage and atomic force measurements, these authors observed nanoscale switching centers, which are interpreted in terms of conductive filaments that almost bridge the two electrodes. (From Ref. [127].)

image in Figure 8.20b). These peaks were interpreted as "nanoasperities", in which the effective distance between the top and bottom electrode is reduced. These nanoasperities might be due to conducting filaments that almost bridge the electrodes.

Besides filaments, there are other sources of artifacts, such as oxidation of the metal electrode (i.e. titanium), which can also induce hysteretic current–voltage responses [129] or the formation of charge traps at the electrode/molecule interface [130]. Similarly, nanogaps produced by the electromigration technique (see Section 8.3.5) can, even in the absence of molecules, show "molecular" features, for example coulomb blockade effects with addition energies that are in the range as would expected for single molecules [82]. These effects are ascribed to small metallic grains within the junction.

To rule out these artifacts and to verify that the observed effect is truly molecular, several strategies have been followed [126]. One straightforward approach is systematically to vary the composition of the molecule (e.g. its length) and to study the influence of this variation on the current versus voltage characteristic [96]. Other strategies are to use a variety of test set-ups to rule out any systematic errors due to the measurement set-up [93], to completely avoid metallic electrodes and thereby eliminate the possibility of metallic filaments [131].

An alternative approach is to study the vibrational states of the molecule by using *inelastic electron tunneling spectroscopy* (IETS). This technique is highly sensitive to the molecular vibrations which open additional inelastic channels through which electrons can tunnel. However, these vibration states smear out in energy at room

temperature, and can only be observed at cryogenic temperatures. In the current versus voltage characteristic, they are visible as peaks in the second derivative [46, 132]. For three-terminal devices, the second derivative can be plotted in the source voltage (V_S) versus gate voltage (V_G) plane, and the vibrational modes are then visible as lines running parallel to the diamond edges (cf. Figure 8.18c) [122].

The vibrational states are an intrinsic property of the molecule, and thus IETS provides a "molecular fingerprint" to prove that the molecule is actively involved in current transport.

8.4.7 Conclusions

Molecular electronic devices provide a wealth of functions. The appeal of molecular electronics is, amongst other things, based on the variety of these functions. The target is to make use of new effects that appear at these small dimensions (e.g. coulomb blockade or resonant tunneling effects, conformational changes of the molecule) for novel electronic devices.

Whilst it has been shown how these molecular devices can be combined to form small logical gates (e.g. in the form of diode–diode logic or the crossbar latch), this raises one important question: How can molecules be assembled so that these gates are formed?

8.5 Building Logical Circuits: Assembly of a Large Number of Molecular Devices

In the previous sections it has been shown how single (or at least a low number of) molecules can be contacted, and which functionalities these molecules can provide. Furthermore, it has been described, how these single molecules can be combined to small logic gates, for example as AND, OR, XOR gates, or as the crossbar latch. In this section, the discussion proceeds one step further to determine how a large number of devices can be assembled. And what implications does the use of single molecules have for the architecture of future logic circuits? As already discussed (in Section 8.3), for single-molecule devices there is at present no method available to deterministically place a single molecule on a chip. Thus, reliance must be placed on statistical processes and the ability of the molecules to self-organize, for example in the form of SAMs.

This dependence on self-organization bears the first implication for the architecture of molecular devices. Self-organization will always result in very regular structures, which have only a low information content [133]. In comparison to current CMOS circuits, in which a huge number of transistors is connected statically to other transistors, and which include therefore a high information content, this lack of information must be fed into the molecular circuit by an additional, post-fabrication training step. In other words, the technology and architecture has to be re-configurable.

A second implication of the self-organization process is given by the fact that these circuits will always contain defective parts, and the high yields necessary for CMOS architectures are not feasible. Again, the defective sites on the molecular chip must be identified and isolated in a post-fabrication step.

Several architectures have been proposed for future molecular circuits, and these differ in how strongly they rely on self-organization. PLAs based on crossbars, for example, use self-organization only for the preparation of the SAMs. The electrodes that contact the SAM are commonly defined by lithography (although techniques are available to prepare crossbars completely by self-organization, e.g. [102]). In contrast to PLAs, the *Nano-Cell* architecture (see Section 8.5.2) relies completely on self-organization.

8.5.1
Programmable Logic Arrays Based on Crossbars

In Section 8.4.2.4 it was described, how simple AND and OR gates can be implemented using diodes and resistances only. In the following it will be shown, how large arrays of these gates can be implemented using crossbars, as described in Section 8.3.6.

The equivalent circuit of a crossbar is shown in Figure 8.21a and b. A SAM of rectifying diodes is contacted by orthogonal top and bottom electrodes. These arrays can easily be converted into AND and OR gates. For an AND circuit, the vertical electrodes must be connected to the high voltage, and the horizontal lines to the input variables (see Figure 8.21a). Similarly, for OR, the horizontal lines are connected to the low voltage and the vertical lines are the signal lines.

However, these circuits have one drawback, in that only a single disjunction or conjunction can be implemented. If, for example, a simple AND connection of two input variables is built, all other horizontal input lines must be set to the high voltage. Therefore, at the end of each vertical line, only the conjunction of A and B is computed.

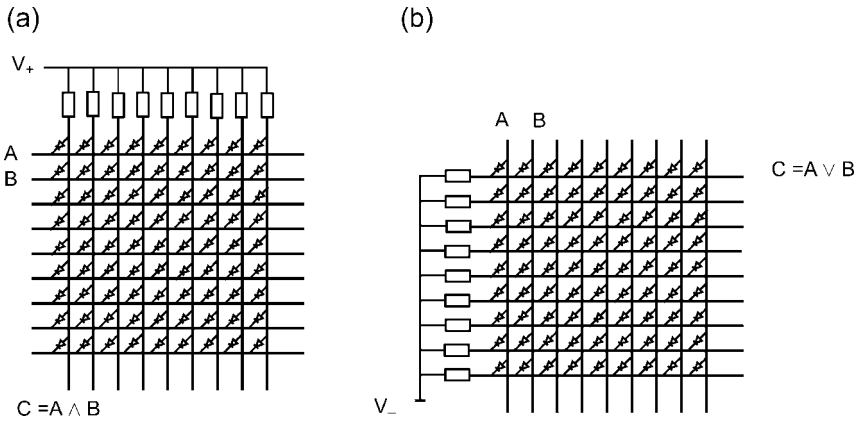


Figure 8.21 Crossbar structure implementing AND (a) and OR (b).

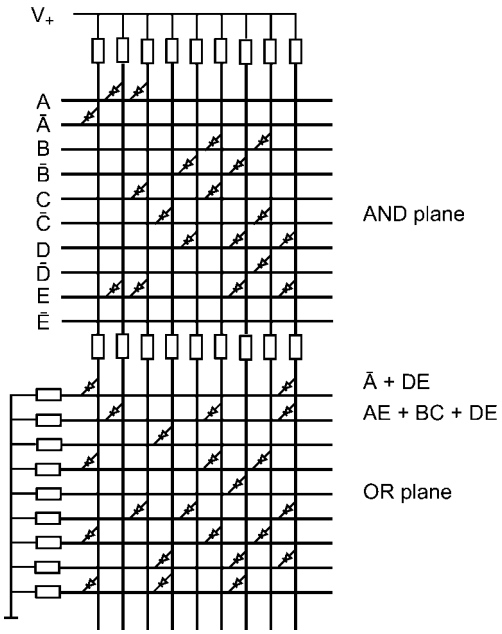


Figure 8.22 A programmable logic array consisting of an AND and an OR plane.

This problem can be circumvented if some diodes can be switched off – that is, if some input lines can be isolated from the output lines. Thus, a combination of a hysteretic switch (as described in Section 8.4.4) and a diode, for example by asymmetrically coupling a bistable molecule to the electrodes, would be highly beneficial. By using these switches, individual crossing points could be switched "off" and "on" by the application of a high-voltage pulse. The state of the molecules at the crossing points will therefore determine the logical function that is computed.

The AND and the OR gate can be combined to a PLA, as shown in Figure 8.22. The output of the AND plane is fed into the OR plane. Most diodes at the crossing points are switched off, so that certain Boolean functions are realized. The Boolean functions of the first two horizontal lines in the OR plane are given in Figure 8.22.

Such a PLA can compute every Boolean function if not only the input variables but also the negation of them is supplied to the PLA. Therefore, the negation of each variable must be computed, which can be done by using NDR-diodes as described in Section 8.4.3.1. Again, these NDR diodes can be implemented in a crossbar, so that all negations of all input signals can be realized simultaneously. As an alternative, the negation can be supplied by the surrounding CMOS circuitry, which would result in hybrid molecular/CMOS circuits.

Based on hysteretic switching diodes, the PLA is re-configurable. Therefore, the logic functions are programmed in a post-fabrication step; defective junctions can be

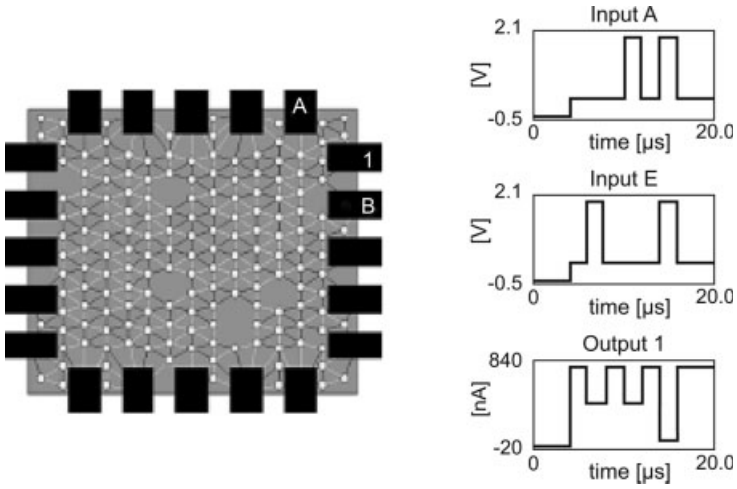


Figure 8.23 NanoCell trained as NAND gate, as proposed by Tour *et al.* (From Ref. [134].)

identified and disregarded in the circuit. This makes the PLA architecture a promising architecture for molecular electronics. A more detailed explanation of architectures based on crossbars is provided in Chapter 11 of this volume.

8.5.2 NanoCell

Although the crossbar set-up relies on self-organization, a "quasi-regular" structure is imposed by the orthogonal top and bottom electrode [133]. An architecture, which consists completely of random patterns, is the so-called NanoCell.

The structure of a NanoCell is shown in Figure 8.23 [134, 135]. It consists of a self-assembled, two-dimensional network of metallic particles which are randomly interlinked with molecules (cf. Section 8.3.8). In order to provide inverting logic, molecules exhibiting an NDR effect are used (see Section 8.4.3.1). The network is contacted by large metallic leads at its sides.

Due to the random arrangement of metal particles and molecules, the NanoCell must be trained or programmed to fulfill a certain task. To train this circuit, the molecules must be bistable and, similar to the PLAs, the molecules can be switched off and on by large voltage pulses. Therefore, the molecules must exhibit a combination of NDR effect and hysteretic switching. In Figure 8.23, the molecules are represented by lines connecting two metal particles; a white line represents an open switch, and a black line a closed switch.

In real applications, the molecular network is completely random – that is, neither are the positions of the individual molecules known, nor can a certain single molecule be addressed. The only knowledge about the circuit can be obtained through the contact pins, and most probably only bundles of molecules and not individual molecules can be switched by the application of voltage pulses to the contacts.

However, for a proof of concept, Tour *et al*. assumed the case of omnipotent programming [134], which means that the position of each molecule is known and that it can be individually programmed to its low or high state.

Based on this assumption, the network can be trained for a certain task; for example, to perform a NAND operation, as shown in Figure 8.23. The state of the network can be described by a list of the switching states of all molecules, typically, which molecule is open or which is closed. A function can be defined, which evaluates by how much the output ("1" in Figure 8.23) resembles a NAND combination of the inputs A and B. The task of training is now reduced to find a network state, which sufficiently minimizes (or, depending on the definition, maximizes) the evaluation function.

Tour and coworkers have used a genetic algorithm to identify such a network state – that is, to determine which molecules must be switched on, and which off. The output signal is determined by a SPICE simulation and compared to the desired functionality. Using this genetic algorithm, Tour and colleagues were able to train NanoCells as inverters, NAND gates, or complete 1-bit adders [134].

8.6 Challenges and Perspectives

Molecular electronics represents an exciting and promising field of research, but it imposes huge demands on the technology and is at the border of what is currently feasible. Many challenges remain for further research, as well in the design of molecules and in the assembly and architecture of future devices. For example, molecular interconnects must be found that can conduct current across larger distances, the rectification ratio of molecular diodes must be increased and, as a key element, bistable switching molecules must be identified and optimized. Future architectures of molecular devices will have to incorporate the statistical nature of the assembly of molecules. The optimum architecture will start from a random arrangement of molecules and will be trained the molecular network, as has been attempted with the NanoCell set-up. However, this training step is mathematically highly complex and has only been solved for the simplifying case of omnipotent programming. It seems that there is a price to pay for the low information content of these random structures with a complex training algorithm.

In CMOS, molecular electronics has a very strong competitor. CMOS has been so successful in the past, because it combines high integration densities, a high switching speed, and a low power consumption [133]. If molecular electronics is to at least complement CMOS in the future, it must outperform CMOS in one or more of these key characteristics. It is proposed that molecular electronics will be small in size and will exhibit low switching energies [1, 133], but the switching speed will be low. However, these predictions are very uncertain and will need to be substantiated by further research.

References

1 International Technology Roadmap for Semiconductors; 2005 Edition. public. itrs.net. 2006.

2 A. Aviram, M. Ratner, *Chem. Phys. Lett.* 1974, **29**, 277–283.

3 C. P. Collier, G. Mattersteig, E. W. Wong, Y. Luo, K. Beverly, J. Sampaio, F. M. Raymo, J. F. Stoddart, J. R. Heath, *Science* 2000, **289**, 1172–1175.

4 J. Chen, M. A. Reed, A. M. Rawlett, J. M. Tour, *Science* 1999, **286**, 1550–1552.

5 R. F. Service, *Science* 2001, **294**, 2442–2443.

6 R. F. Service, *Science* 2003, **302**, 556.

7 S. Datta, *Nanotechnology* 2004, **15**, S433–S451.

8 S. Datta, *Quantum Transport – Atom to Transistor*, Cambridge University Press, 2005.

9 M. Zharnikov, S. Frey, H. Rong, Y. J. Yang, K. Heister, M. Buck, M. Grunze, *Phys. Chem. Chem. Phys.* 2000, **2**, 3359–3362.

10 J. C. Love, D. B. Wolfe, R. Haasch, M. L. Chabinyc, K. E. Paul, G. M. Whitesides, R. G. Nuzzo, *J. Am. Chem. Soc.* 2003, **125**, 2597–2609.

11 L. Patrone, S. Palacin, J. P. Bourgoin, J. Lagoute, T. Zambelli, S. Gauthier, *Chem. Phys.* 2002, **281**, 325–332.

12 P. Morf, F. Raimondi, H. G. Nothofer, B. Schnyder, A. Yasuda, J. M. Wessels, T. A. Jung, *Langmuir* 2006, **22**, 658–663.

13 J. M. Wessels, H. G. Nothofer, W. E. Ford, F. von Wrochem, F. Scholz, T. Vossmeyer, A. Schroedter, H. Weller, A. Yasuda, *J. Am. Chem. Soc.* 2004, **126**, 3349–3356.

14 J. Chen, L. C. Calvet, M. A. Reed, D. W. Carr, D. S. Grubisha, D. W. Bennett, *Chem. Phys. Lett.* 1999, **313**, 741–748.

15 A. Ulman, *Chem. Rev.* 1996, **96**, 1533–1554.

16 S. Ranganathan, I. Steidel, F. Anariba, R. L. McCreery, *Nano Lett.* 2001, **1**, 491–494.

17 M. R. Kosuri, H. Gerung, Q. M. Li, S. M. Han, P. E. Herrera-Morales, J. F. Weaver, *Surface Sci.* 2005, **596**, 21–38.

18 J. Moreland, J. W. Ekin, *J. Appl. Physics* 1985, **58**, 3888–3895.

19 M. A. Reed, C. Zhou, C. J. Muller, T. P. Burgin, J. M. Tour, *Science* 1997, **278**, 252–254.

20 J. Reichert, R. Ochs, D. Beckmann, H. B. Weber, M. Mayor, H. von Lohneysen, *Phys. Rev. Lett.* 2002 88.

21 H. B. Weber, J. Reichert, F. Weigend, R. Ochs, D. Beckmann, M. Mayor, R. Ahlrichs, H. von Lohneysen, *Chem. Phys.* 2002, **281**, 113–125.

22 H. B. Weber, J. Reichert, R. Ochs, D. Beckmann, M. Mayor, H. von Lohneysen, *Physica E: Low-Dimensional Systems Nanostructures* 2003, **18**, 231–232.

23 J. Reichert, H. B. Weber, M. Mayor, H. von Lohneysen, *Appl. Phys. Lett.* 2003, **82**, 4137–4139.

24 M. Mayor, C. von Hanisch, H. B. Weber, J. Reichert, D. Beckmann, *Angew. Chem. Int. Ed.* 2002, **41**, 1183–1186.

25 M. Mayor, H. B. Weber, J. Reichert, M. Elbing, C. von Hanisch, D. Beckmann, M. Fischer, *Angew. Chem. Int. Ed.* 2003, **42**, 5834–5838.

26 M. Elbing, R. Ochs, M. Koentopp, M. Fischer, C. von Hanisch, F. Weigend, F. Evers, H. B. Weber, M. Mayor, *Proc. Natl. Acad. Sci. USA* 2005, **102**, 8815–8820.

27 D. Dulic, S. J. van der Molen, T. Kudernac, H. T. Jonkman, J. J. D. de Jong, T. N. Bowden, J. van Esch, B. L. Feringa, B. J. van Wees, *Phys. Rev. Lett.* 2003 91.

28 E. Lortscher, J. W. Ciszek, J. Tour, H. Riel, *Small* 2006, **2**, 973–977.

29 R. H. M. Smit, Y. Noat, C. Untiedt, N. D. Lang, M. C. van Hemert, J. M. van Ruitenbeek, *Nature* 2002, **419**, 906–909.

30 R. Waser (Ed.), *Nanoelectronics and Information Technology*, 2nd edition. Wiley-VCH, 2005.

31 D. J. Wold, C. D. Frisbie, *J. Am. Chem. Soc.* 2001, **123**, 5549–5556.

32 D. J. Wold, C. D. Frisbie, *J. Am. Chem. Soc.* 2000, **122**, 2970–2971.

33 D. J. Wold, R. Haag, M. A. Rampi, C. D. Frisbie, *J. Phys. Chem. B* 2002, **106**, 2813–2816.

34 A. M. Rawlett, T. J. Hopson, L. A. Nagahara, R. K. Tsui, G. K. Ramachandran, S. M. Lindsay, *Appl. Phys. Lett.* 2002, **81**, 3043–3045.

35 F. Schreiber, *Prog. Surface Sci.* 2000, **65**, 151–256.

36 J. M. Beebe, V. B. Engelkes, L. L. Miller, C. D. Frisbie, *J. Am. Chem. Soc.* 2002, **124**, 11268–11269.

37 K. Moth-Poulsen, L. Patrone, N. Stuhr-Hansen, J. B. Christensen, J. P. Bourgoin, T. Bjornholm, *Nano Lett.* 2005, **5**, 783–785.

38 X. L. Li, B. Q. Xu, X. Y. Xiao, X. M. Yang, L. Zang, N. J. Tao, *Faraday Disc.* 2006, **131**, 111–120.

39 B. Q. Xu, X. Y. Xiao, N. J. Tao, *J. Am. Chem. Soc.* 2003, **125**, 16164–16165.

40 D. I. Gittins, D. Bethell, D. J. Schiffrin, R. J. Nichols, *Nature* 2000, **408**, 67–69.

41 Y. Yasutake, Z. J. Shi, T. Okazaki, H. Shinohara, Y. Majima, *Nano Lett.* 2005, **5**, 1057–1060.

42 L. A. Bumm, J. J. Arnold, M. T. Cygan, T. D. Dunbar, T. P. Burgin, L. Jones, D. L. Allara, J. M. Tour, P. S. Weiss, *Science* 1996, **271**, 1705–1707.

43 B. Lussem, L. Muller-Meskamp, S. Karthauser, R. Waser, M. Homberger, U. Simon, *Langmuir* 2006, **22**, 3021–3027.

44 A. S. Blum, J. G. Kushmerick, S. K. Pollack, J. C. Yang, M. Moore, J. Naciri, R. Shashidhar, B. R. Ratna, *J. Phys. Chem. B* 2004, **108**, 18124–18128.

45 J. G. Kushmerick, J. Naciri, J. C. Yang, R. Shashidhar, *Nano Lett.* 2003, **3**, 897–900.

46 J. G. Kushmerick, J. Lazorcik, C. H. Patterson, R. Shashidhar, D. S. Seferos, G. C. Bazan, *Nano Lett.* 2004, **4**, 639–642.

47 J. G. Kushmerick, D. B. Holt, J. C. Yang, J. Naciri, M. H. Moore, R. Shashidhar, *Phys. Rev. Lett.* 2002, 89.

48 H. Park, A. K. L. Lim, A. P. Alivisatos, J. Park, P. L. Mceuen, *Appl. Phys. Lett.* 1999, **75**, 301–303.

49 S. Kronholz, S. Karthauser, A. van der Hart, T. Wandlowski, R. Waser *Microelectronics J.* 2006, **37**, 591–594.

50 S. Kronholz, S. Karthauser, G. Meszaros, T. Wandlowski, A. van der Hart, R. Waser, *Microelectronic Eng.* 2006, **83**, 1702–1705.

51 C. Z. Li, H. X. He, N. J. Tao, *Appl. Phys. Lett.* 2000, **77**, 3995–3997.

52 S. Boussaad, N. J. Tao, *Appl. Phys. Lett.* 2002, **80**, 2398–2400.

53 S. Kubatkin, A. Danilov, M. Hjort, J. Cornil, J. L. Bredas, N. Stuhr-Hansen, P. Hedegard, T. Bjornholm, *Nature* 2003, **425**, 698–701.

54 E. Ruttkowski, R. J. Luyken, Y. Mustafa, M. Specht, F. Hofmann, M. Städele, W. Rösner, W. Weber, R. Waser, L. Risch, *Proceedings, 2005 5th IEEE Conference on Nanotechnology*, Volume 1, pp. 438–441, 2005.

55 J. Park, A. N. Pasupathy, J. I. Goldsmith, C. Chang, Y. Yaish, J. R. Petta, M. Rinkoski, J. P. Sethna, H. D. Abruna, P. L. Mceuen, D. C. Ralph, *Nature* 2002, **417**, 722–725.

56 W. J. Liang, M. P. Shores, M. Bockrath, J. R. Long, H. Park, *Nature* 2002, **417**, 725–729.

57 H. Park, J. Park, A. K. L. Lim, E. H. Anderson, A. P. Alivisatos, P. L. Mceuen, *Nature* 2000, **407**, 57–60.

58 A. V. Danilov, S. E. Kubatkin, S. G. Kafanov, T. Bjornholm, *Faraday Disc.* 2006, **131**, 337–345.

59 A. N. Pasupathy, J. Park, C. Chang, A. V. Soldatov, S. Lebedkin, R. C. Bialczak, J. E. Grose, L. A. K. Donev, J. P. Sethna, D. C. Ralph, P. L. Mceuen, *Nano Lett.* 2005, **5**, 203–207.

60 S. Kubatkin, A. Danilov, M. Hjort, J. Cornil, J. L. Bredas, N. Stuhr-Hansen, P. Hedegard, T. Bjornholm, *Curr. Appl. Physics* 2004, **4**, 554–558.

61 M. J. Tarlov, *Langmuir* 1992, **8**, 80–89.

62 G. C. Herdt, D. R. Jung, A. W. Czanderna, *Prog. Surface Sci.* 1995, **50**, 103–129.

63 T. Ohgi, H. Y. Sheng, H. Nejoh, *Appl. Surface Sci.* 1998, **132**, 919–924.

64 G. L. Fisher, A. E. Hooper, R. L. Opila, D. L. Allara, N. Winograd, *J. Phys. Chem. B* 2000, **104**, 3267–3273.

65 G. L. Fisher, A. V. Walker, A. E. Hooper, T. B. Tighe, K. B. Bahnck, H. T. Skriba, M. D. Reinard, B. C. Haynie, R. L. Opila, N. Winograd, D. L. Allara, *J. Am. Chem. Soc.* 2002, **124**, 5528–5541.

66 A. V. Walker, T. B. Tighe, B. C. Haynie, S. Uppili, N. Winograd, D. L. Allara, *J. Phys. Chem. B* 2005, **109**, 11263–11272.

67 A. V. Walker, T. B. Tighe, O. M. Cabarcos, M. D. Reinard, B. C. Haynie, S. Uppili, N. Winograd, D. L. Allara, *J. Am. Chem. Soc.* 2004, **126**, 3954–3963.

68 B. de Boer, M. M. Frank, Y. J. Chabal, W. R. Jiang, E. Garfunkel, Z. Bao, *Langmuir* 2004, **20**, 1539–1542.

69 T. B. Tighe, T. A. Daniel, Z. H. Zhu, S. Uppili, N. Winograd, D. L. Allara, *J. Phys. Chem. B* 2005, **109**, 21006–21014.

70 Y. Tai, A. Shaporenko, H. Noda, M. Grunze, M. Zharnikov, *Adv. Mater.* 2005, **17**, 1745–1749.

71 Y. Tai, A. Shaporenko, W. Eck, M. Grunze, M. Zharnikov, *Langmuir* 2004, **20**, 7166–7170.

72 H. B. Akkerman, P. W. M. Blom, D. M. de Leeuw, B. de Boer, *Nature* 2006, **441**, 69–72.

73 H. Haick, M. Ambrico, J. Ghabboun, T. Ligonzo, D. Cahen, *Phys. Chem. Chem. Phys.* 2004, **6**, 4538–4541.

74 Y. L. Loo, D. V. Lang, J. A. Rogers, J. W. P. Hsu, *Nano Lett.* 2003, **3**, 913–917.

75 K. T. Shimizu, J. D. Tabbri, J. J. Jelincic, N. A. Melosh, *Adv. Mater.* 2006, **18**, 1499–1504.

76 B. Q. Xu, X. Y. Xiao, X. M. Yang, L. Zang, N. J. Tao, *J. Am. Chem. Soc.* 2005, **127**, 2386–2387.

77 T. Albrecht, K. Moth-Poulsen, J. B. Christensen, A. Guckian, T. Bjornholm, J. G. Vos, J. Ulstrup, *Faraday Disc.* 2006, **131**, 265–279.

78 T. Albrecht, K. Moth-Poulsen, J. B. Christensen, J. Hjelm, T. Bjornholm, J. Ulstrup, *J. Am. Chem. Soc.* 2006, **128**, 6574–6575.

79 W. Haiss, H. van Zalinge, S. J. Higgins, D. Bethell, H. Hobenreich, D. J. Schiffrin, R. J. Nichols, *J. Am. Chem. Soc.* 2003, **125**, 15294–15295.

80 X. Y. Xiao, L. A. Nagahara, A. M. Rawlett, N. J. Tao, *J. Am. Chem. Soc.* 2005, **127**, 9235–9240.

81 A. R. Champagne, A. N. Pasupathy, D. C. Ralph, *Nano Lett.* 2005, **5**, 305–308.

82 H. S. J. van der Zant, Y. V. Kervennic, M. Poot, K. O'Neill, Z. de Groot, J. M. Thijssen, H. B. Heersche, N. Stuhr-Hansen, T. Bjornholm, D. Vanmaekelbergh, C. A. van Walree, L. W. Jenneskens, *Faraday Disc.* 2006, **131**, 347–356.

83 M. Brust, M. Walker, D. Bethell, D. J. Schiffrin, R. Whyman, *J. Chem. Soc. - Chem. Commun.* 1994, 801–802.

84 C. J. Kiely, J. Fink, M. Brust, D. Bethell, D. J. Schiffrin, *Nature* 1998, **396**, 444–446.

85 J. Liao, L. Bernard, M. Langer, C. Schonenberger, M. Calame, *Adv. Mater.* 2006, **18**, 2444–2447.

86 J. M. Tour, L. Cheng, D. P. Nackashi, Y. X. Yao, A. K. Flatt, S. K. St Angelo, T. E. Mallouk, P. D. Franzon, *J. Am. Chem. Soc.* 2003, **125**, 13279–13283.

87 T. Hassenkam, K. Moth-Poulsen, N. Stuhr-Hansen, K. Norgaard, M. S. Kabir, T. Bjornholm, *Nano Lett.* 2004, **4**, 19–22.

88 T. Hassenkam, K. Norgaard, L. Iversen, C. J. Kiely, M. Brust, T. Bjornholm, *Adv. Mater.* 2002, **14**, 1126–1130.

89 K. Norgaard, T. Bjornholm, *Chem. Commun.* 2005, 1812–1823.

90 D. P. Long, C. H. Patterson, M. H. Moore, D. S. Seferos, G. C. Bazan, J. G. Kushmerick, *Appl. Phys. Lett.* 2005, **86**, 153105.

91 T. Dadosh, Y. Gordin, R. Krahne, I. Khivrich, D. Mahalu, V. Frydman, J. Sperling, A. Yacoby, I. Bar-Joseph, *Nature* 2005, **436**, 677–680.

92 R. Haag, M. A. Rampi, R. E. Holmlin, G. M. Whitesides, *J. Am. Chem. Soc.* 1999, **121**, 7895–7906.

93 A. S. Blum, J. G. Kushmerick, D. P. Long, C. H. Patterson, J. C. Yang, J. C.

Henderson, Y. X. Yao, J. M. Tour, R. Shashidhar, B. R. Ratna, *Nature Mater.* 2005, **4**, 167–172.

94 J. Chen, M. A. Reed, *Chem. Phys.* 2002, **281**, 127–145.

95 M. A. Reed, J. Chen, A. M. Rawlett, D. W. Price, J. M. Tour, *Appl. Phys. Lett.* 2001, **78**, 3735–3737.

96 W. Y. Wang, T. Lee, M. A. Reed, *Phys. Rev. B* 2003, 68.

97 C. Zhou, M. R. Deshpande, M. A. Reed, L. Jones, J. M. Tour, *Appl. Phys. Lett.* 1997, **71**, 611–613.

98 S. C. Goldstein, D. Rosewater, Solid-State Circuits Conference 2002. Digest of Technical Papers, ISSCC 2002, Volume 1, p. 204.

99 M. Mayor, H. B. Weber, R. Waser,in: R. Waser (Ed.), *Nanoelectronics and Information Technology* Wiley-VCH Weinheim 2003, pp. 503–525.

100 M. J. Crossley, P. L. Burn, *J. Chem. Soc. - Chem. Commun.* 1991, 1569–1571.

101 R. L. Carroll, C. B. Gorman, *Angew. Chem. Int. Ed.* 2002, **41**, 4379–4400.

102 A. Ismach, E. Joselevich, *Nano Lett.* 2006, **6**, 1706–1710.

103 A. Aviram, M. Ratner, *Chem. Phys. Lett.* 1974, **29**, 277–283.

104 P. E. Kornilovitch, A. M. Bratkovsky, R. S. Williams, *Phys. Rev. B* 2002, **66**, 165436.

105 A. Aviram, M. Ratner, *Chem. Phys. Lett.* 1974, **29**, 277–283.

106 R. M. Metzger, *Chem. Phys.* 2006, **326**, 176–187.

107 G. J. Ashwell, J. R. Sambles, A. S. Martin, W. G. Parker, M. Szablewski, *J. Chem. Soc. - Chem. Commun.* 1990, 1374–1376.

108 A. S. Martin, J. R. Sambles, G. J. Ashwell, *Phys. Rev. Lett.* 1993, **70**, 218–221.

109 C. Krzeminski, C. Delerue, G. Allan, D. Vuillaume, R. M. Metzger, *Phys. Rev. B* 2001, 6408.

110 R. M. Metzger, B. Chen, U. Hopfner, M. V. Lakshmikantham, D. Vuillaume, T. Kawai, X. L. Wu, H. Tachibana, T. V. Hughes, H. Sakurai, J. W. Baldwin, C. Hosch, M. P. Cava, L. Brehmer, G. J. Ashwell, *J. Am. Chem. Soc.* 1997, **119**, 10455–10466.

111 G. J. Ashwell, R. Hamilton, L. R. H. High, *J. Mater. Chem.* 2003, **13**, 1501–1503.

112 J. C. Ellenbogen, J. C. Love, *Proc. IEEE* 2000, **88**, 386–426.

113 M. A. Reed, *Proc. IEEE* 1999, **87**, 652–658.

114 J. Taylor, M. Brandbyge, K. Stokbro, *Phys. Rev. B* 2003, 68.

115 P. E. Kornilovitch, A. M. Bratkovsky, R. S. Williams, *Phys. Rev. B* 2002, **66**, 245413.

116 K. Norgaard, B. W. Laursen, S. Nygaard, K. Kjaer, H.-R. Tseng, A. H. Flood, J. F. Stoddart, T. Bjornholm, *Angew. Chem. - Int. Ed.* 2005, **44**, 7035–7039.

117 J. W. Choi, A. H. Flood, D. W. Steuerman, S. Nygaard, A. B. Braunschweig, N. N. P. Moonen, B. W. Laursen, Y. Luo, E. DeIonno, A. J. Peters, J. O. Jeppesen, K. Xu, J. F. Stoddart, J. R. Heath, *Chemistry - A European Journal* 2005, **12**, 261–279.

118 P. J. Kuekes, D. R. Stewart, R. S. Williams, *J. Appl. Physics* 2005, **98**, 049901.

119 P. Hedegard, T. Bjornholm, *Chem. Phys.* 2005, **319**, 350–359.

120 K. K. Likharev, *Proc. IEEE* 1999, **87**, 606–632.

121 K. Uchida, in: R. Waser (Ed.), *Nanoelectronics and Information Technology*, Wiley-VCH, Weinheim, 2006, pp. 425–443.

122 E. A. Osorio, K. O'Neill, N. Stuhr-Hansen, O. F. Nielsen, T. Bjornholm, H. S. J. van der Zant *Adv. Mater.* 2007, **19**, 281–285.

123 K. Uchida, J. Koga, R. Ohba, A. Toriumi, *IEEE Trans. Electron Devices* 2003, **50**, 1623–1630.

124 K. Ishibashi, D. Tsuya, M. Suzuki, Y. Aoyagi, *Appl. Phys. Lett.* 2003, **82**, 3307–3309.

125 D. R. Stewart, D. A. A. Ohlberg, P. A. Beck, Y. Chen, R. S. Williams, J. O. Jeppesen, K. A. Nielsen, J. F. Stoddart, *Nano Lett.* 2004, **4**, 133–136.

126 A. H. Flood, J. F. Stoddart, D. W. Steuerman, J. R. Heath, *Science* 2004, **306**, 2055–2056.

127 C. N. Lau, D. R. Stewart, R. S. Williams, M. Bockrath, *Nano Lett.* 2004, **4**, 569–572.

128 V. V. Zhirnov, R. K. Cavin, *Nature Mater.* 2006, **5**, 11–12.

129 W. R. McGovern, F. Anariba, R. L. McCreery, *J. Electrochem. Soc.* 2005, **152**, E176–E183.

130 C. A. Richter, D. R. Stewart, D. A. A. Ohlberg, R. S. Williams, *Appl. Physics A - Mater. Sci. Process.* 2005, **80**, 1355–1362.

131 J. L. He, B. Chen, A. K. Flatt, J. J. Stephenson, C. D. Doyle, J. M. Tour, *Nature Mater.* 2006, **5**, 63–68.

132 W. Y. Wang, T. Lee, I. Kretzschmar, M. A. Reed, *Nano Lett.* 2004, **4**, 643–646.

133 M. R. Stan, P. D. Franzon, S. C. Goldstein, J. C. Lach, M. M. Ziegler, *Proc. IEEE* 2003, **91**, 1940–1957.

134 J. M. Tour, W. L. Van Zandt, C. P. Husband, S. M. Husband, L. S. Wilson, P. D. Franzon, D. P. Nackashi, *IEEE Trans. Nanotechnol.* 2002, **1**, 100–109.

135 C. P. Husband, S. M. Husband, J. S. Daniels, J. M. Tour, *IEEE Trans. Electron Devices* 2003, **50**, 1865–1875.

9
Intermolecular- and Intramolecular-Level Logic Devices

Françoise Remacle and Raphael D. Levine

9.1
Introduction and Background

Today, there is an intense research activity in the field of nanoscale logic devices towards miniaturization and qualitative improvement in the performance of logic circuits [1–16]. A radical and potentially very promising approach is the search for quantum computing [17–24], and this is reviewed as an emerging technology in Ref. [25]. Other alternatives are based on neural networks [26], on DNA-based computing [27–31], or on molecular quantum cellular automata [32–37]. Single-electron devices should also be mentioned because if they use chemically synthesized quantum dots (QDs) they are "molecular" in nature [38–40]. Devices that have been implemented rely on the ability to use molecules as switches and/or as wires, an approach known as "molecular electronics" [5, 12, 41–44]. This approach is currently being extended in several interesting directions, including the modification of the electronic response of the molecule through changing its Hamiltonian [45–47]. In this chapter, these topics are first reviewed, after which ongoing studies on an alternative computational model, where the molecule acts not as a switch but as an entire logic circuit, are discussed. Both, electrical and optical inputs and outputs are considered. Advantage is then taken of the discrete quantum states of molecules to endow the circuits with memory, such that a molecule acts as a finite state logic machine. Speculation is also made as to how such machines can be programmed. Finally, the potential concatenation of molecular logic circuits either by self-assembly or by directed synthesis so as to produce an entire logic array, is discussed. In this regard, directed deposition is also a possible option [48–52].

9.1.1
Quantum Computing

Quantum computing can be traced to Feynman, who advocated [53] the use of a quantum computer instead of a classical computer to simulate quantum systems. The rational is that quantum systems, when simulated classically, are very demanding in computing resources. A quantum state is described by two "numbers" – its amplitude

Nanotechnology. Volume 4: Information Technology II. Edited by Rainer Waser

ISBN: 978-3-527-31737-0

and its phase – and the number of quantum states, N, grows exponentially with the number of degrees of freedom of the system. The sizes of the matrices necessary to describe a system quantum mechanically scales as N^2, and for large systems this number becomes rapidly prohibitively large. A computer that operates quantum mechanically will require far less resources because the computations can be massively parallel due to the superposition principle of quantum mechanics. Conceptually, to compute quantum mechanically required the extension of classical Boolean logic to quantum logic [19, 54] and to set up quantum logic gates [55–61] that operate reversibly [62, 63].

In quantum computing, the logic is processed via the coupling structure between the levels of the Hamiltonian. Typically, this coupling is induced by external electrical and magnetic fields. Nuclear magnetic resonance (NMR) is a particularly promising direction for both pump and probe [20, 64]. Quantum implementations are very encouraging for search algorithms [65–68], where a power law reduction in the number of queries can be obtained. For operations where the answer is more complex than a YES or NO – such as Fourier transform operations – the read-out remains a key problem because of the collapse of the wave function when one of its component is read. One very successful outcome of quantum computing and quantum information is *cryptography* [69], where the very effective factorization algorithms (public key cryptography, Shor algorithm [70, 71]) show the potential that is available. Quantum computing has very much caught the popular imagination, and several excellent introductory books (e.g. Ref. [72]) are now available.

9.1.2
Quasiclassical Computing

The essential difference between quantum computing and the approach discussed here is that a quantum gate operates on both the amplitude and the phase of the quantum state. The phase is very sensitive to noise, and quantum computing theorists have devised various ways to protect the phase from external unwanted perturbations [73–79] or to seek to correct a corrupted phase. Because the authors' background is in molecular dynamics and coherent control, they are aware that the phase of quantum states is extremely difficult to protect, and in this chapter adopt a quasi-classical approach [80] where, while the time evolution of the molecular system is quantal, what matters in terms of inputs and outputs are the populations of the states – that is, the square modulus of the amplitudes. This approach relies on classical logic and does not require reversible gates. There are two special characteristics of quantum computing: parallelism and entanglement. Currently, the authors' investigations center on understanding the potential of the quasi-classical approach in terms of parallelism.

9.1.3
A Molecule as a Bistable Element

Another very successful approach is molecular electronics, which aims to provide molecular-based computing by using the molecule as a switch [5, 11, 12, 42, 43, 81, 82].

In the following sections, it is explained that the essential difference with what is advocated in this chapter is that the molecule can do – and has been shown to do – much more than act as a switch.

Quantum cellular automata (QCA) represents another promising approach to molecular computing where the molecules are not used as switches but as structured charge containers and information transmitted via coupled electric fields [32–34]. The charge configuration of a cell composed of a few QDs – and, more recently, of molecular complexes – is the support for encoding the binary information. Most studies on QCA consists of theoretical design, with very few experimental implementations. While the QD-based implementations [33, 34] operate at very low temperatures, theoretical modeling predicts that molecular quantum cellular automata could be operated at room temperature [36, 37, 83].

9.1.4 Chemical Logic Gates

Early proposed chemical implementations of logic gates were based on the response of molecules in solution to light or changes in chemical species concentrations [84–86]. Using photo-induced electron transfer where the emitted fluorescence is modulated by the concentrations of ions species in solution, different kinds of realizations of uni and binary gates (i.e. OR, NOT, AND, XOR, etc.) have been proposed [11, 14, 87, 88], and it has been shown that AND and XOR gates can be combined to lead to half adder and/or subtractor [4, 8, 30, 89–93] and full adder/subtractor [94–98] implementations. Other well-studied systems that lead to similar levels of logic complexity are those built on molecular motors (catenane, rotaxane), where the inputs can be communicated photochemically and electrochemically [11], or on DNA oligonucleotides [29–31]. These ways of providing the inputs and reading the outputs are rather slow, however, and do not exploit the complexity of the quantum level structure and intra- and inter-molecular couplings. Nor is it clear how to reduce the size of such devices. All of these approaches were largely limited to two bit operations (half adder or subtractors). In 2001 [94, 95], the level of the three bits operation with an optical full adder had already been reached, while more recently a cyclable full adder on a QD by electrical addressing was proposed [99]. This last example shows that QDs, and not only molecules proper, can be used to act as more than a switch.

The emphasis in this chapter is on demonstrating single-molecule rather than chemical computing. The proposal is that it is possible to use the complexity of molecules to integrate logic circuits of increasing sophistication on a single molecule or on supramolecular assemblies. Clearly, at the present time signal-to-noise considerations mean that the independent response of more than a single molecule is needed. Beyond the independent molecule, however, the proposal is to concatenate single molecules in the sense that the logic output from one molecule is the logic input to another. This is achieved by (rapid) intermolecular coupling.

9.1.5
Molecular Combinational Circuits

Combinational circuits are the simplest logic units, being built of logic gates (the most common are NOT, AND, OR, XOR, NOR, NAND) and providing a specific set of outputs, given a specific set of inputs. These circuits have no memory. In transistor-based computer circuits even the simpler such gates require to be built as a network of switches. Studies on molecular combinational circuits forms part of an intense research effort aimed at recasting into logic functions the fact that molecules can respond selectively to stimuli (inputs) of different forms (chemical, optical, electrical) and produce outputs (chemical, optical, electrical). The advantage is that a molecule which implements a combinational circuit acts not as a single switch but as an entire gate. However, most of molecular gates proposed until now have been based on chemistry in solution, and use at one stage or another a chemical stimulus (e.g. concentration of ions such as H^+, Na^+) coupled to optical or electrical stimuli. This leads rather slow rates of processing of the information and concatenation. In contrast, the authors' studies do not involve chemical inputs, and this allows faster rates to be reached. It has already been shown that, within less than 1 ps ($=10^{-12}$ s), it is possible to implement combinational circuits on a single molecule using selective photo-excitation for providing the inputs and the intra- and inter-molecular dynamics for processing the information. (See Refs. [100–102] for an example of sub-ps logic gates using femtosecond ($=10^{-15}$ s) pulses as input, and Refs. [94, 103] for an example of concatenation by intermolecular coupling.)

One advantage of using molecules in the gas phase is that far fewer molecules are required than in solution. In the preliminary gas-phase schemes, the reading of the outputs is achieved by detecting fragments of the molecule used to implement logic, which means that this particular molecule is not available for a new computation. This is not intrinsically a problem because about 10 molecules are needed to obtain a good signal-to-noise ratio; hence, the computation can be continued with other molecules present on the sample. This represents a problem for cycling, however, and is why the target is to explore the possibilities offered by non-destructive optical and electrical reading. Another route to explore is to increase the number of operations (more than 32 bits) performed on a single molecule by using the fast intramolecular dynamics for concatenation, which decreases the need for cycling and the need for I/O. As discussed above and further elaborated below, this will have major implications for the miniaturization of logic devices by implementing compound logic at the hardware level.

In this way it was possible to implement logic functions on a single molecule up to and including the ability to program – that is, to use the same physics to realize different computations.

In terms of technology transfer it is clear, however, that what the industry would very much prefer is some form of extension to the CMOS technology. So, the need is to combine the many advantages of working with molecules in the gas phase with the need to anchor molecules on the surface. In this connection, studies with self-assembled small arrays of QDs is of central interest.

9.1.6
Concatenation, Fan-Out and Other Aspects of Integration

It is an essential characteristic of modern logic circuits that the gates are integrated. Specifically, the output of one gate can be accepted as the input of the next gate; the two gates are thereby *concatenated*. Very simple examples of concatenation are the NotAND (=NAND) or NotOR, etc. gates. It should also be possible to deliver the output of one gate as input to several gates. In 2001, it was shown [94] how to concatenate the logic performed by two molecules using electronic energy transfer as the vehicle for the information forwarding. It is clear that electron transfer – including electron transfer between QDs, proton transfer, and vibration energy transfer – can all be used for this purpose. However, this remains an as-yet poorly traveled course and further studies are clearly called for as it is a possible key to high-density circuits.

9.1.7
Finite-State Machines

So far, only combinational logic has been discussed, where combinational gates combine the inputs to produce an output. A *finite-state machine* does more – and the "more" is very essential and is also something well suited to what a molecule can do. A finite-state machine accepts inputs to produce an output that is dependent both on the inputs *and* on the current state of the machine. In addition to producing an output, such a machine will update the state of the machine so that it is ready for the next operation. Technically, what the finite-state machine has is a memory, and the circuit has as many different states of memory as states of the machine (see Figure 9.1). It has been shown possible to build very simple finite-state and Turing machines at the molecular level [101, 104]. Molecules have a high capacity for memory because of their many quantum states; for example, in conformers, after radiationless transition, the molecule can undergo relaxation and be in a stable state for a long time. Retrieval of the information may then be effected by either optical or electrical pulses (see below).

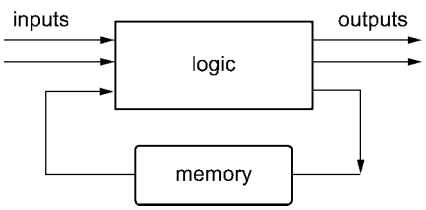


Figure 9.1 A schematic diagram of a finite-state machine. This consists of a combinational circuit ("logic") to which a memory element is connected to form a feedback path. The molecular implementation of this simple machine has been described in the authors' exploratory studies.

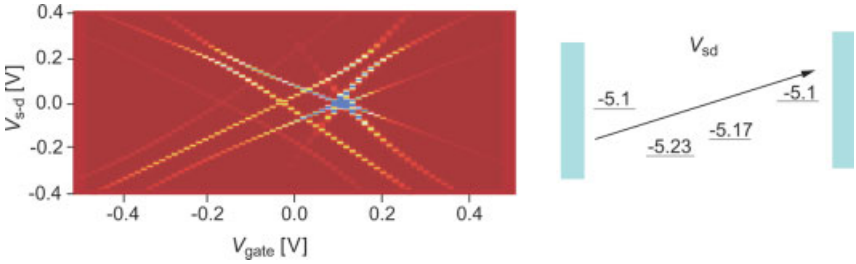


Figure 9.2 Differential conductance, d*I*/d*V* plotted as a function of the gate (V_{gate}) and the source-drain ($V_{s\text{-}d}$) voltages. The computation is for the four-level system shown where the coupling between the sites is 0.05 eV. The Stark shift of the site energies due to the source-drain bias is included in the Hamiltonian. The effect of the gate voltage is to shift the energies of system with respect to the Fermi energy of the electrodes.

A finite-state molecular machine has been proposed that can be cycled based on laser pulses [101, 104]. In recent investigations, the same optical finite-state machine has been used to implement a cyclable full adder and subtractor [96]. In a different direction, recently a molecular finite-state machine in the gas phase has been proposed by showing that a linear sequential machine can be designed using the vibrational relaxation process of diatomic molecules in a buffer gas. An alternative to optical pulses for providing inputs is to apply *voltage pulses*. In a recent study on QD assemblies (e.g. Refs. [105, 106]), it was shown that the application of a gate voltage allows the molecular orbitals to be tuned in resonance with the nanoelectrodes so that a current flows, whereas there is no current in the absence of the gate voltage (see Figures 9.2 and 9.3). A similar control is also possible on a single molecule [107–109] and for single QDs or several coupled dots.

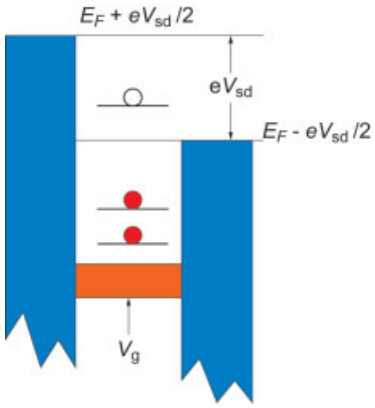


Figure 9.3 Schematic representation of a three-terminal device. The gating voltage is applied to a dieletric layer perpendicularly to the direction of the source-drain voltage. This scheme, which is equally useful for molecules and quantum dots, shows that either can be made to act as a finite-state device.

9.1.8
Multi-Valued Logic

So far, it has been taken as given that an input or an output can have one of two possible values. A laser beam can be on or off, a molecule can fluoresce or not, and a charge can transfer or not, and so on. Therefore, two-valued or Boolean logic is being implemented. Ever since Shannon showed, in 1938, the equivalence between a Boolean logic gate and a network of switches, computer circuits have been assembled from switches. First, the switches were electromechanical in nature, followed by vacuum tubes and then the transistor. Now that the discrete nature of the carriers of electrical charge has made it unclear as to how the size of the transistor might be further reduced, a major – and very serious – effort has been made to use a molecule as a switch.

Previously, several ways have been discussed of using a molecule as an entire logic gate (= a connected net of switches), or even as a finite-state machine rather than simply as a switch. There is, however, another possible generalization, which also is well-suited to what molecules are and what they can do – namely, to go to *multi-valued logic* [110–112]. What this means is that there are more than two allowed mutually exclusive answers to every question. This makes the numbers to be dealt with shorter (e.g. 1001 is the binary number that is written as 9 in base 10). In their studies, the present authors took three as a compromise between shorter numbers (e.g. 10 is ternary for 9) and the errors that can result in making a choice between too many alternatives. The molecular physics is straightforward, to take zero, one or two electrons in the valence orbital [113]. But clearly, there are other choices such as a Raman transition to several different final vibrational states. Molecules are willing and able to be pumped and to be probed to multiple states. It is not clear, however, if the industry is willing to learn to go beyond two.

9.2
Combinational Circuits by Molecular Photophysics

Combinational circuits are made of concatenated combinational gates. A logic gate accepts inputs and implements a particular function of the inputs to provide an output. In most implementations to date, the inputs and outputs are Boolean (binary) variables that can take one of two values, that is 0 or 1. Boolean logic gates can be one input–one output gates, like the NOT gate or two input–one output gates. There are in total 16 functions of two binary variables, and among these the AND, OR, XOR and INH gates are among the most commonly used in combinational circuits. The way in which half adders and full adders can be implemented are discussed in the following section, but initially attention is centered on elementary gates such as AND or OR to set the scene. The truth tables of these gates are provided in Table 9.1, and correspond respectively to taking the MIN and the MAX of the two Boolean inputs x and y. The AND gate can also be viewed as the product of the two inputs, $x \times y$. This gives an output 1 only if the two inputs are both 1. It can be implemented by two switches in series, while two switches in

Table 9.1 Truth tables of the binary AND, OR, XOR and INH gates.

INPUT		OUTPUT			
x	*y*	*x* AND *y*	*x* OR *y*	*x* XOR *y*	*x* INH *y*
0	0	0	0	0	0
1	0	0	1	1	1
0	1	0	1	1	0
1	1	1	1	0	0

parallel correspond to the OR gate. Another binary gate which is often used in combinational circuits is the XOR gate (see Table 9.1) which corresponds to the addition of the binary inputs modulo 2, $x \times y$.

The addition modulo 2 of two binary numbers works exactly as would be expected, in base 10. When the two binary inputs are both 1, the result of their sum modulo 2 is 0, but a 1 (the carry digit) must be reported in the next column of binary digits [114] (see Table 9.2).

When using the XOR and the AND gates, it is possible to build a half adder. This has two inputs, the two numbers to be added modulo 2, and two outputs, the sum modulo 2 of the two inputs, realized by a XOR gate and the carry, realized by an AND gate, the truth tables for which are shown in Table 9.1. A half adder is an important building block for molecular combinational circuits because, by concatenating two half adders, a full adder can be built. Usually, as shown by an example in Table 9.2, binary numbers of length more than one digit must be added. A full adder is more complex than a half adder, in that it takes the carry digit from the addition of the previous two digits into account (called the "carry in"). A full adder has therefore three inputs – the two digits to be added and the carry digit from the previous addition – and two outputs – the sum modulo 2 of the three inputs and the carry digit for the next addition (see Table 9.2) – called the "carry out".

In photophysics logic implementations, the inputs are provided by laser pulses, their physical action being to excite the molecule, typically to electronic excited states. The logical value 1 is encoded as the laser being "on", and the logical value 0 as the laser being "off". The molecules in the sample act independently of one another, and there is uncertainty as to whether every molecule absorbs light. It is therefore not the case that a single molecule suffices to provide an output. When working with an ensemble of molecules, there is no need to read a strict "yes" or a strict "no" from each molecule.

Table 9.2 Addition of the two binary numbers $x = 010$ and $y = 111$.

Carry	**1**	**1**		
x		0	1	0
y		1	1	1
Sum	1	0	0	1

What is needed is to excite enough molecules to be above the threshold for the detection of light absorption. The detection of the outputs is by fluorescence and/or by detecting ions. The detection of ions is relatively easy, so by monitoring the absorption by photoionization of the molecule, it is sufficient if only a small number of molecules respond to the input. An excited molecular ion typically fragments, and the detection of ionic fragments can also be used to encode outputs. Although ionic fragments can be detected very efficiently, the price is that the molecule self-destructs at the end of the computation and so the gate cannot be cycled. The details of an optically addressed half and full adder that can be cycled are provided in Section 9.1.3 [96].

First, however, the implementation of a half adder will be discussed and two approaches for doing this will be compared. It will then be shown how to implement a fault-tolerant full adder on a single molecule [115] using photophysics in the gas phase of the 2-phenylethyl-*N*,*N*-dimethylamine (PENNA) molecule [116–118]. This implementation follows the lines of the 2001 implementation of a full adder on the NO molecule [95]. Finally, the realization of a full adder by concatenation of two half adders is discussed, where the logic variables are transmitted between the two half adders by energy transfer between two aromatic molecules that are photoexcited [94] in solution.

9.2.1
Molecular Logic Implementations of a Half Adder by Photophysics

A discussed above, a half adder has two outputs: Addition modulo 2 is implemented by the XOR gate, and the carry digit is the result of an AND operation. A molecular realization of the logical XOR operation is challenging because the output must be 0 when both inputs are applied. For a photophysics implementation of the inputs this means that when both lasers are on, there is not the output that is observed when only one laser is on. The realization of the AND gate is comparatively easier because an output is produced only if both inputs are on, and so any reproducible experiment, which requires two inputs for generating an output can implement an AND gate [103]. The truth table for the implementation of a half-adder by optical excitation is given in Table 9.3.

Two-photon resonance-enhanced absorption by aromatic chromophores has been used as an effective way to implement an XOR operation on optical inputs at the molecular level [94, 95, 104]. The two photons (each of a somewhat different color and

Table 9.3 Truth table for an implementation of a half-adder by photoexcitation.

x(laser 1)	*y* (laser 2)	Carry (AND)	Sum (XOR)	(carry, sum)
0	0	0	0	(0,0)
1	0	0	1	(0,1)
0	1	0	1	(1,0)
1	1	1	0	(1,1)

therefore distinguishable) represent the possible inputs. For an XOR gate to be physically realizable, the following conditions must be satisfied. First, for aromatic molecules or for molecules with aromatic chromophores the resonant level, typically the first optically bright electronically excited state, S_1, has a fairly broad absorption band consisting of many vibronic transitions. So, two photons of different frequencies (and therefore distinguishable) can be absorbed with a similar cross-section. This provide the OR part of the XOR gate (second and third lines of Table 9.3): if laser light of either frequency is on, the output can be identified as the fluorescence. The exclusive part results from the following effect: it is often the case that, having absorbed one photon, the cross-section of an aromatic chromophore to absorb a second photon is higher. That the bottleneck for two-UV photon absorption is often the absorption of the first photon has been realized since the earliest days of visible/UV multiphoton ionization/dissociation [119, 120], and has been used extensively since then. Therefore, when two light pulses are applied the system need not remain in S_1 for a significant length of time because it can absorb a second photon. Whether it preferentially does so depends on the particular molecule. Either input laser can excite from the ground state to S_1. Therefore, the fluorescence signal will increase when both lasers are on. The input in this case is 1,1 and from Table 9.3 the need is to detect the output 1,0. In the presence of two photons the fluorescence from S_1 (a rather slow, >5 ns scale, process in PENNA) may fall, but the more secure detection of the presence of both input beams is the increase of the number of ions. There is an increase rather than pure onset because a single laser can also cause ionization, but this occurs with low intensity. A high ionization intensity corresponds to a simultaneous input of both lasers. In practice, the two events "high ionization efficiency" and "low ionization efficiency" can easily be distinguished by the use of discriminators and analog electronics. A simulation of the temporal response of the molecule to the laser pulses, so as to show that this is possible, is available in Ref. [115]. To conclude, the 1,1 input is identified as a high ionization signal.

The use of two lasers of different colors is dictated by the need to represent two distinct inputs, but there is a clear physical advantage if the two frequencies differ by more than a vibrational frequency in S_1: because the S_1 and S_0 vibrational frequencies are not exactly the same, the down-pumping by the other laser is not resonance enhanced and thus improbable, as shown schematically in Figure 9.4.

There are therefore two ways to implement a half adder (see Table 9.4). The direct way is to detect separately the sum and the carry and to assign them to a different experimental probe. In the case of the PENNA molecule (see Figure 9.5) the inputs are encoded as UV lasers being off or on. For the sum digit, the experimental probe is the detection of the fluorescence from the S_1 electronic state. The carry digit is the result of the absorption of two photons. In the case of PENNA, the absorption of two UV photons at the chromophore end causes local ionization, followed by charge migration to the N-end of the chain. The carry digit is therefore encoded as the detection of N-fragments. As discussed above, the absorption of a second photon decreases the intensity of the fluorescence from S_1, but typically does not quench it enough so that detection of the carry digit through N-end ions is preferable.

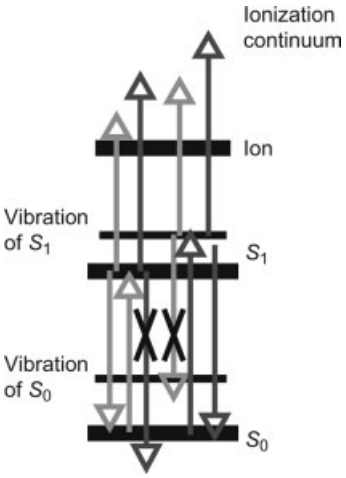


Figure 9.4 Two-photon ionization with the two inputs being lasers of unequal frequency can be made fault-tolerant to stimulated emission. This is particularly so if one laser operates on the 0,0 transition while the other pumps a vibrationally excited level of S_1. For details, see the text.

One way of implementing a fault-tolerant half adder is to combine the result of the carry and the sum digit into one word (carry, sum). This is shown in the fifth column of Table 9.3. As can be seen, the four distinct pairs of inputs of the half adder – that is, (0,0), (1,0), (0,1) and (1,1) – corresponds only to three distinct outputs of (0,0), (0,1) and (1,0). This is because in addition modulo 2, if the carry is 1, the value of the sum is necessarily 0. Therefore, instead of assigning a separate physical probe for the sum and the carry, it is possible to assign a physical probe for the three different logical values of the word (carry, sum). It should be noted that the binary meaning of the word (carry, sum) corresponds to the number of inputs with value 1. In the case of PENNA, the choice was made to assign the word (0,1) to the presence above threshold

Table 9.4 Truth table and experimental probes used for the two ways of implementing a half-adder on PENNA.

x (UV(1))	*y* (UV(2))	Carry (AND)	Sum (XOR)	Probe for carry (AND)	Probe for XOR	Probe for words (carry,sum)
0	0	0	0	no N-end fragment	no fluo from S_1	no output signal (0,0)
1	0	0	1	no N-end fragment	fluo from S_1	fluorescence from S_1 (0,1)
0	1	0	1	no N-end fragment	fluo from S_1	fluorescence from S_1 (0,1)
1	1	1	0	N-end fragment	no fluo from S_1	N-end fragment (1,0)

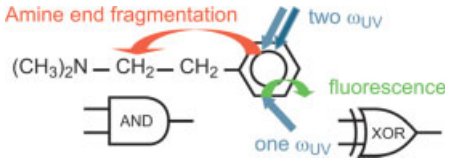


Figure 9.5 Schematic representation of the two-photon excitation scheme used to implement a half adder on PENNA.

of a fluorescence signal from S_1, and the word (1,0) to the detection above a threshold value of a N-end fragments. The fault tolerance of this scheme arises from the fact that in case of inputs (1,1), a N-end fragment ion detection will be reported, irrespective of the intensity of the fluorescence signal from S_1. The scheme is fault tolerant with respect to the extent of fluorescence from S_1 in case of two-photon excitation. Detecting the (0,0) output is straightforward as no excitation is provided. The two ways of implementing a half adder on PENNA are summarized in Table 9.4.

This half-adder self-destructs at the end of the computation because the local ionization at the chromophore end causes the PENNA ion to fragment. However, this scheme allows for a remarkable sensitivity in the detection of the outputs because very few ions can already be detected with a good signal-to-noise ratio. Although this involves an ensemble of molecules, not all of which provide an answer,. However, response is needed from only 100 molecules to obtain acceptable statistics, and this response occurs quite rapidly.

In a full adder, the word (1,1) as an output is allowed, so that a full adder has four distinct binary words as outputs, namely (0,0), (0,1), (1,0) and (1,1). As discussed above, it has also one more input than the half adder, the carry digit. It is shown in the next section how a full adder can be implemented on the PENNA molecule using the same fault-tolerant scheme for probing the outputs. This manner of implementation of a full adder is contrasted with the implementation based on the concatenation of two half adders, in which the sum and the carry are detected separately.

9.2.2
Two Manners of Optically Implementing a Full Adder

A full adder has three inputs and produces two outputs, the sum which is the addition modulo 2 of the two inputs and the carry in digit and the carry out, which for the next cycle of computation becomes the carry in:

$$\text{sum out} = x \oplus y \oplus \text{carry in} \tag{9.1}$$

$$\begin{aligned}\text{carry out} &= (x \otimes y) + ((x \oplus y) \otimes \text{carry in}) \\ &= (x \otimes y) + (x \otimes \text{carry in}) + (y \otimes \text{carry in})\end{aligned} \tag{9.2}$$

where $\oplus$ means addition modulo 2 (XOR), $\otimes$ means binary product (AND) and $+$ means OR. The carry out logic equation can be simplified to

$$\text{carry out} = x \otimes y + x \otimes \text{carry in} + y \otimes \text{carry in} \tag{9.3}$$

In the implementation on PENNA that was proposed above, the two inputs x and y are encoded as for the half adder, by two UV photons with slightly different wavelengths. The carry in digit is encoded as a laser pulse of green light which is intense enough that two photons can be absorbed to allow the transition to the S_1 state by a non-resonance-enhanced two-green photon transition. The four outputs words (carry out, sum): (0,0), (0,1), (1,0) and (1,1) are detected each by a distinct experimental probe. As in the half adder implementation of PENNA, the output word (0,1) is detected as fluorescence from the S_1 state while the output (1,0) as the presence of N-end fragment ions. The detection of the output (0,0) is straightforward as it corresponds to no inputs. The output (1,1) corresponds to the three inputs having the value 1; that is, the PENNA molecule is excited by two UV (x and y) and two green photons (carry in). Experimentally, this amount of energy is causing fragmentation at the C-end (instead of fragmentation at the N-end which occurs when only two inputs are 1, in which case it is only the equivalent in energy of two UV photons). The presence of C-end fragment ions above a given threshold is therefore used to detect the (1,1) output. The excitation scheme and experimental probes of the outputs are shown in Figure 9.6 and the corresponding truth table in Table 9.5. Note that, as in the case of the half adder, the output word (carry, sum) counts in binary how many inputs are 1.

Another way to implement a full adder is by concatenation of two half adders. The corresponding combination logic circuit is shown in Figure 9.7.

The physical implementation, by photoexcitation of a donor–acceptor complex in solution[94], is an example of intermolecular concatenation of two half adders by energy transfer. It demonstrates that one molecule is able to communicate its logical output to another molecule. The implementation is on a specific pair (rhodamine 6G–azulene), for which considerable data are available, but the scheme is general enough to allow a wide choice of donor and acceptor pairs. The first half-adder is realized on rhodamine 6G, and the second half adder on azulene. The midway sum is transmitted from the first to the second half adder by electronic energy transfer between rhodamine 6G and azulene.

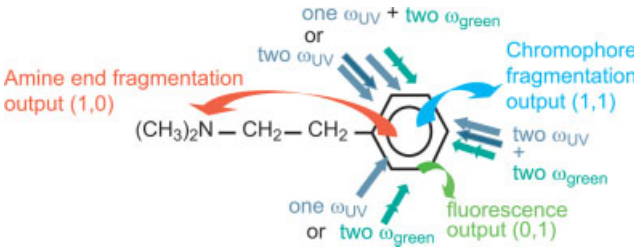


Figure 9.6 Excitation scheme of PENNA for implementing a full adder. Fluorescence from the S_1 state of the phenyl ring codes for carry-out = 0, sum-out = 1, fragmentation at the amine end codes for carry out = 1, sum out = 0 and fragmentation at the chromophore end for carry out = 1, sum out = 1 (see also Table 9.5). The two binary digits to be added are encoded as the two UV photons being on or off. The carry-in is encoded as excitation by two green photons.

Table 9.5 Truth table and detection scheme for the outputs for the optical implementation of a full adder on PENNA.

x (UV(1))	*y* (UV(2))	Carry in (vis, two-photon)	Carry out	Sum out	Output word (carry,sum)	Probe for output word
0	0	0	0	0	(0,0)	No signal
1	0	0	0	1	(0,1)	fluorescence from S_1
0	1	0	0	1	(0,1)	Fluorescence from S_1
1	1	0	1	0	(1,0)	N-end fragment
0	0	1	0	1	(0,1)	Fluorescence from S_1
1	0	1	1	0	(1,0)	N-end fragment
0	1	1	1	0	(1,0)	N-end fragment
1	1	1	1	1	(1,1)	C-end fragment

The physical realization of the two half adders relies (as in the case of the half adder implementation on PENNA discussed above) on the fact that the absorption of one or two UV photons (inputs x and y) by the donor (acceptor) molecule leads to distinct outputs. However here, unlike the case of PENNA, the absorption of the second photon does not lead to ionization, but rather to absorption by a second excited electronic state, S_2. In general, a molecular half adder will be available for molecules which have a detectable one-photon and a detectable two-photon absorption. This seems to go against Kasha's rule [121], but in fact there are enough exceptions. Azulene and many of its derivatives provide one class. The emission from the second electronically excited state, S_2, is often as strong or stronger as the fluorescence from S_1 [122, 123]. More in general, emission from S_2 is not forbidden; rather, due to competing non-radiative processes it often has a low quantum yield but it is definitely detectable, particularly so as it is much to the blue as compared to the emission from S_1. If necessary, the emission from S_2 can be detected by photon counting. There is, therefore, a case where the outputs of the XOR gate and the AND gate that constitute the half adder can be probed separately, with sufficient fidelity. The output of the XOR gate of the first half adder is encoded as populating the S_1 state of rhodamine 6G, while the output of the XOR gate of the second half adder (the sum out) is encoded as detecting fluorescence of the S_1 state of azulene. The output of the first AND gate

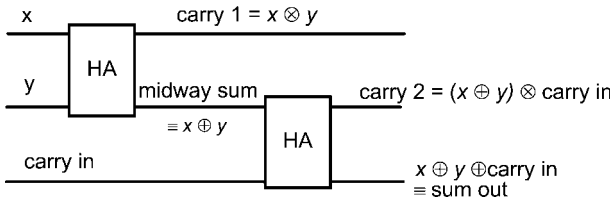


Figure 9.7 Combinational circuit of a full adder implemented by concatenation of two half adders.

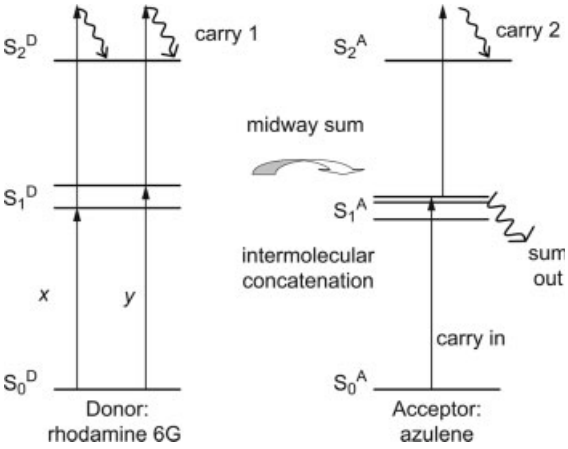


Figure 9.8 Photophysics of the implementation of a full adder by concatenation of two half adders on the donor–acceptor complex rhodamine 6G–azulene in solution.

(carry 1 in Figure 9.7) is probed by fluorescence from the S_2 of rhodamine 6G and correspondingly, the output of the AND of second half-adder (carry 2 in Figure 9.7) by fluorescence of the S_2 state of azulene. The concatenation between the two half adders is performed by the fairly rapid [124] intermolecular electronic energy transfer. Specifically, the well-characterized [125–127] transfer from the S_1 level of rhodamine 6G to the S_1 of azulene was proposed.

The photophysical scheme for the implementation of a full adder on the donor–acceptor complex rhodamine 6G–azulene is summarized in Figure 9.8. The S_1 level of rhodamine 6G can be readily pumped with photons absorbed within the $S_0 \rightarrow S_1$ band. The frequencies ω_1 ($\equiv$ input x) $= 18\,797\,\text{cm}^{-1}$ (the second harmonic of the Nd-YAG laser) and ω_2 ($\equiv$ input y) $= 18\,900\,\text{cm}^{-1}$ are taken. This is not needed for the full adder, but the absorption to S_1 can be detected through its emission at about $17\,500\,\text{cm}^{-1}$ [125, 126]. This emission is logically equivalent to $x \oplus y$ because if the intensity is high enough due to two photons being present, the donor will be pumped either directly to S_2 or to higher levels, followed by ultrafast non-radiative relaxation to S_2. The large absorption cross-section of $2.5 \times 10^{-18}\,\text{cm}^2\,\text{molecule}^{-1}$ [128] for the $S_1 \rightarrow S_n$ ($n \geq 2$) of rhodamine 6G ensures efficient pumping of S_2. The emission from S_2 is at about $23\,250\,\text{cm}^{-1}$, with a quantum yield of about 10^{-4} [126]. It is this emission which serves to logically implement the left AND gate, and it is equivalent to $x \otimes y$ (denoted as carry 1). The S_1 level of the donor transfers the energy, via the Forster mechanism [124] to the azulene acceptor, the S_1 level of which is at $14\,400\,\text{cm}^{-1}$, that emits in the 13 400 to $11\,000\,\text{cm}^{-1}$ range [129]. This emission provides the logical sum output [Eq. (9.1)]. The S_2 level of azulene has its absorption origin at $28\,300\,\text{cm}^{-1}$, and so it can be reached from S_1 by a third photon of frequency $14\,400\,\text{cm}^{-1}$. The same photon can also pump ground-state azulene to its S_1 level. Emission (or lack thereof) from S_2 of azulene at $26\,670\,\text{cm}^{-1}$ [127] provides the carry 2 bit. The carry 1 and the carry 2 cannot be equal to 1 together, because if the carry 1 is 0,

the midway sum is 0, meaning that even if the carry in is 1, the carry 2 cannot be 1. In other words:

$$carry\,out = carry\,1 + carry\,2 \tag{9.4}$$

The carry out is therefore physically probed by monitoring the fluorescence from the S_2 states of rhodamine 6G and azulene, which logically corresponds to the first line of Eq. (9.2).

The advantage of an all-optical scheme for the full adder implementation compared to the implementation on PENNA as discussed above is that the adder does not self-destruct at the end of the computation. Another advantage is that it operates relatively rapidly. The energy transfer rate for a solution of 10^{-3} M azulene, estimated using the S_1 fluorescence spectrum of rhodamine 6G and the absorption spectrum of azulene, is about $10^{10}\,s^{-1}$. This rate is sufficient for present needs, but it can be increased [130] if the two chromophores are incorporated within a single molecular unit using a short bridge to connect them [124]. The increase in the rate will be particularly significant (five orders of magnitude) if the bridge is rigid [130, 131]. It should be emphasized that a rigid bridge is required to achieve a very high rate. Many other couples based on commonly used laser dyes as donors and azulene derivatives [128, 132] may also be utilized for implementation of the logic gate [133, 134].

9.3 Finite-State Machines

Finite-state (also called sequential) machines are combinational circuits with memory capability. The memory registers are the internal state(s) of the machine [135, 136]. As in a combinational circuit, the outputs of the machine depend on the inputs, but in addition the output also depends on the current state of the machine. It is this dependence of the output on the state of the machine that endows finite-state machines with a "memory". The memory of the machine corresponds to the state of the experimental system, and this state can be changed by applying suitable perturbations, such as optical or voltage pulses. As in the other logic schemes, the "logic" part is an encoding of the subsequent dynamics of the system.

The finite-state machine computational model takes advantage of two aspects that are natural for quantum systems:

- A physical quantum system has discrete internal states and its response to perturbation will in general depend on what state it is in.
- Perturbations can be applied sequentially, so that the machine can be cycled.

By taking advantage of the two points above, the implementation of several forms of finite-state machine was proposed: a simple set-reset that can be either optically [101] or electrically addressed [106]; an optical flip-flop [104] and full adder and subtractor [96]; an electrically addressed full adder [106] and a electro-optically addressed counter [137]. Beyond that it has been shown, using optical addressing,

that a molecule can be programmed and behaves (almost) like a Turing machine [101]. The caveat "almost" is introduced because a molecule can have only a finite number of quantum states, whereas a Turing machine has an unlimited memory. Possibly this is not a true limitation since if indeed the number of quantum states of the universe is finite – sometimes known as the "holographic bound" – then no physical system can strictly act as a Turing machine.

In this section, a review is conducted of optically addressed finite-state machines, up to a full adder (Section 9.3.1) and an electrically addressed machine (Section 9.3.2). If molecules and/or supramolecular assemblies are to offer an inherent advantage over the paradigm of switching networks, it will likely be through each molecule acting as a finite-state unit.

9.3.1
Optically Addressed Finite-State Machines

Laser pulses are used to optically address atomic or molecular discrete quantum states. All of the schemes discussed here are based on the Stimulated Raman Adiabatic passage (STIRAP) pump-probe control scheme, that allows the population of the quantum states of atoms or molecules to be manipulated. The advantages of the STIRAP control scheme for implementing finite-state machines are that the external perturbation can induce a change of state with a very high efficiency (close to 100%), and that the residual noise which accumulates when the machine is cycled can be erased by resetting it. Moreover, the perturbation has a distinctly different effect on the system depending on the initial state. These advantages are supported by experimental results for atomic (i.e. Ne [138]) and molecular systems (i.e. SO_2 [139], NO [140]), and the dynamics is well-described by solving the quantum mechanical time-dependent Schrödinger equation [141–143].

Here, the operation of finite-state machines using quantum simulations on a three-level system with a Λ-level scheme is described (see Figure 9.9). The pump pulse, with photons of frequency ω_P, is nearly resonant (up to a detuning Δ_P) with the $1 \rightarrow 2$ transition, while the Stokes pulse, with photons of frequency ω_S, is nearly resonant (up to a detuning Δ_S) with the $2 \rightarrow 3$ transition. Levels 1 and 3 are long-lived, but level 2 is metastable because it can fluoresce. The spontaneous emission from level 2 provides a readable output. The important feature of this level structure is that there are two routes for going from level 1 to level 3. The first route is a kinetic or

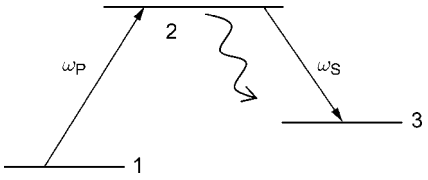


Figure 9.9 The Λ-level scheme is a STIRAP experiment. The pump transition driven by ω_P is between levels 1 and 2, while the Stokes or dump transition is between levels 2 and 3 induced by ω_S. The population from level 2 is detected by its fluorescence.

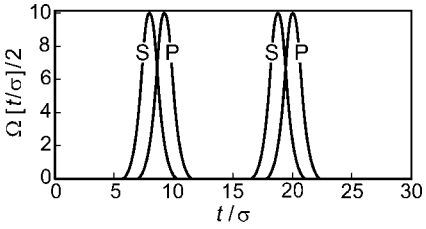


Figure 9.10 Resistance–time profile of an individual SP pulse and of the sequence of the two pulses, as used in the simulation reported in Figure 9.11. The plot is as a function of time in reduced units t/σ, where σ is the width of the S and the P pulses and the two widths are taken as equal.

"intuitive" pump scheme where first a pulse of frequency ω_P is applied so that population is transferred from level 1 to the excited level 2. From this level, the population can fluoresce or be transferred to level 3 by induced emission, using a Stokes pulse of frequency ω_S. The second route is the counter-intuitive or STIRAP route that takes population from level 1 to level 3 with almost no population in level 2, and therefore no spontaneous emission from level 2. In this counter-intuitive route, the Stokes (S) pulse of frequency ω_S is applied first and the Pump (P) pulse of frequency ω_P is somewhat delayed and applied subsequent to the dump pulse, preferably so that its front still overlaps the tail of the pulse of frequency ω_S (see also Figure 9.10).

For the purpose of constructing a finite machine, the following observation is used. Suppose that passage occurs from level 1 to level 3 by the counter-intuitive route; this means that the S pulse is on first and the P pulse is delayed. This promotes (almost) all molecules from level 1 to level 3. The same set of two pulses can now be applied, keeping their respective order in time, and this will drive almost all molecules back to level 1. However, since the system is in level 3 this order of pulses constitutes a kinetic route, and therefore state 2 is populated as an intermediate state. In the simulation of finite-state machines, the pulse conditions are chosen such that the kinetic route leaves a few percent of molecules in level 2. The level 2 will fluoresce, and this will serve as the signature of the kinetic route. In other words, the set of two pulses – Stokes followed immediately by pump – drives the molecule from level 1 to level 3, or vice-versa, depending on what state it is in. The fluorescence from level 2 is the signature of which transition was driven. It should be noted that, by using the S and the P pulses as two distinct inputs, with either one being on or off, it could be shown possible to implement a set–reset finite-state machine and also a programmable Turing machine on the Λ-level structure [94]. In order to keep the discussion as simple as possible, these schemes will not be discussed here. In the present discussion the input is the presence or absence of the superposition of both pulses.

For the purpose of implementing a flip-flop and a full adder and subtractor, the optical input is defined as a Stokes pulse at the frequency ω_S, followed in time by a pump pulse at frequency ω_P. The input is referred to as a SP pulse (see Figure 9.10). The overlap in time between the two pulses and their intensity is adjusted such that if

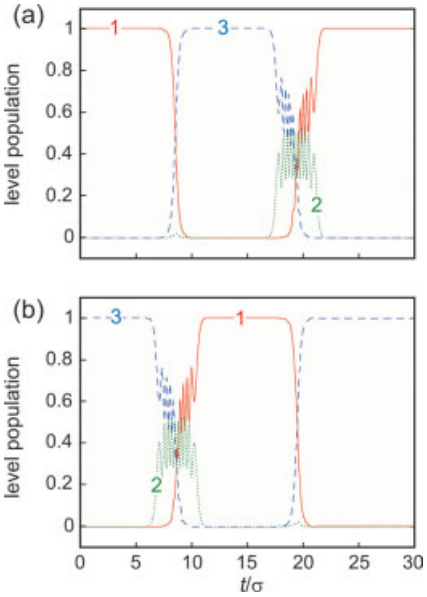


Figure 9.11 Population of the levels 1 (red), 2 (green) and 3 (blue) as a function of time when two SP pulses are applied (see Figure 9.10). The reason why such a sequence of two pulses is useful as a way of reading the state of the machine is explained in the text. The populations are computed as $|c_i(t/\sigma)|^2$ by numerical integration of the time-dependent Schrödinger equation. (a) The system is initially in level 1. The first SP pulse takes it to level 3 via the STIRAP route, with essentially no transient population in level 2. The second SP pulse takes the population in level 3 back to level 1. This second transition occurs via the kinetic route and significant transient population in level 2. (b) The system is initially in level 3. The population transfer from level 3 to level 1 goes via the kinetic route while the second SP pulse takes the system back to level 3 via the STIRAP route.

the system is initially in level 1, the SP pulse transfers to level 3 via the STIRAP route, without any significant population in level 2. On the other hand, if the system is initially in level 3, the SP pulse will pump it down to level 1 via the kinetic route, that can be detected by spontaneous emission from level 2. The time profile of the SP pulse is shown in Figure 9.10.

The quantum simulations illustrating the two routes that are possible using a SP pulse as an input are shown in Figure 9.11. The purpose of the simulation is to show that, by either route, the SP pulse achieves an essentially 100% population transfer between levels 1 and 3. The main source of noise is the spontaneous emission from level 2 that can end up either in level 1 or 3 or to yet another level, in which case the molecule is lost from the ensemble. To achieve a population transfer close to 100% between levels 1 and 3, rather intense pulses are needed, with the result that in the kinetic route the population in level 2 remains low (see Figure 9.10). Only a few photons are necessary to detect the output, so that detecting the output does not introduce too much noise. After a few cycles, the noise accumulation can be corrected for by resetting the machine (see Ref. [104]).

For the three-level structure shown in Figure 1.9, the Hamiltonian in the rotating wave approximation takes the form [141, 144, 145]:

$$\mathbf{H} = \frac{1}{2}\begin{pmatrix} 2\omega_1 & \Omega_P(t)\exp(i\omega_P t) & 0 \\ \Omega_P(t)\exp(-i\omega_P t) & 2\omega_2 & \Omega_S(t)\exp(-i\omega_S t) \\ 0 & \Omega_S(t)\exp(i\omega_S t) & 2\omega_3 \end{pmatrix} \tag{9.5}$$

where the two pairs of levels are coupled by nearly resonant transient laser pulses. The Rabi frequency [141, 146] is denoted as $\Omega(t)$. It is given by the product of the amplitude of the laser pulse, $E(t)$ and the transition dipole, μ: $\Omega(t) = \mu\, E(t)/h$. The central frequency of the Pump and Stokes lasers is almost resonant with the $1 \rightarrow 2$ and the $2 \rightarrow 3$ transitions; that is, $\omega_P = \omega_2 - \omega_1 - \Delta_P$ and $\omega_S = \omega_2 - \omega_3 - \Delta_S$ and the detunings are small and taken to be equal in the simulation, $\Delta_P = \Delta_S = \Delta$. Therefore, the two lasers are off resonance for the transitions for which they are not intended. In the rotating-wave approximation, the Hamiltonian couples between levels using only the component of the oscillating electrical field that is in resonance or nearly so for the two levels. The Hamiltonian [Eq. (9.5)] is that used in earlier studies of STIRAP [141, 147–149].

The Hamiltonian [Eq. (9.5)] can be recast in the interaction picture where it takes the form:

$$\tilde{\mathbf{H}} = \frac{1}{2}\begin{pmatrix} 0 & \Omega_P(t)\exp(-i\Delta_P t) & 0 \\ \Omega_P(t)\exp(i\Delta_P t) & 0 & \Omega_S(t)\exp(i\Delta_S t) \\ 0 & \Omega_S(t)\exp(-i\Delta_S t) & 0 \end{pmatrix} \tag{9.6}$$

The wavefunction of the system, $\psi(t)$, is a linear combination of the three levels, with time-dependent coefficients:

$$\psi(t) = \sum_{i=1}^{3} \tilde{c}_i(t)|i\rangle, \quad \tilde{c}_i(t) = c_i(t)\exp(-i\omega_i t) \tag{9.7}$$

where $\tilde{c}(t)$ are the coefficients in the interaction picture. These satisfy the matrix equation of the time-dependent Schrödinger equation, $i\, d\tilde{\mathbf{c}}/dt = \mathbf{H}\tilde{\mathbf{c}}$, which is solved numerically without invoking the adiabatic approximation [150]. The total probability, $\mathbf{c}^T\mathbf{c} = \tilde{\mathbf{c}}^T\tilde{\mathbf{c}}$, is conserved because the Hamiltonian is Hermitian.

Figure 1.11 shows the effect of acting with two SP pulses successively, for the system being initially in level 1 [panel a and in level 3 (panel b)]. The time profile of the sequence of two SP pulses in shown in Figure 9.10.

The simulations start with the molecule either in level 1 (Figure 9.11, panel a), or level 3 (panel b). The sequence of pulses as shown in Figure 9.10 returns the system to the level it started from. In a single cycle of the machine the SP pulse is applied only once. Parameters of the simulation given in reduced time units (t/σ) are: $\Omega_P(t/\sigma) = \Omega_S(t/\sigma) = 20.05\ \exp\ (-((t/\sigma)-\tau_i)^2/2)$, with $\tau_{s1} = 8$, $\tau_{p1} = 9.25$, $\tau_{s2} = 18.75$, $\tau_{p2} = 20$. The detuning $\Delta = \Delta_S = \Delta_P = 4(\sigma/t)$. The area of the pulse, $A(t) = \int \Omega(t/\sigma)d(t/\sigma)$ is 6.38 π. These details are quoted since the achievement of an essentially complete population transfer by the kinetic route (as shown in Figure 9.11) is sensitive to the intensity of the pulse and also to the detuning.

The physics shown in Figure 9.11 is all that is required to implement finite-state machines. The implementation of a full adder and a full subtractor are discussed

below; these are implemented in a cyclable manner, with each full addition or subtraction requiring two steps. The inputs x and y are both encoded as an SP pulse. The duration of a computer time step is taken to be somewhat longer than the duration of the input SP pulse. Here (unlike Section 9.3.1, where the combinational circuits implement a full adder) there is no need for concatenation because the carry (borrow) is encoded in the state of the machine for the first step and the midway sum (the midway difference) is encoded in the state of the machine for the second step. This is a major advantage of finite-state machines. The state of the machine encodes intermediate values needed for the computation.

A logic value of 0 is encoded for the carry-in or of the borrow digit as the molecule being in level 1, and a logic value of 1 as the molecule being in level 3. During the course of the discussion, it will also be shown (see Table 9.8) how the first cycle of the operation can also be logically interpreted as a T-flip-flop [136] (T for toggle) machine.

In order to cycle an adder after two optical inputs, the machine should be in a state that corresponds to the carry for the next addition, so that it is ready for the next operation. At present, a scheme which does exactly that cannot be devised, as two more operations are required in order for the machine to be ready for the next cycle. The reason for this is that, as shown below, at the end of the two cycles, the sum out is encoded in the state of the machine. So, the first requirement is to read the state of the machine in order to obtain the sum out as an output. This can be readily done by applying a SP pulse (as explained in Ref. [104] and shown in Figure 9.11). If the machine is in logical state 1 (level 3), an output from level 2 will be obtained, whereas if it is in logical state 0 (level 1) there will be no input. The machine is then restored to state 0 (level 1) by applying a second SP pulse if needed. Next, the carry out must be encoded in the internal state of the machine. If fluorescence was observed either in the first step or in the second step of the addition, it means that the carry is 1 and a SP pulse must be input in order to bring the machine to internal state 1 (level 3). Depending on the value of the sum and the carry out, the preparation of the machine for the next cycle may be automatic, in the sense that reading the sum out can coincide with encoding the carry in.

In a full addition, the order into which the three inputs, x, y and carry in, are added does not matter. This is unlike the case of a full subtractor, where the order does matter – that is, $x - y$ and $y - x$ differ by a sign. In order that the first step is the same for the full addition and the full subtraction discussed below, the process is started by adding the carry in and the y input digit. The finite-state machine implementation of a full adder goes along lines similar to the combinational circuit implementation by concatenation of two half adders. It is simply the order of adding the three inputs that differs, in order to take advantage of the memory provided by the internal state of the machine. The first step can be summarized by the Boolean equations:

$$\begin{aligned} state(t+1) &= carry\,in \oplus y \\ carry\,1 &= carry\,in \otimes y \end{aligned} \tag{9.8}$$

The corresponding truth table is given in Table 9.6.

The carry 1 is logically represented as the output of the machine after the first step (at time $t + 1$) and if its value is 1, fluorescence is detected from level 2. This only

Table 9.6 Truth table for the first half addition.

***state*(*t*) ≡ carry in**	***y*(*t*) ≡ SP pulse**	***state*(*t* + 1) (XOR) ≡ midway sum**	***output*(*t* + 1) (AND) ≡ carry 1**
0 (level 1)	0	0 (level 1)	0
0 (level 1)	1	1 (level 3)	0
1 (level 3)	0	1 (level 3)	0
1 (level 3)	1	0 (level 1)	1

occurs if the input is (1,1) – that is, the carry in was 1 (system in level 3) and the input SP pulse is 1, so that is induces a transition from level 3 to level 1 via the kinetic route. At the next interval the second digit, *x*, is input as a SP pulse. The truth table is given in Table 9.7, and corresponds to the following logic equations.

State ($t + 2$) is the XOR sum of the three inputs (*x*, *y*, and the *carry in*):

$$state(t+2) = state(t+1) \oplus x = (state(t) \oplus y) \oplus x = carry\,in \oplus y \oplus x \quad (9.9)$$

and corresponds the sum out given by Eq. (9.1) above. Using a bar to denote negation

$$\begin{aligned} carry2 &= state(t+1) \otimes x = (carry\,in \oplus y) \otimes x \\ &= (\overline{carry\,in} \otimes y + carry\,in \otimes \bar{y}) \otimes x \\ &= x \otimes y \otimes \overline{carry\,in} + x \otimes \bar{y} \otimes carry\,in \end{aligned} \quad (9.10)$$

The carry out is obtained by reading fluorescence from level 2, either at time $t + 1$ or at $t + 2$, *carry out* = *carry* 1 + *carry* 2, which corresponds to Eq. (9.4) above.

It can now be shown how encoding level 1 as the logical value 1 of the internal state of the machine and level 3 as the logical value 0 and still using a SP pulse as the input, known as *y*, leads to different state equations and different machines. With this convention, the following logic equations are obtained:

$$state(t+1) = \bar{y}\,(t) \otimes state(t) + y(t) \otimes \overline{state(t)} \quad (9.11)$$

$$Output(t) = y(t) \otimes \overline{state(t)} \quad (9.12)$$

Table 9.7 Truth table for the second half addition.

State(*t* + 1) ≡ midway sum	***x*(*t* + 1) ≡ SP pulse**	**State(*t* + 2) (XOR) ≡ sum**	**Output(*t* + 2) (AND) ≡ carry 2**
0 (level 1)	0	0 (level 1)	0
0 (level 1)	1	1 (level 3)	0
1 (level 3)	0	1 (level 3)	0
1 (level 3)	1	0 (level 1)	1

Table 9.8 Truth table for the operation of the machine with logical encoding level 1 ≡ 1 and level 3 ≡ 0.

State (t)	$\gamma(t)$ (SP pulse)	*State* ($t+1$)	*Output* (t)
0 (level 3)	0	0 (level 3)	0
0 (level 3)	1	1 (level 1)	1 (kinetic)
1 (level 1)	0	1 (level 1)	0
1 (level 1)	1	0 (level 3)	0 (STIRAP)

The equation for the next state corresponds to a XOR operation identical to Eq. (9.8), while the logical equation for the output corresponds to an INH gate (see Table 9.1). The truth table corresponding to the logic Eqs. (9.11) and (9.12) is given in Table 9.8.

This machine can be logically interpreted in two different ways. The first approach is to note that the machine's output monitors the direction of the change of state as induced by the input. The output is 1 if the pulse induces the logical change of state is $0 \rightarrow 1$. For the change $1 \rightarrow 0$ there is no output. Viewed in this manner [104], the machine is a flip-flop because it maintains a binary state until directed by the input to switch state. Specifically, the machine is similar to a T flip-flop [136] because a single input toggles the state. Flip-flops are key components as they provide a memory element for storing one bit. The data in Table 9.8 show that the state indeed flips, but the machine has no provision for knowing what is the present state of the machine. As discussed above, knowledge of the state of the machine can be readily implemented by applying two SP pulses – one to interrogate the state and one to restore the machine to its initial state. This is in the sense that the machine is endowed with memory.

Another way in which to view the machine represented by Eqs. [9.11–9.12] and Table 9.8 is to see it as a half subtractor, where the minuent digit x is encoded in the state of the machine and the subtrahend γ as a SP pulse, so that the machine computes $x - \gamma$. In a half subtractor, the difference is given by the XOR of the two digits (so that it is equivalent to the sum) but instead of a carry a borrow is needed, which is given by the INH function (see Table 9.1).

$$diff = x \oplus \gamma \tag{9.13}$$

$$borrow = x \otimes \bar{\gamma} \tag{9.14}$$

Therefore, the state at $t + 1$ [Eq. (9.11)] gives the difference while the output [Eq. (9.12)] gives the borrow. Another way of implementing a half subtractor is discussed in Ref. [96], where the initial convention of level $1 \equiv 0$ and level $3 \equiv 1$ is maintained but the input is "reverse" and is now a PS pulse. It can be readily checked that by encoding x in the state and γ as a PS pulse, Eqs. [9.11–9.12] are obtained for the next state and for the output.

There are two ways to implement a full subtractor (see Ref. [96] for details). The first method is by combining two half subtractors, along the lines used for the full adder discussed above. The other method is more interesting because it closely mimics the

implementation of the full adder, which means that the same logic device can be used, either to add or to subtract. This is what is meant by the ability to program a molecule: the same set of levels and of inputs can be used to implement different logic operations.

9.3.2 Finite-State Machines by Electrical Addressing

Until now, only the implementation of combinational circuits and finite-state machines in the gas phase or in solution have been discussed. Here, attention is focused on logic machine implementations on QD arrays. Because of the confinement induced by their nanometer size, QDs have also discrete quantum states but otherwise they are closer to solid-state devices. This is particularly the case for lithographic QDs where the confinement is induced by external voltages that confine electrons in a finite region embedded in a solid-state semi-conductor layer with a high dielectric constant. In this case, the electrons in the QD behave in good approximation as a 2-D electron gas[151]. At this point, interested is centered on a more "chemical" form of QDs – that is, metallic or semiconducting nanosize clusters passivated by organic ligands, for example thioalkane chains. The role of the ligands is to prevent aggregation of the colloidal nanoparticles and to ensure confinement of the electrons in the nanocluster. These QDs are prepared using a wet chemical method, and typically present a size dispersion of at best 5 to 10% in diameter of the cluster. When the size dispersion is narrow enough, they can self-assemble into ordered chains or arrays, and in that sense that they become closer to solid-state devices. They behave in many respects like artificial atoms [152–155] and can be used to make artificial solids [156–161]. When an ordered domain [105] or a chain of QDs [40]can be tethered between electrodes, they can be electrically addressed and probed. This is this type of arrangement used for implementing logic.

Although the details of the system matter a great deal, when discussing the principle of operation of the first logic implementation the only observation needed is that there can be one or more discrete level(s) that can be accessed by varying the electric potential across the dot. The second example discussed in this subsection is built on a system of coupled QDs, and such assemblies have been realized experimentally [162].

Initially, a three-terminal device is considered (see Figure 9.3) so that both a source-drain voltage, V_{sd}, can be applied across the system, and a gate voltage, V_g, in the perpendicular direction. Advantage is taken of the discrete level structure of the QD tethered between the three electrodes of the device to perform more complex operations at the hardware level than is usually done on a transistor.

The implementation of a set–reset finite-state machine is discussed in detail at this point. This is a machine with two logical states, that can accept two inputs, a set input and a reset input. The role of the set input is to bring the machine to logical state 1 if it was in logical state 0, and to do nothing if it is already in state 1. The role of the reset input is to bring the machine back to logical state 0 if it was in state 1 and to do nothing

Table 9.9 Operation of a set–reset machine.

Present state	Set input	Reset input	Name of action	Next State
0	0	0	No change	0
0	1	0	set	1
1	1	0	set	1
1	0	0	No change	1
1	0	1	reset	0
0	0	1	reset	0

if it is already in 0. The case where the two inputs are both 1 is not defined. The operation of the set–reset machine is summarized in Table 9.9, and a state diagram is shown in Figure 9.12.

Here, a single QD tethered in a three-terminal device is considered, that is submitted to a source-drain and to a gate bias. Its discrete level structure is described using the "orthodox" theory [163–165], which assumes that discrete level structure of the QD is due solely to quantization of charge on the dot. The one-electron level spacing of the dot is assumed to be continuous because it is much smaller than the change in electrostatic energy of the QD that occurs when an electron is added to or removed from it by varying the source-drain or the gate bias. The electrostatic energy of a QD with N electrons in a three-terminal device is given by

$$U_N = \frac{Q^2}{2C_T} = \frac{N^2e^2}{2C_T} + \frac{Ne}{C_T}\sum_i C_iV_i + \frac{1}{2C_T}\left(\sum_i C_iV_i\right)^2 \tag{9.15}$$

where Q the effective charge on the dot is given by

$$Q = C_T\Phi = Ne + \sum_{i=l,r,g} C_iV_i \tag{9.16}$$

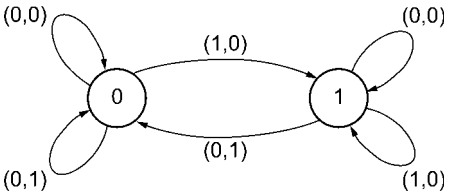


Figure 9.12 State diagram of a set–reset machine. The two possible values of the logical state of the machine are represented by the two circles denoted as 0 or 1. The arrows show the state changing transitions induced by the inputs. The inputs are given next to the arrows as (set,reset).

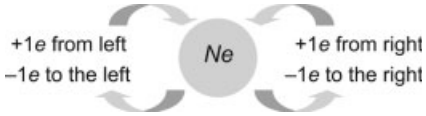


Figure 9.13 Electron transfer to/from the left and right electrode possible for a *N* QD in a three-terminal device.

In Eqs. (9.16) and (9.15), Φ is the electrostatic potential, C_T is the total capacitance of the system ($C_T = C_l + C_r + C_g$ where $C_{l,r}$ are the capacitances of the junctions to the left and right electrodes), and C_g is the capacitance of the gate electrode. $V_{l,r}$ are the source and the drain voltages. For a given gate voltage, an electron will be transferred to the dot or will leave the dot to the left or the right electrode when one of its discrete level falls within the energy window opened by the source-drain bias, V_{sd}, which is the difference between the bias of the right and on the left electrode. As shown in Figure 9.13, if the dot possesses initially N electrons, there are therefore four resonance conditions for electron transfer to/from the left and the right electrodes.

The four resonance conditions are:

$$\begin{aligned}
\Delta E_{l\to QD} &= \frac{e}{C_T}\left(\frac{e}{2} + (Ne + C_g V_g)\right) - e\frac{V}{2} \\
\Delta E_{QD\to l} &= \frac{e}{C_T}\left(\frac{e}{2} - (Ne + C_g V_g)\right) + e\frac{V}{2} \\
\Delta E_{r\to QD} &= \frac{e}{C_T}\left(\frac{e}{2} + (Ne + C_g V_g)\right) + e\frac{V}{2} \\
\Delta E_{QD\to R} &= \frac{e}{C_T}\left(\frac{e}{2} - (Ne + C_g V_g)\right) - e\frac{V}{2}
\end{aligned} \tag{9.17}$$

where $V = V_{sd}$ and a symmetric junction is assumed so that $C_l = C_r$ and $V_l = V_r = V/2$. $\Delta E = U_N - U_{N\pm1}$ must be ≤ 0 for the process to be allowed. It is the free energy difference for adding or removing an electron to the QD. Note that when only the charge on the dot is quantized, $U_N - U_{N\pm1}$ varies linearly with the applied source-drain and gate bias. The threshold for transferring an electron is given by the resonance condition, $\Delta E = 0$, which allows stability maps to be drawn of the charged QD as a function the gate and the source-drain bias. A stability map for $N = 0, 1$ electrons on the dot is shown in Figure 9.14. The areas in gray are the zones where the number of electrons on the QD is stable.

In the "orthodox" theory [164] the rates of transfer from the QD to the source and the drain electrodes are given by

$$\Gamma = \frac{2}{e^2 R}\,\frac{-\Delta E}{1+\exp(\Delta E/kT)} \overset{T\to 0K}{\longrightarrow} \frac{2}{e^2 R}\,|\Delta E|\theta(-\Delta E) \tag{9.18}$$

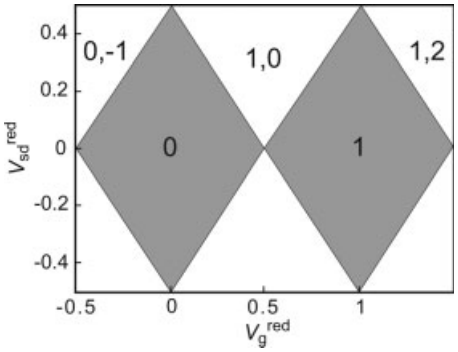


Figure 9.14 Stability map for a quantum dot with $N = 0$ and $N = 1$ electrons, plotted using Eq. (9.17) as a function of the V_g and V_{sd} in reduced units. $V_g^{red} = C_g V_g / e$ corresponds to the number of electrons on the dot, $V_g^{red} = V_{sd} C_T / 2e$.

where R is the resistance of the junction through which the electron passes, and is inversely proportional to the coupling between the QD and the electrode.

For implementation of the set–reset, two charges states of the QD are used, namely $N-1$ and N, where N is the number of extra electrons on the QD. For the simulation shown below, $N=0$ and $N=1$ were utilized. The logical state 0 of the set–reset machine was encoded as the QD with $N=0$ extra electrons, and the logical state 1 of the machine was encoded as the QD with $N=1$ extra electrons. From Figure 9.13, it can be seen that there are two rates for adding an electron to a QD with $N=0$, $\Gamma_{r\rightarrow QD}$ and $\Gamma_{l\rightarrow QD}$, and two rates for removing an electron from a $N=1$ QD, $\Gamma_{QD\rightarrow r}$ and $\Gamma_{QD\rightarrow l}$. Their analytical forms at $T=0$ K are

$$\begin{aligned}\Gamma_{QD\rightarrow l,l\rightarrow QD} &\propto \pm\left(\frac{e}{2C_T}+\frac{C_g V_g}{C_T}\right)\mp\frac{V}{2}\\ \Gamma_{QD\rightarrow r,r\rightarrow QD} &\propto \pm\left(\frac{e}{2C_T}+\frac{C_g V_g}{C_T}\right)\pm\frac{V}{2}\end{aligned} \tag{9.19}$$

These four rates are plotted in Figure 9.15 for a fixed gate voltage as a function of the source-drain bias, V.

In order for the set–reset machine to operate properly, a set voltage must be chosen such that the rate of transfer of an electron from the left electrode to the QD with $N=0$ is much larger than the rate for leaving the dot with $N=1$ to the right electrode, so that an electron is added to the dot and stays on the dot for a finite time. For the reset voltage, it is sufficient that the rate of leaving the dot with $N=1$ to the left electrode is significant. The rate $\Gamma_{r\rightarrow QD}$ corresponds to adding an electron to the QD with $N=0$. The operation of the set–reset device is more robust if the resistance of the right junction is much larger than that of the left one.

To check that the set–reset machine operates properly, the probability of getting an extra electron on the QD is monitored as a function of time while applying a time-dependent source-drain bias. By defining Q as the probability for having $N=1$ extra

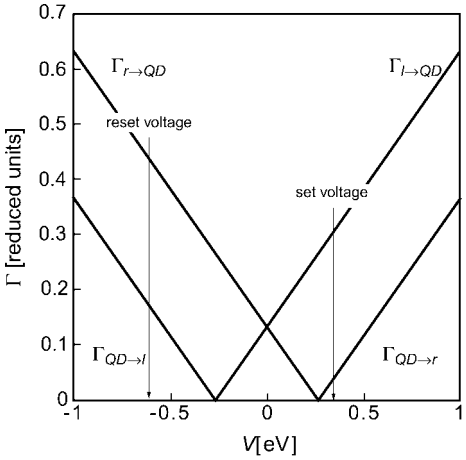


Figure 9.15 The four rates relevant for a QD with $N = 0$ and $N = 1$ electrons computed as a function of the source-drain bias V with a value of the gate voltage $V_g = -0.8$ V. $C_g = C_l = C_r = 0.2$ aF. The reduced units for Γ are $2/e^2R$, where R is the resistance of the junction. A symmetric junction is considered so that $R = R_l = R_r$. R is inversely proportional to the coupling between the QD and the left (right) electrode.

electrons on the dot, and P_l and P_r as the probabilities for this extra electron to be on the left and on right electrode, respectively, the following kinetic scheme is obtained:

$$\begin{aligned}
\frac{dP_l}{dt} &= \Gamma_{QD\to l}Q(t) - \Gamma_{l\to QD}P_l(t) \\
\frac{dQ}{dt} &= \Gamma_{l\to QD}P_l(t) + \Gamma_{r\to QD}P_r(t) - (\Gamma_{QD\to l} + \Gamma_{QD\to r})Q(t) \\
\frac{dP_r}{dt} &= \Gamma_{QD\to r}Q(t) - \Gamma_{r\to QD}P_r(t)
\end{aligned} \tag{9.20}$$

The time profile of the applied source-drain bias (a) and result of the integration of the kinetic scheme (b) are shown in Figure 9.16. The logical state 0 of the device is defined as $P_l \gg Q$, while state 1 is defined as $Q \gg P_l$. It can be seen that the effect of the set voltage is to fill the dot with one extra electron, whilst applying the reset pulse empties the dot of that extra electron. This shows that a single QD with two electrically addressable discrete levels can operate as a set–reset machine. It has also been shown that varying not only the source-drain but also the gate bias allows a full adder to be implemented on this system [106].

This subsection is concluded with a description of the implementation of another form of finite-state machine, a counter, on an array of two QDs anchored on a surface. This implementation is based on a scheme for addressing and/or reading the states of the dots electrically or optically that has been experimentally realized and characterized [162] (see Figure 9.17).

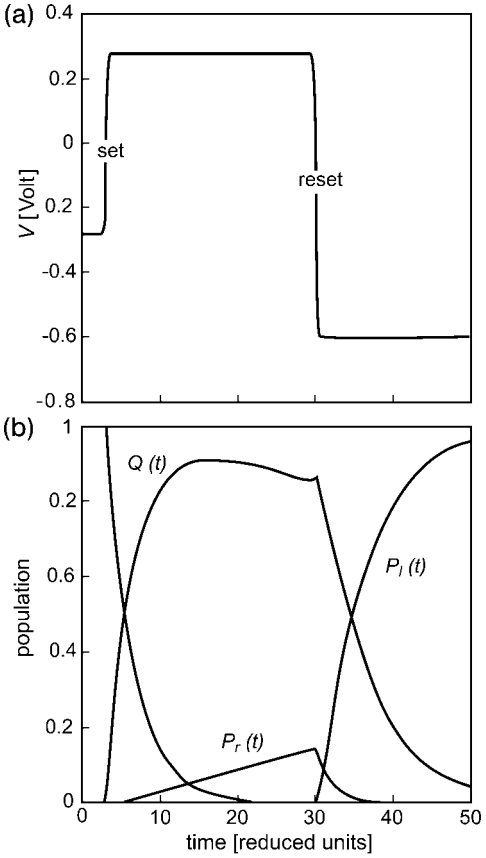


Figure 9.16 (a) Voltage–time profile and (b) time-dependent probability for the extra electron to be localized on the QD, $Q(t)$, on the left electrode, $P_l(t)$, and on the right electrode, $P_r(t)$. In Eq. (9.20), the rates are expressed in reduced units and the time is scaled correspondingly.

A counter [136] is a machine that is able to accept N inputs and to provide an output for every N inputs. The states of the counter are S_i, $i = 0, 1, 2, \ldots, N-1$ and transition can occur between successive states only when an input i, $i = 0, 1, 2, \ldots, N-1$, is received. After the count of N inputs, the state S_{N-1} is reached, an output is produced and the next input resets the counter to its initial state, S_0.

In the implementation based on the device shown in Figure 9.17 the index of the state is determined by the number of extra electrons on the Au QD. This number can be controlled optically or electrically (further details may be found in Refs. [137, 162]).

The counter functions as follows. First, the CdS QD is irradiated in a solution of TEA (10^{-2} M) so as to charge the Au QD. The irradiation is then stopped. The initial state, S_0, corresponds physically to the Au QD charged with four extra electrons. This number can be determined using surface plasmon resonance spectroscopy, by measuring the shift of the plasmon resonance of the gold surface due to the charging

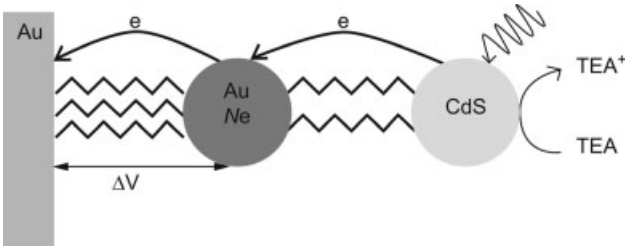


Figure 9.17 An experimentally realized [162] opto-electrically addressed array of two QDs anchored on a surface. The device can be used as a counter. The Au QDs are linked to the gold surface by a long-chain alkyl monolayer, and covalently to a semi-conducting CdS QD. The optical excitation of the CdS QD induces an electron transfer to the Au QD that is compensated by triethanolamine (TEA) present in the electrolyte solution. While the optical excitation is on, extra electrons accumulate on the Au QD and a potential drop is maintained across the junction between the Au QDs and the conductive gold surface. The charging of the Au QD can be optically monitored by changes in the resonance spectrum of the surface plasmon.

of the Au QD. On each occasion that an input is to be provided an input the index of the state must be incremented; this is done by decreasing the potential applied to the Au surface by a step sufficient to discharge an electron onto the surface. The magnitude of the required voltage drop is determined by the capacitance of the Au QD – that is, by the energy needed to charge or discharge the dot by one electron [137]. In the experiment the Au QD is passivated by a ligand, tiopronin, that has a high dielectric constant (16; see Ref. [162]), so that the charging energy is exceptionally low ($\approx$30 meV for Au QD of 2.3 ± 0.5 nm diameter). When the dot is fully discharged, after four voltage drops, S_4 is reached; this is the last state of the counter for which there are no extra electrons on the dot. At this point, the counter must be returned to the state S_0, so that it is available for the next counting cycle, and an output signal must then be provided. Unlike the usual system for counters, in this scheme the last input does not reset the counter to the state S_0. For this reason, even though for a dot with four extra electrons, there are five states – S_0, S_1, S_2, S_3, and S_4 – and modulo 4 is counted rather than modulo 5, the last step, $S_3 \rightarrow S_4$, is used to produce the output and reset the counter. The output is produced by monitoring the disappearance of the plasmon angle shift or by measuring the value of the surface potential. To reset the counter, the CdS dot is irradiated again. It should be noted that, in principle, the maximal number of four extra electrons on the dot is not a limitation, but up to 15 oxidation states of monolayer-protected Au QDs have been reported [166]. It is possible, therefore, to implement counters with a higher value of N.

9.4 Perspectives

The entire discussion in this chapter is based on the premise that there is a desire to design molecule-based logic circuits and not only switches. Results to date that have

been validated by *proof of concept* experiments, include:

- the implementation of *combinational circuits on a single molecule;*
- the *concatenation of logic operations*, whether performed on different molecules (intermolecular) or performed within the same molecule by communicating results carried out on different functional groups (intramolecular);
- the implementation of a *finite-state logic machine on a single molecule* and beyond that, *programming* of a single molecule; and
- using both electrical and optical addressing and readout with the advantage that it is not necessary to be able to address many states, because with two states a full adder can already be performed.

Technical studies in progress exploit these results towards increasing the logical capacity and depth (= number of switches) that can be implemented on a single molecule, or on a supramolecular assembly by the application to multifunctional group molecules where the intramolecular dynamics are used to concatenate the logical operations carried separately by the different groups. Next, in the order of integration is the assembly and concatenation of an array of molecules or an array of quantum dots.

Further studies are also needed to take even greater advantage of the large number of quantum states available, in a hierarchical order (electronic, backbone vibrations, torsions, rotations), which allows the processing in one cycle of far more information than a binary (classical or quantum) gate and, in the same direction, the use of more sophisticated optical and electrical inputs and readouts.

The first results to reach the level of technological implementation will most likely be the use of a single molecule not as a switch but rather as a combinational circuit. This will likely happen in the context of the architecture of a 2-D array cross-bar, which is the favored device geometry as foreseen by Hewlett Packard and others. However, even this progress will take time before it becomes a technology. The essential difference to be advocated is that at each node is placed not a switch but a single molecule acting as the equivalent of an entire network of switches. The very fast logic is conducted within the node, but the slower, wire-mediated communication between the nodes will remain. In the second round, communication between the nodes will be carried out by concatenation through self-assembly of the array using molecular recognition. Part of this endeavor is to achieve realistic programming abilities with special reference to selective intramolecular dynamics.

The key further breakthroughs that are currently required include:

- The design of molecular logic circuits that can be cycled reliably many times, and to explore whether this can be done using all-optical schemes.
- Input/output operations that reduce dissipation and allow fan-out and macroscopic interface, with special reference to the use of pulse shaping, electrical read/write and integrate storage within the logic unit.
- Beyond what is already available, it will be necessary to improve concatenation in order to reduce not only the need for cycling but also for interfacing with the macroscopic world. This will in turn lead to a need for molecular systems with special reference to devices on surfaces and their application as logic units.

Acknowledgments

These studies were supported by the EC FET-Open project MOLDYNLOGIC, the US-Israel Binational Science Foundation, BSF, Jerusalem, Israel and the EC NoE FAME.

References

1 C. Joachim, J. K. Gimzewski, A. Aviram, *Nature* 2000, **408**, 541.
2 C. P. Collier, E. W. Wong, M. Belohradsk, F. M. Raymo, J. F. Stoddart, P. J. Kuekes, R. S. Williams, J. R. Heath, *Science* 1999, **285**, 391.
3 C. P. Collier, G. Mattersteig, E. W. Wong, Y. Luo, K. Beverly, J. Sampaio, F. M. Raymo, J. F. Stoddart, J. R. Heath, *Science* 2000, **289**, 1172.
4 P. R. Ashton, R. Ballardini, V. Balzani, A. Credi, K. R. Dress, E. Ishow, C. J. Kleverlaan, O. Kocian, J. A. Preece, N. Spencer, J. F. Stoddart, M. Venturi, S. Wenger, *Chem-Eur. J.* 2000, **6**, 3558.
5 M. A. Reed, J. M. Tour, *Sci. Am.* 2000, **282**, 86.
6 R. M. Metzger, *Acc. Chem. Res.* 1999, **32**, 950.
7 R. M. Metzger, *J. Mater. Chem.* 2000, **10**, 55.
8 A. P. de Silva, N. D. McClenaghan, *J. Am. Chem. Soc.* 2000, **122**, 3965.
9 A. P. de Silva, Y. Leydet, C. Lincheneau, N. D. McClenaghan, *J. Phys. - Cond. Mater.* 2006, **18**, S1847.
10 Y. Luo, C. P. Collier, J. O. Jeppesen, K. A. Nielsen, E. Delonno, G. Ho, J. Perkins, H.-R. Tseng, T. Yamamoto, J. F. Stoddart, J. R. Heath, *ChemPhysChem* 2002, **3**, 519.
11 V. Balzani, A. Credi, M. Venturi, *ChemPhysChem* 2003, **4**, 49.
12 J. M. Tour, *Molecular Electronics*, World Scientific, River Edge, USA, 2003.
13 T. Nakamura, *Chemistry of Nanomolecular Systems: Towards the Realization of Molecular Devices*, Volume 70, Springer, Berlin, 2003.
14 F. M. Raymo, *Adv. Mater.* 2002, **14**, 401.
15 F. M. Raymo, M. Tomasulo, *Chem.- Eur. J.* 2006, **12**, 3186.
16 R. Waser, *Nanoelectronics and Information Technology*, Wiley-VCH, Weinheim, 2003.
17 M. A. Nielsen, I. L. Chuang, *Quantum Computation and Quantum Information*, Cambridge University Press, Cambridge, 2000.
18 C. H. Bennett, D. P. DiVincenzo, *Nature* 2000, **404**, 247.
19 D. Deutsch, *Proc. R. Soc. Lond.* 1985, **A 400**, 97.
20 N. Gershenfeld, I. L. Chuang, *Sci. Am.* 1998, **278**, 66.
21 S. J. Glaser, T. Schulte-Herbruggen, M. Sieveking, O. Schedletzky, N. C. Nielsen, O. W. Sorensen, C. Griesinger, *Science* 1998, **280**, 421.
22 A. Steane, *Nature* 2003, **422**, 387.
23 D. R. Glenn, D. A. Lidar, V. A. Apkarian, *Mol. Phys.* 2006, **104**, 1249.
24 M. A. Nielsen, M. R. Dowling, M. Gu, A. C. Doherty, *Phys. Rev. A* 2006, 73.
25 D. Bouwmeester, A. Ekert, A. Zeilinger, *The Physics of Quantum Information*, Springer, Berlin, 2000.
26 S. Haykin, *Neural Networks*, Prentice-Hall, Upper Saddle River, 1999.
27 J. Chen, D. H. Wood, *Proc. Natl. Acad. Sci. USA* 2000, **97**, 1328.
28 Y. Benenson, R. Adar, T. Paz-Elizur, Z. Livneh, E. Shapiro, *Proc. Natl. Acad. Sci. USA* 2003, **100**, 2191.
29 M. N. Stojanovic, D. Stefanovic, *J. Am. Chem. Soc.* 2003, **125**, 6673.
30 D. Margulies, G. Melman, C. E. Felder, R. Arad-Yellin, A. Shanzer, *J. Am. Chem. Soc.* 2004, **126**, 15400.

31 A. Okamoto, K. Tanaka, I. Saito, *J. Am. Chem. Soc.* 2004, **126**, 9458.

32 P. D. Tougaw, C. S. Lent, *J. Appl. Phys.* 1994, **75**, 1818.

33 A. O. Orlov, I. Amlani, G. H. Bernstein, C. S. Lent, G. L. Snider, *Science* 1997, **277**, 928.

34 I. Amlani, A. O. Orlov, G. Toth, G. H. Bernstein, C. S. Lent, G. L. Snider, *Science* 1999, **284**, 289.

35 L. Boni, M. Gattobigio, G. Iannaccone, M. Macucci, *J. Appl. Phys.* 2002, **96**, 3169.

36 H. Qi, S. Sharma, Z. H. Li, G. L. Snider, A. O. Orlov, C. S. Lent, T. P. Fehlner, *J. Am. Chem. Soc.* 2003, **125**, 15250.

37 J. Twamley, *Phys. Rev. A* 2003, **67**, 052328.

38 D. L. Klein, R. Roth, A. K. L. Lim, A. P. Alivisatos, P. L. McEuen, *Nature* 1997, **389**, 699.

39 W. Liang, M. P. Shores, J. L. Long, M. Bockrath, H. Park, *Nature* 2002, **417**, 725.

40 D. N. Weiss, X. Brokmann, L. E. Calvet, M. A. Kastner, M. G. Bawendi, *Appl. Phys. Lett.* 2006, **88**, 143507.

41 A. Nitzan, M. A. Ratner, *Science* 2003, **300**, 1384.

42 J. R. Heath, M. A. Ratner, *Physics Today* 2003, **56**, 43.

43 A. H. Flood, J. F. Stoddart, D. W. Steuerman, J. R. Heath, *Science* 2004, **306**, 2055.

44 C. Joachim, M. A. Ratner, *Nanotechnology* 2004, **15**, 1065.

45 J. Fiurasek, N. J. Cerf, I. Duchemin, C. Joachim, *Physica E* 2004, **24**, 161.

46 S. Ami, M. Hliwa, C. Joachim, *Chem. Phys. Lett.* 2003, **367**, 662.

47 R. Stadler, S. Ami, M. Forshaw, C. Joachim, *Nanotechnology* 2002, **13**, 424.

48 A. Bezryadin, C. Dekker, *Appl. Phys. Lett.* 1997, **71**, 1273.

49 S. Karthauser, E. Vasco, R. Dittmann, R. Waser, *Nanotechnology* 2004, **15**, S122.

50 C. R. Barry, J. Gu, H. O. Jacobs, *Nano Lett.* 2005, **5**, 2078.

51 B. Lussem, L. Muller-Meskamp, S. Karthauser, R. Waser, *Langmuir* 2005, **21**, 5256.

52 J. J. Urban, D. V. Talapin, E. V. Shevchenko, C. B. Murray, *J. Am. Chem. Soc.* 2006, **128**, 3248.

53 R. P. Feynman, *Feynman Lectures on Computations, reprint with corrections*, Perseus Publishing, Cambridge, MA, 1999.

54 D. Deutsch, *Proc. R. Soc. Lond.* A 1989, **425**, 73.

55 D. P. DiVincenzo, *Proc. R. Soc. Lond. A* 1998, **454**, 261.

56 R. Cleve, A. Ekert, C. Macchiavello, M. Mosca, *Proc. R. Soc. Lond. A* 1998, **454**, 339.

57 A. Ekert, R. Jozsa, *Proc. R. Soc. Lond. A* 1998, **356**, 1769.

58 R. Jozsa, *Proc. R. Soc. Lond. A* 1998, **454**, 323.

59 D. Loss, D. P. DiVincenzo, *Phys. Rev. A* 1998, **57**, 120.

60 G. Burkard, D. Loss, D. P. DiVincenzo, *Phys. Rev. B* 1999, **59** 2070.

61 K. R. Brown, D. A. Lidar, K. B. Whaley, *Phys. Rev. A* 2001, **65**, 012307.

62 C. H. Bennett, *IBM J. Res.* 1973, **17**, 525.

63 C. H. Bennett, *Int. J. Theoret. Phys.* 1982, **21**, 905.

64 D. Cory, A. Fahmy, T. Havel, *Proc. Natl. Acad. Sci. USA* 1997, **94**, 1634.

65 L. K. Grover, in Proceedings 28th ACM Symposium on the Theory of Computing, 1996.

66 D. Deutsch, R. Jozsa, *Proc. R. Soc. Lond. A* 1992, **439**, 553.

67 L. K. Grover, *Phys. Rev. Lett.* 1997, **79**, 4709.

68 T. Tulsi, L. K. Grover, A. Patel, *Quant. Inf. Comp.* 2006, **6**, 483.

69 C. H. Bennett, F. Bessette, G. Brassard, L. Slavail, J. Smolin, *J. Crypt.* 1992, **5**, 3.

70 P. W. Shor, in: S. Goldwasser (Ed.), Proceedings, 35th Annual Symposium on the Foundations of Computer Science, IEEE Computer Society Press, Los Alamitos, CA, 1994.

71 A. Ekert, R. Jozsa, *Rev. Mod. Phys.* 1996, **68**, 733.

72 J. Brown, *The Quest for Quantum Computer*, Simon & Schuster, New York, 2001.

73 J. Preskill, *Proc. R. Soc. Lond. A* 1998, **454**, 385.

74 J. Preskill, *Physics Today* 1999, **52**, 24.

75 E. Knill, R. Laflamme, L. Viola, *Phys. Rev. Lett.* 2000, **84**, 2525.

76 D. P. DiVincenzo, D. Bacon, J. Kempe, G. Burkard, K. B. Whaley, *Nature* 2000, **408**, 339.

77 D. Bacon, K. R. Brown, K. B. Whaley, *Phys. Rev. Lett.* 2001, **87**, 247902.

78 E. S. Myrgren, K. B. Whaley, *Quant. Inf. Proc.* 2004, **2**, 309.

79 G. Schaller, S. Mostame, R. Schutzhold, *Phys. Rev. A* 2006, 73.

80 F. Remacle, R. D. Levine, *Proc. Natl. Acad. Sci. USA* 2004, **101**, 12091.

81 A. Aviram, M. A. Ratner, *Chem. Phys. Lett.* 1974, **29**, 277.

82 J. C. Ellenbogen, J. C. Love, *Proc. IEEE* 2000, **88**, 386.

83 C. S. Lent, B. Isaksen, M. Lieberman, *J. Am. Chem. Soc.* 2003, **125**, 1056.

84 A. Credi, V. Balzani, S. J. Langford, J. F. Stoddart, *J. Am Chem. Soc.* 1997, **119**, 2679.

85 A. P. de Silva, I. M. Dixon, H. Q. N. Gunaratne, T. Gunnlaugsson, P. R. S. Maxwell, T. E. Rice, *J. Am. Chem. Soc.* 1999, **121**, 1393.

86 V. Balzani, A. Credi, F. M. Raymo, J. F. Stoddart, *Angew. Chem. Int. Ed.* 2000, **39**, 3349.

87 A. P. de Silva, N. D. McClenaghan, *Chem. Eur. J.* 2004, **10**, 574.

88 F. M. Raymo, R. J. Alvarado, S. Giordani, M. A. Cejas, *J. Am. Chem. Soc.* 2003, **125**, 2361.

89 G. J. Brown, A. P. de Silva, S. Pagliari, *Chem. Commun.* 2002, 2461.

90 X. F. Guo, D. Q. Zhang, G. X. Zhang, D. B. Zhu, *J. Phys. Chem. B* 2004, **108**, 11942.

91 J. Andreasson, G. Kodis, Y. Terazono, P. A. Liddell, S. Bandyopadhyay, R. H. Mitchell, T. A. Moore, A. L. Moore, D. Gust, *J. Am. Chem. Soc.* 2004, **126**, 15926.

92 R. Baron, O. Lioubashevski, E. Katz, T. Niazov, I. Willner, *Angew. Chem.* 2006, **45**, 1572.

93 K. Szacilowski, W. Macyk, G. Stochel, *J. Am. Chem. Soc.* 2006, **128**, 4550.

94 F. Remacle, S. Speiser, R. D. Levine, *J. Phys. Chem. A* 2001, **105**, 5589.

95 F. Remacle, E. W. Schlag, H. Selzle, K. L. Kompa, U. Even, R. D. Levine, *Proc. Natl. Acad. Sci. USA* 2001, **98**, 2973.

96 F. Remacle, R. D. Levine, *Phys. Rev. A* 2006, **73**, 033820.

97 H. Lederman, J. Macdonald, D. Stefanovic, M. N. Stojanovic, *Biochemistry* 2006, **45**, 1194.

98 D. Margulies, G. Melman, A. Shanzer, *J. Am. Chem. Soc.* 2006, **128**, 4865.

99 F. Remacle, J. R. Heath, R. D. Levine, *Proc. Natl. Acad. Sci. USA* 2005, **102**, 5653.

100 K. L. Kompa, R. D. Levine, *Proc. Natl. Acad. Sci. USA* 2001, **98**, 410.

101 F. Remacle, R. D. Levine, *J. Chem. Phys.* 2001, **114**, 10239.

102 T. Witte, C. Bucher, F. Remacle, D. Proch, K. L. Kompa, R. D. Levine, *Angew. Chem.* 2001, **40**, 2512.

103 F. Remacle, R. Weinkauf, D. Steinitz, K. L. Kompa, R. D. Levine, *Chem. Phys.* 2002, **281**, 363.

104 D. Steinitz, F. Remacle, R. D. Levine, *ChemPhysChem.* 2002, **3**, 43.

105 F. Remacle, K. C. Beverly, J. R. Heath, R. D. Levine, *J. Phys. Chem. B* 2003, **107**, 13892.

106 F. Remacle, J. R. Heath, R. D. Levine, *Proc. Natl. Acad. Sci. USA* 2005, **102**, 5653.

107 F. Remacle, R. D. Levine, *Faraday Disc.* 2006, **131**, 46.

108 A. W. Ghosh, T. Rakshit, S. Datta, *Nano Lett.* 2004, **4**, 565.

109 X. Li, B. Xu, X. Xiao, X. Yang, L. Zang, N. Tao, *Faraday Disc.* 2006, **131**, 111.

110 S. L. Hurst, *IEEE Trans. Comp.* 1984, **C-33**, 1160.

111 D. C. Rine, *Computer Science and Multiple-Valued Logic*, North-Holland, Amsterdam, 1977.

112 Wikipedia, http://en.wikipedia.org/wiki/Multi-valued_logic 2006.

113 G. C. Schatz, M. A. Ratner, *Quantum Mechanics in Chemistry*, Prentice-Hall, New York, 1993.

114 M. M. Mano, C. R. Kime, *Logic and Computer Design Fundamentals*, Prentice-Hall, Upper Saddle River, NJ, 2000.

115 F. Remacle, R. Weinkauf, R. D. Levine, *J. Phys. Chem. A* 2006, **110**, 177.

116 W. Cheng, N. Kuthirummal, J. Gosselin, T. I. Solling, R. Weinkauf, P. Weber, *J. Phys. Chem. A* 2005, **109**, 1920.

117 L. Lehr, T. Horneff, R. Weinkauf, E. W. Schlag, *J. Phys. Chem. A* 2005, **109**, 8074.

118 R. Weinkauf, L. Lehr, A. Metsala, *J. Phys. Chem. A* 2003, **107**, 2787.

119 R. B. Bernstein, *J. Phys. Chem.* 1982, **86**, 1178.

120 R. B. Bernstein, *Chemical Dynamics via Molecular Beam and laser techniques*, Oxford University Press, New York, 1982.

121 G. Wiswanath, M. Kasha, *J. Chem. Phys.* 1956, **24**, 574.

122 M. Beer, H. C. Longuett-Higgins, *J. Chem. Phys.* 1955, **23**, 1390.

123 J. W. Sidman, D. S. McClure, *J. Chem. Phys.* 1955, **24**, 757.

124 S. Speiser, *Chem. Rev.* 1996, **96** 1953.

125 I. Kaplan, J. Jortner, *Chem. Phys. Lett.* 1977, **52**, 202.

126 I. Kaplan, J. Jortner, *Chem. Phys.* 1978, **32**, 381.

127 S. Speiser, *Appl. Phys. B* 1989, **49**, 109.

128 S. Speiser, N. Shakkour, *Appl. Phys. B* 1985, **38**, 191.

129 M. Orenstein, S. Kimel, S. Speiser, *Chem. Phys. Lett.* 1978, **58**, 582.

130 N. Lokan, M. N. Paddow-Row, T. A. Smith, M. LaRosa, K. P. Ghiggino, S. Speiser, *J. Am. Chem. Soc.* 1999, **121**, 2917.

131 S. Speiser, F. Schael, *J. Mol. Liq.* 2000, **86**, 25.

132 S. Speiser, *Opt. Commun.* 1983, **45**, 84.

133 U. Peskin, M. Abu-Hilu, S. Speiser, *Optical Mater.* 2003, **24**, 23.

134 S. Speiser, *J. Luminescence* 2003, **102**, 267.

135 T. L. Booth, *Sequential Machines and Automata Theory*, Wiley, New York, 1968.

136 Z. Kohavi, *Switching and Finite Automata Theory*, Tata McGraw-Hill, New Delhi, 1999.

137 F. Remacle, I. Willner, R. D. Levine, *ChemPhysChem.* 2005, **6**, 1.

138 J. Martin, B. W. Shore, K. Bergmann, *Phys. Rev. A* 1996, **54**, 1556.

139 T. Halfmann, K. Bergmann, *J. Chem. Phys.* 1996, **104**, 7068.

140 A. Kuhn, S. Steuerwald, K. Bergmann, *Eur. Phys. J. D* 1998, **1**, 57.

141 B. W. Shore, *The Theory of Coherent Atomic Excitation: Multilevel Atoms and Incoherence*, Wiley, New York, 1990.

142 B. W. Shore, K. Bergmann, J. Oreg, S. Rosenwaks, *Phys. Rev. A* 1991, **44**, 7442.

143 N. V. Vitanov, B. W. Shore, K. Bergmann, *Eur. Phys. J. D* 1998, **4**, 15.

144 V. S. Malinovsky, D. J. Tannor, *Phys. Rev. A* 1997, **56**, 4929.

145 S. A. Rice, M. Zhao, *Optical Control of Molecular Dynamics*, Wiley, New York, 2000.

146 C. Cohen-Tannoudji, J. Dupont-Roc, G. Grynberg, *Atom-Photon Interactions*, Wiley, New York, 1992.

147 K. Bergmann, H. Theuer, B. W. Shore, *Rev. Mod. Phys.* 1998, **70**, 1003.

148 D. J. Tannor, *Introduction to Quantum Mechanics: A Time Dependent Perspective*, University Science Books, Sausalito, CA, 2005.

149 N. V. Vitanov, M. Fleischhauer, B. W. Shore, K. Bergmann, *Adv. At. Mol. Opt. Phys.* 2001, **46**, 55.

150 M. P. Fewell, B. W. Shore, K. Bergmann, *Aust. J. Phys.* 1997, **50**, 281.

151 L. P. Kouwenhoven, D. G. Austing, S. Tarucha, *Rep. Prog. Phys.* 2001, **64**, 701.

152 R. C. Ashoori, *Nature* 1996, **379**, 413.

153 M. A. Kastner, *Physics Today* 1993, **46**, 24.

154 U. Banin, Y. W. Cao, D. Katz, O. Millo, *Nature* 1999, **400**, 542.

155 M. Reed, *Sci. Am.* 1993, **268**, 98.

156 C. P. Collier, R. J. Saykally, J. J. Shiang, S. E. Henrichs, J. R. Heath, *Science* 1997, **277** 1978.

157 C. P. Collier, T. Vossmeyer, J. R. Heath, *Annu. Rev. Phys. Chem.* 1998, **49**, 371.

158 C. J. Kiely, J. Fink, J. G. Zheng, M. Brust, D. Bethell, D. J. Schiffrin, *Adv. Mater.* 2000, **12**, 640.

159 G. Markovich, C. P. Collier, S. E. Henrichs, F. Remacle, R. D. Levine, J. R. Heath, *Acc. Chem. Res.* 1999, **32**, 415.

160 G. Schmid, U. Simon, *Chem. Commun.* 2005, 697.

161 A. Taleb, V. Russier, A. Courty, M. P. Pileni, *Phys. Rev. B* 1999, **59**, 13350.

162 M. Zayats, A. B. Kharitonov, S. P. Pogorelova, O. Lioubashevski, E. Katz, I. Willner, *J. Am. Chem. Soc.* 2003, **125**, 16006.

163 R. I. Shekhter, *Sov.-Phys. JETP* 1973, **36**, 747.

164 D. V. Averin, A. N. Korotkov, K. K. Likharev, *Phys. Rev. B* 1991, **44**, 6199.

165 I. O. Kulik, R. I. Shekhter, *Sov. Phys. JETP* 1975, **41**, 308.

166 B. M. Quinn, P. Liljeroth, V. Ruitz, T. Laaksonen, K. Kontturi, *J. Am. Chem. Soc.* 2003, **125**, 6644.

II
Architectures and Computational Concepts

Nanotechnology. Volume 4: Information Technology II. Edited by Rainer Waser

ISBN: 978-3-527-31737-0

10
A Survey of Bio-Inspired and Other Alternative Architectures

Dan Hammerstrom

10.1
Introduction

Since the earliest days of the electronic computer, there has always been a small group of people who have seen the computer as an extension of biology, and have endeavored to build computing models and even hardware that are inspired by, and in some cases are direct copies of, biological systems. Although biology spans a wide range of systems, the primary model for these early efforts has been neural circuits. Likewise, in this chapter the discussion will be limited to neural computation.

Several examples of these early investigations include McCulloch-Pitts *Logical Calculus of Nervous System Activity* [2], Steinbuch's *Die Lernmatrix* [3], and Rosenblatt's *Perceptron* [4]. At the same time, an alternate approach to intelligent computing, *Artificial Intelligence* (AI), that relied on higher-order symbolic functions, such as structured and rule-based representations of knowledge, began to demonstrate significantly greater success than the neural approach. In 1969, Minsky and Papert [5] of the Massachusetts Institute of Technology published a book that was critical of the then current "bio-inspired" algorithms, and which succeeded in eventually ending most research funding for that approach. Consequently, significant research funding was directed towards AI, and the field subsequently flourished. The AI approach, which relied on symbolic reasoning often represented by a first-order calculus and sets of rules, began to exhibit real intelligence, at least on toy problems. One reasonably successful application was the "expert" system, and there was even the development of a complete language, ProLog, dedicated to logical rule-based inference.

A few expert system successes were also enjoyed in actually fielded systems, such as Soar [6], the development of which was started by Alan Newell's group at Carnegie Mellon University. However, by the 1980s AI in general was beginning to lose its luster after 40 years of funding with ever-diminishing returns.

Since the 1960s, however, there have always been groups that continued to study biologically inspired algorithms, and two such projects – mostly as a result of their

Nanotechnology. Volume 4: Information Technology II. Edited by Rainer Waser

ISBN: 978-3-527-31737-0

being in the right place at the right time – had a huge impact which re-energized the field and led to an explosion of research and funding. The first project incorporated the investigations [7] of John Hopfield, a physicist at Caltech, who proposed a model of auto-associative memory based on physical principles such as the Ising theory of spin-glass. Although Hopfield nets were limited in capability and size, and others had proposed similar algorithms previously, Hopfield's formulation was both clean and elegant. It also succeeded in bringing many physicists, armed with sophisticated mathematical tools, into the field. The second project was the "invention" of the back-propagation algorithm by Rumelhart, Hinton, and Williams [8]. Although there too similar studies had been conducted previously [9], the difference with Rumelhart and colleagues was that they were cognitive scientists creating a set of techniques called parallel distributed processing (PDP) models of cognitive phenomena, where back-propagation was a part of a larger whole.

At this point, it would be useful to present some basic neuroscience, followed by details of some of the simpler algorithms inspired by this biology. This information will provide a strong foundation for discussing various biologically inspired hardware efforts.

10.1.1 Basic Neuroscience

In simplified terms, neural circuits consist of large numbers of parallel processing components, the *neurons*. These tend to be slow in operation, with typical switching times on the order of milliseconds, and consequently the brain uses significant parallelism rather than speed to perform its complex tasks. Adaptation comes in short- and long-term versions, and can result from a variety of complex interactions.

Although most neurons are exceptions to the canonical neuron shown in Figure 10.1, the neuron illustrated is sufficiently complex to demonstrate the basic principles. Via various ion channels, neurons maintain a constant negative voltage of approximately -70 mV relative to the ambient environment. This neuron consists of a *dendritic* tree for taking inputs from other neurons, a body or *soma*, which basically

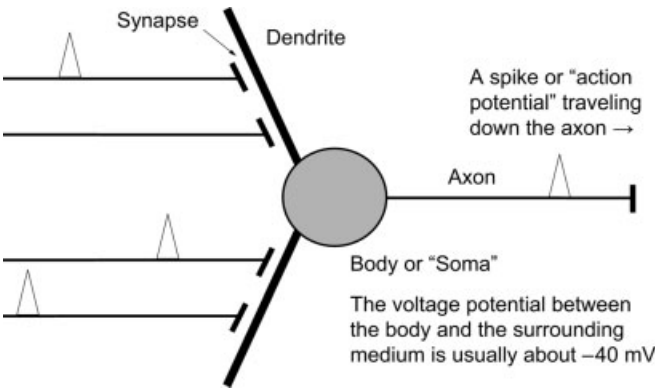


Figure 10.1 An abstract neuron.

performs charge summation, and an output, the *axon*. Inter-neuron communication is generally via pulses or spikes. Axons form synapses on dendrites and signal the dendrite by releasing small amounts of neurotransmitter, which is taken up by the dendrite.

Axons from other neurons connect via synapses onto the dendritic tree of each neuron. When an axon fires it releases a neurotransmitter into the junction between the *presynaptic* axon and the *postsynaptic* dendrite. The neurotransmitter causes the dendrite to depolarize slightly, and this charge differential eventually reaches the body or soma of the neuron, depolarizing the neuron.

When a neuron is sufficiently depolarized it passes a threshold which causes it to generate an *action potential*, or output spike, which moves down the axon to the next synapse. When an output spike is traveling down an axon, it is continuously regenerated allowing for arbitrary fan-out.

While the dendrites are depolarizing the neuron, the resting potential is slowly being restored, creating what is known as a "leaky integrator." Unless enough action potentials arrive within a certain time window of each other, the depolarization of the soma will not be sufficient to generate an output action potential.

In addition to accumulating signals and creating new signals, neurons also learn. When a spike arrives via an axon at a synapse, it "presynaptically" releases a neurotransmitter, which causes some depolarization of the postsynaptic dendrite. Under certain conditions, the effect of a single spike can be modified, typically increased or "facilitated", where it causes a greater depolarization at the synapse. When the effect that an action potential has on the postsynaptic neurons is enhanced, the synapse is said to be *potentiated*, and learning has occurred. One form of this is called long-term potentiation (LTP), as such potentiation has been shown to last for several weeks at a time and may possibly last much longer. LTP is one of the more common learning mechanisms, and has been shown to occur in several areas of the brain whenever the inputs to the neuron and the output of the neuron correlate within some time window. Learning correlated inputs and outputs is also called Hebb's law, named after Donald Hebb who proposed it in 1947. Synapses can also lose their facilitation; one example of this is a similar mechanism called long-term depression (LTD), which generally occurs when an output is generated and there is no input at a particular synapses.

Postsynaptic excitation can either be *excitatory* (an excitatory postsynaptic potential or EPSP), which leads to accumulating even more charge in the soma, or *inhibitory* (an inhibitory postsynaptic potential or IPSP), which tends to drain charge off of the soma, making it harder for a neuron to fire an action potential. Both capabilities are needed to control and balance the activation of groups of neurons. In one model, the first neuron that fires tends to inhibit the others in the group leading to what is called a "winner-take-all" function.

10.1.2 A Very Simple Neural Model: The Perceptron

One of the first biologically inspired models was the perceptron which, although very simple, was still based on biological neurons. The primary goal of a perceptron is to

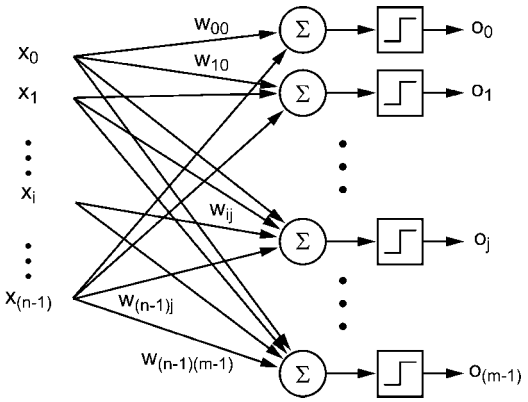


Figure 10.2 A single-layer perceptron.

do classification. Perceptron operation is very simple, as it has a one-dimensional synaptic weight vector and takes another, equal size, one-dimensional vector as input. Normally the input vector is binary and the weight vectors positive or negative integers. During training, a "desired" signal is also presented to the perceptron. If the output matches the training signal, then no action is taken; however, if the output is incorrect and does not match the training signal, then the weights in the weight vector are adjusted accordingly. The perceptron learning rule, which was one of the first instances of the "delta rule" used in most artificial neural models, incremented or decremented the individual weights depending on whether they were predictive of the output or not. A single-layer perceptron is shown in Figure 10.2. Basic perceptron operation is

$$o_j = f\left(\sum_{i=1}^{n} w_{ij}x_i\right) = f(W_j^T X)$$
$$O = f(W^T X)$$

where $f()$ is an activation function, and is a step function here (if $sum > 0$, then $f(sum) = 1$); however, $f()$ can also be a smooth function (see below). A "layer" has some number (two or more) of perceptrons, each with its own weight vector and individual output value, leading to a weight matrix and an output vector. In a single "layer" of perceptrons, each one sees the same input vector.

The Delta Rule, which is used during learning is,

$$\Delta w_{ij} = \alpha(d_j - o_j)x_i$$

where d_j is the desired output, and o_j is the actual output.

The delta rule is fundamental to most adaptive neural network algorithms. Rosenblatt proved that if a data set is linearly separable, the perceptron will eventually find a plane that separates the set. Figure 10.3 shows the two-dimensional (2-D) layout of data for, first, a linearly separable set of data and, second, for a non-linearly separable set. Unfortunately, if the data are not linearly separable the perceptron fails miserably, and this was the point of "...the book that killed neural networks",

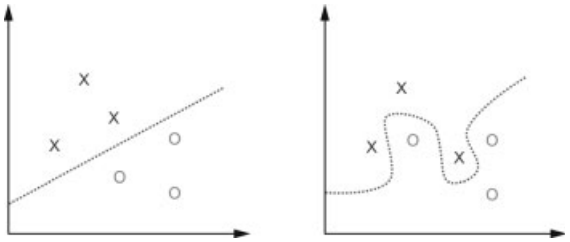


Figure 10.3 Linear (left) and non-linear (right) classification.

Perceptrons by Marvin Minsky and Seymour Papert (1968). Perceptrons cannot solve non-linearly separable problems; neither do they function in the kind of multiple layer structures that may be able to solve non-linear problems, as the algorithm is such that the output layer cannot tell the middle layer what its desired output should be. Attention is now turned to a description of the multi-layer perceptron.

10.1.3 A Slightly More Complex Neural Model: The Multiple Layer Perceptron

The invention of the multi-layer perceptron constituted, to some degree, a "victory" against the "evil empire" of symbolic computing. The Minsky and Papert book focused primarily, through numerous sophisticated examples, on how perceptrons, as envisioned at that time, could not solve non-linear problems, an excellent example being the XOR problem. Although there are a number of variations, back-propagation (BP) allowed the use of multiple, non-linear layers, and is sometimes referred to as a multi-layer perceptron (MLP).

Although the derivation of BP is beyond the scope of this chapter, some of the characteristics that allowed it to extend perceptron-like learning to multiple layers can be briefly summarized. First, instead of a discrete step function for output, a continuous activation function is used. As with the perceptron, there is also a training or "desired" signal, and by actually quantifying the error of the output as a function (generally least means square) an error surface is created. The gradient of the error surface can then be found and used to adjust the weights. By using the chain rule from calculus the error can then be back-propagated through the various levels.

In summary, the steps of the BP algorithm are:

- Present an input vector to the first layer of the network.
- Calculate the output for each layer, moving forward to network output.
- Calculate the error delta at the output (using actual output and the externally supplied target output, d).
- Use the computed error deltas at the output to compute the error deltas at the next layer, etc., moving backward through the net.
- After all error deltas have been computed for every node, use the delta rule to incrementally update all the weights in the network. (Note: the feedforward input activation values for each connection must be remembered.)

Even two-level networks can approximate complex, non-linear functions. Moreover, this technique generally finds good solutions, which are compact, leading to fast, feed-forward (non-learning) execution time. Although it has been shown to approximate Bayesian decisions (i.e. it results in a generally good estimate of where Baye's techniques would put the decision surface), it can have convergence problems due to many non-local minima. It is also computationally intensive, often taking days to train with complex large feature sets.

10.1.4 Auto-Association

Another important family of networks are associative networks, one example of which (as given above) is the Hopfield net. Here, details of a simple associative network developed by G. Palm [10] will be presented (this was in fact developed before the studies of Hopfield). One useful variation implements an *auto-associative network* that is an overly simplistic approximation of the circuits in the mammalian neocortex. Auto-associative networks have been studied extensively. In its simplest form, an associative memory maps an input vector to an output vector. When an input is supplied to the memory, its output is a "trained" vector with the closest match, assuming some metric, to the given input. Auto-association results when the input and output vectors are in the same space, with the input vector being a corrupted version of one of the training vectors. With *best-match association*, when a noisy or incomplete input vector is presented to the network, the "closest" training vector can usually be recalled reliably. In auto-association the output is fed back to the input, which may require several iterations to stabilize to a trained vector.

Here, it is assumed that the vectors are binary and the distance metric between two vectors is simply the number of bits that are different – that is, the Hamming distance.

The Palm associative network maps input vectors to output vectors, where the set of input vector to output vector mappings are noted as $\{(x^\mu, y^\mu), \mu = 1, 2, \ldots, M\}$. There are M mappings, and both x^μ and y^μ are binary vectors with size of m and n respectively. x^μ and y^μ are sparsely encoded, with $\sum_{i=1}^{m} x_i = l\,(l \ll m)$ and $\sum_{j=1}^{n} y_j = k$ $(k \ll n)$. Here, l and k are the numbers of active nodes (non-zero) in the input and output vectors, respectively. In training, the "clipped" Hebbian "outer-product," learning rule is generally used, and a binary weight matrix W is formed by $W = \vee_{\mu=1}^{M}[y^\mu \cdot (x^\mu)^T]$. Such batch computation has the weights computed off-line and then down-loaded into the network. It is also possible to learn the weights adaptively [11].

During recall, a noisy or incomplete input vector $\tilde{x}$ is applied to the network, and the network output is computed by $\tilde{y} = f(W \cdot \tilde{x} - \theta)$, θ is a global threshold, and $f()$ is the Heaviside step function, where an output node will be 1 (active) if its dendritic sum $x_i = \sum_{j=1}^{m} w_{ij}\tilde{x}_j$ is greater than the threshold θ; otherwise, it is 0. To set the threshold, the "k winners take all (k-WTA) rule" is used, where k is the number of active nodes in an output vector. The threshold, θ, is set so that only those nodes that have the k maximum dendritic sums are set to "1", and the remaining nodes are set to "0". The k-WTA threshold is very close to the minimum error threshold. The k-WTA

operation plays the role of competitive lateral inhibition, which is a major component in all cortical circuits. In the BCPNN model of Lansner and his group [11], the nodes are divided into hypercolumns, typically $\sqrt{N}$ nodes in each of the $\sqrt{N}$ columns, with 1-WTA being performed in each column.

An auto-associative network starts with the associative model just presented and feeds the output back to the input, so that the x and y are in the same vector space and $l = k$. This auto-associative model is called an *attractor model* in that its state space creates an energy surface with most minima ("attractor basins") occurring when the state is equal to a training vector. Under certain conditions, given an input vector x', then the output vector y that has the largest conditional probability $P(x'|y)$ is the most likely training vector in a Bayesian sense. It is possible to define a more complex version with variable weights, as would be found during dynamic learning, which also allows the incorporation of prior probabilities [12].

10.1.5
The Development of Biologically Inspired Hardware

With BP and other non-linear techniques in hand, research groups began to solve more complex problems. Concurrent to this there was an explosion in neuroscience that was enabled by high-performance computing and sophisticated experimental technologies, coupled with an increasing willingness in the neuroscience community to begin to speculate about the function of the neural circuits being studied. As a result, research into artificial neural networks (ANNs) of all types gained considerable momentum during the late 1980s, continuing until the mid-1990s when the research results began to slow down. However, like AI before it – and fuzzy logic, which occurred concurrently – ANNs had trouble in scaling to solve the difficult problems in intelligent computing. Nevertheless, ANNs still constitute an important area of research, and ANN technologies play a key role in a number of real-world applications [13, 14]. In addition, they are responsible for a number of important breakthroughs.

During the heady years of the late 1980s and early 1990s, while many research groups were investigating theory, algorithms, and applications, others began to examine hardware implementation. As a consequence, there quickly evolved three schools of thought, though with imprecise dividing lines between them:

- The first concept was to build very specialized analog chips where, for the most part, the algorithms were hard-wired into silicon. Perhaps the best known was the aVLSI (low-power analog VLSI) technology developed by Carver Mead and his students at Caltech.

- The second concept was to build more general, highly parallel digital, but still fairly specialized chips. Many of the ANN algorithms were very computer-intensive, and it seemed that simply speeding up algorithm execution – and especially the learning phase – would be a big help in solving the more difficult problems and the commercialization of ANN technology. During the late 1980s and early 1990s these chips were also significantly faster than mainstream desktop technology; however, this second group of chips incorporated less biological realism than the analog chips.

- The third option was to use off-the-shelf hardware, digital signal processing (DSP) and media chips, and this ultimately was the winning strategy. This approach was successful because the chips were used in a broader set of applications and had manufacturing volume and software inertia in their favor. Their success was also assisted by Amdahl's law (see Section 10.2.1).

The aim of this chapter is to review examples of these biologically inspired chips in each of the main categories, and to provide detailed discussions of the motivation for these chips, the algorithms they were emulating, and architecture issues. Each of the general categories presented is discussed in greater detail as appropriate. Finally, with the realm of nano- and molecular-scale technology rapidly approaching, the chapter concludes with a preview of the future of biologically inspired hardware.

10.2 Early Studies in Biologically Inspired Hardware

The hardware discussed in this chapter is based on neural structures similar to those presented above, and, as such, is designed to solve a particular class of problems that are sometimes referred to as "intelligent computing". These problems generally involve the transformation of data across the boundary between the real world and the digital world, in essence from sensor readings to symbolic representations usable by a computer; indeed, this boundary has been called "the digital seashore".[1)] Such transformations are found wherever a computer is sampling and/or acting on real-world data. Examples include the computer recognition of human speech, computer vision, textual and image content recognition, robot control, optical character recognition (OCR), automatic target recognition, and so on. These are difficult problems to solve on a computer, as they require the computer to find *complex structures and relationships* in massive quantities of low-precision, ambiguous, and noisy data. These problems are also very important, and an inability to solve them adequately constitutes a significant barrier to computer usage. Moreover, the list of ideas has been exhausted, as neither AI, ANNs, fuzzy logic, nor Bayesian networks[2)] have yet enabled robust solutions.

At the risk of oversimplifying a complex family of problems, the solution to these problems will, somewhat arbitrarily, be partitioned into two domains: the "front end" and the "back end" (see Figure 10.4):

- *Front-end* operations involve more direct access to a signal, and include filtering and feature extraction.

1) Hiroshi Ishii, MIT Media Lab.

2) "A Bayesian network (or a belief network) is a probabilistic graphical model that represents a set of variables and their probabilistic independencies". Wikipedia, http://www.wikipedia.org/.

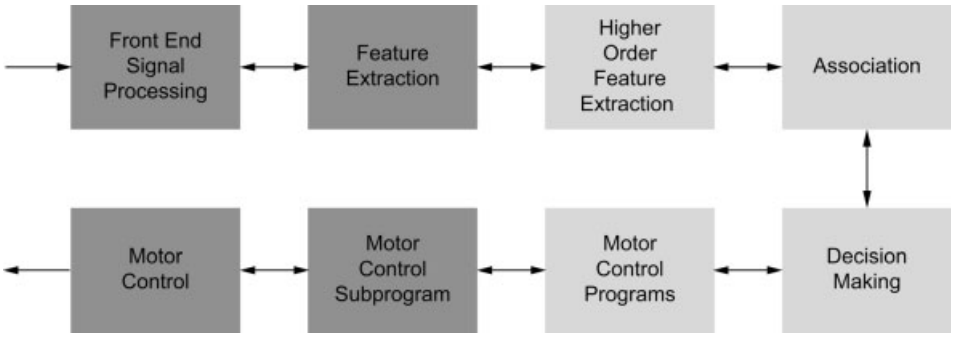


Figure 10.4 A canonical system.

- *Back-end* operations are more "intelligent", and include storing abstract views of objects or inter-word relationships.

In moving from front end to back end, the computation becomes increasingly interconnect driven, leveraging ever-larger amounts of diffuse data at the synapses for the connections. Much has been learned about the front end, where the data are input to the system and where there are developments in traditional as well as neural implementations. Whilst these studies have led to a useful set of tools and techniques, they have not solved the whole problem, and consequently more groups are beginning to examine the back-end – the realm of the cerebral cortex – as a source of inspiration for solving the remainder of the problem. Moreover, as difficult as the front-end problems are, the back-end problems are even more so. One manifestation of this difficulty is the "perception gap" discussed by Lazzaro and Wawrzynek [15], where the feature representations produced by more biologically inspired front-end processing are incompatible with existing back-end algorithms.

A number of research groups are beginning to refer to this "backend" as intelligent signal processing (ISP), which augments and enhances existing DSP by incorporating contextual and higher level knowledge of the application domain into the data transformation process. Simon Haykin (McMaster University) and Bart Kosko (USC) were editors of a special issue of the *Proceedings of the IEEE* [16] on ISP, and in their introduction stated:

> "*ISP uses learning and other 'smart' techniques to extract as much information as possible from signal and noise data.*"

If you are classifying at Baye's optimal rates and you are still not solving the problem, what do you do next? The solution is to add more knowledge of the process being classified to your classification procedures, which is the goal of ISP. One way to do this is to increase the contextual information (e.g. higher-order relationships such as sentence structure and word meaning in a text-based application) available to the algorithm. It is these complex, "higher-order" relationships that are so difficult for us to communicate to existing computers and, subsequently, for them to utilize efficiently when processing signal data.

Humans make extensive use of contextual information. We are not particularly good classifiers of raw data where little or no context is provided, but we are masters of leveraging even the smallest amount of context to significantly improve our pattern-recognition capabilities.

One of the most common contextual analysis techniques in use today is the Hidden Markov Model (HMM) [17]. An HMM is a discrete Markov model with a finite number of states, that can be likened to a probabilistic finite-state machine. Transitions from one state to the next are determined probabilistically. Like a finite-state machine, a symbol is emitted during each state transition. In an HMM the selection of the symbol emission during each state transition is also probabilistic. If the HMM is being used to model a word, the symbols could be phonemes in a speech system. As the symbols are not necessarily unique to a particular state, it is difficult to determine the state that the HMM is in simply by observing the emitted symbols – hence the term "hidden." These probabilities can be determined from data of the real-world process that the HMM is being used to model. One variation is the study of Morgan and Bourlard [18], who used a large BP network to provide HMM emission probabilities.

In many speech-recognition systems, HMMs are used to find simple contextual structure in candidate phoneme streams. Most HMM implementations generate parallel hypotheses and then use a dynamic programming algorithm (such as the Viterbi algorithm) to find a match that is the most likely utterance (or most likely path through the model) based on the phonemes captured by preprocessing and capturing features from the speech stream, and the probabilities used in constructing the HMM. However, HMMs have several limitations:

- They grow very large if the probability space is complex, such as when pairs of symbols are modeled rather than single symbols; yet most human language has very high order structure.
- The "Markov horizon", which is a fundamental definition of Markov models and makes them tractable analytically, also contributes to the inability to capture higher-order knowledge. Many now believe that we have passed the point of diminishing returns for HMMs in speech recognition.

The key point here is that most neuro-inspired silicon – and in particular the analog-based components – is primarily focused on the front-end "DSP" part of the problem, since robust, generic back-end algorithms (and subsequently hardware) have eluded identification. It has been argued by some that if the front end was performed correctly, then the back-end would be easier, but whilst it is always easier to do better in front-end processing the room for improvement is smaller there. Without robust back-end capabilities, general solutions will be more limited.

10.2.1 Flexibility Trade-Offs and Amdhal's Law

During the 1980s and early 1990s, when most of the hardware surveyed in this chapter was created, there was a perception that the algorithms were large and

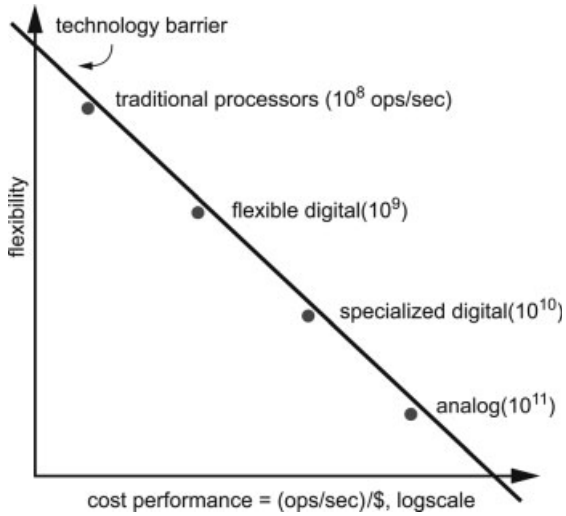


Figure 10.5 The flexibility–cost/performance trade-off.

complex and therefore slow to emulate on the computers of that time. Consequently, specialized hardware to accelerate these algorithms was required for successful applications. What was not fully appreciated by many was that the performance of general-purpose hardware was increasing faster than Moore's law, and that the existing neural algorithms did not scale well to the large sizes that would have fully benefited from special purpose hardware. The other problem was *Amdahl's law*.

As discussed above, these models have extensive concurrency which naturally leads to massively parallel implementations. The basic computation in these models is the multiply-accumulate operation that forms the core of almost all DSP and which can be performed with minimal, fixed point, precision. Also during the early 1990s, when many of the studies on neural inspired silicon were carried out, microprocessor technology was actually not fast enough for many applications.

The problem is that neural network silicon is highly specialized and there are specific risks involved in its development. One way to conceptualize the trade-offs involved in designing custom hardware is shown in Figure 10.5. Although *cost-performance*[3)] can be measured, flexibility cannot be assessed as easily, and so the graph in Figure 10.5 is more conceptual than quantitative. The general idea is that the more a designer hard-wires an algorithm into silicon, the better the cost-performance of the device, but the less flexible.

The line, which is moving slowly to the right according to Moore's law, shows these basic trade-offs and is, incidentally, not likely to be linear in the real world. Another

3) In this chapter, cost-performance is measured by (operations/second)/cost. The cost of producing a silicon chip is directly related (in a complex, non-linear manner) to the *area* of the chip. Larger chips are generally more expensive, as used here. More recently, however, other factors such as *power consumption* have become equally, if not more, important.

assumption is that the algorithms being emulated can be implemented at many different points on that scale, which is not always true either.

An important assumption in this analysis is that most applications require a wide range of computations, although there are exceptions, as in most real-world systems the "recognition" component is only a part of a larger system. So, when considering specialized silicon for such a system, the trade-off shown in Figure 10.5 must be factored into the analysis. More general-purpose chips tend to be useful on a larger part of the problem, but will lead to a sacrifice in cost-performance. On the other hand, the more specialized chips tend to be useful on a smaller part of the problem, but at a much higher cost-performance. Related to this trade-off then is Amdahl's law [19], which has always been a fundamental limitation to fully leveraging parallel computing,

> *Amdahl's law*: the speed-up due to parallelizing a computation is proportional to that portion of the computation that cannot be parallelized.

Imagine, for example, that there is a speech-recognition application, and 20% of the problem can be cast into a simplified parallel form. If a special-purpose chip was available that speeded up that 20% portion by 1000-fold, then the total system performance increase would be about 25%:

$$1/(0.8 + (0.2/1000)) = 1.25$$

Of course, depending on the cost of the 1000-fold chip, the 25% may still be worthwhile. However, if a comparably priced chip was available that was slower but more flexible and could parallelize 80% of the application, albeit with a more moderate speed increase (say 20-fold), then the total system performance would be over 400%:

$$1/(0.2 + (0.8/20)) = 4.17$$

Almost all computationally intensive pattern recognition problems have portions of sequential computation, even if it is just moving data into and out of the system, data reformatting, feature extraction, post-recognition tasks, or computing a final result. Amdahl's law shows that these sequential components have a significant impact on total system performance. As a result, the biggest problem encountered by many early neural network chips was that they tended to speed up a small portion of a large problem by moving to the right in Figure 10.5. For many commercial applications, after all was said and done, the cost of a specialized neural chip did not always justify the resulting modest increase in total system performance.

During the mid-1990s, desktop chips were doubling their performance every 18 to 24 months. Then, during the mid-1990s both Intel and AMD added on-chip SIMD coprocessing in the form of MMX which, for the Intel chips, has eventually evolved to SSE3 [20]. These developments, for the most part, spelled the death of most commercial neural net chips. However, in spite of limited commercial success most neural network chips were very interesting implementations, often with elegant engineering. A representative sample of some of these chips will be examined briefly in the remainder of this chapter.

It should not be concluded from the discussions so far that specialized chips are never economically viable. Rather, the continued success of graphics processors and DSPs are examples of specialized high-volume chips, and some neural networks chips[4] have found very successful niches. Nonetheless, it does illustrate some of the problems involved in architecting a successful niche chip. An example is the commercial DSP chips used for signal processing and related applications, these provide unique cost-performance, efficient power utilization, and just the right amount of specialization in their niche to hold their own in volume applications against general-purpose processors. In addition, they have enough volume and history to justify a significant software infrastructure.

In light of what is now known about Amdahl's law and ISP, the history and state of the art of neuro-inspired silicon can now be surveyed.

10.2.2
Analog Very-Large-Scale Integration (VLSI)

There is no question that the most elegant implementation technique developed for neural emulation is the sub-threshold CMOS technology pioneered by Carver Mead and his students at CalTech [21]. Most MOS field effect transistors (MOSFETs) used in digital logic are operated in two modes, either off (0) and on (1). For the *off* state the gate voltage is more or less zero and the channel is completely closed. For the *on* state, the gate voltage is significantly above the transistor threshold and the channel is saturated. A saturated *on* state works fine for digital, and is generally desired to maximize current drive. However, the limited gain in that regime restricts the effectiveness of the device in analog computation. This is due to the fact that the more gain the device has, the easier it is to leverage this gain to create circuits that perform useful computation and which also are insensitive to temperature and device variability.

However, if the gate voltage is positive (for the nMOS gate) but below the point where the channel saturates, the FET is still on, though with a much lower current. In this mode, which sometimes is referred to as "weak inversion", there is useful gain and the small currents significantly lower the power requirements, though FETs operating in this mode tend to be slower. Carver Mead's great insight was that when modeling biologically inspired circuits, significant computation could be carried out using simple FETs operating in the sub-threshold regime where, like real neurons, performance resulted from parallelism and not the speed of the switching devices. Moreover, as Carver and colleagues have shown, these circuits do a very good job approximating a number of neuroscience functions.

By using analog voltage and currents to represent signals, the considerable expense of converting signal data into digital, computing the various functions in digital, and then converting the signal data back to analog, was eliminated. Neurons operate slowly and are not particularly precise, yet when combined appropriately they perform complex and remarkably precise computations. The goal of the aVLSI

4) One example is General Vision; http://www.general-vision.com/.

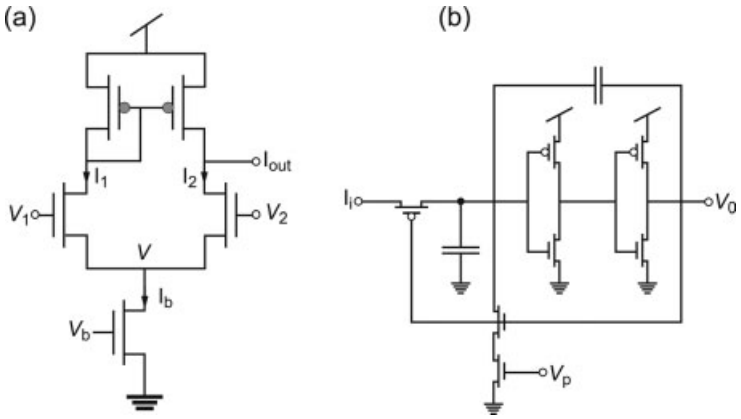


Figure 10.6 Basic aVLSI building blocks. (a) A transconductance amplifier; (b) an integrate and fire neuron.

research community has been to create elegant VLSI sub-threshold circuits that approximate biological computation.

One of the first chips developed by Carver *et al*. was the silicon retina [22]. This was an image sensor that performed localized adaptive gain control and temporal/spatial edge detection using simple "local neighborhood" functional extensions to the basic photosensitive cell. There subsequently followed a silicon cochlea and numerous other simulations of biological circuits.

Two examples of these circuits are shown in Figure 10.6. A transconductance amplifier (voltage input, current output) and an "integrate and fire" neuron are two of the most basic building blocks for this technology. The current state of aVLSI research is very well described by Douglas [23], of the Neuroinformatics Institute, ETH-Zurich:

> *Fifteen years of Neuromorphic Engineering: progress, problems, and prospects*. Neuromorphic engineers currently design and fabricate artificial neural systems: from adaptive single chip sensors, through reflexive sensorimotor systems, to behaving mobile robots. Typically, knowledge of biological architecture and principles of operation are used to construct a physical emulation of the target neuronal system in an electronic medium such as CMOS analog very large scale integrated (aVLSI) technology.
>
> Initial successes of neuromorphic engineering have included smart sensors for vision and audition; circuits for non-linear adaptive control; non-volatile analog memory; circuits that provide rapid solutions of constraint-satisfaction problems such as coherent motion and stereo-correspondence; and methods for asynchronous event-based communication between analog computational nodes distributed across multiple chips.

> *These working chips and systems have provided insights into the general principles by which large arrays of imprecise processing elements could cooperate to provide robust real-time computation of sophisticated problems. However, progress is retarded by the small size of the development community, a lack of appropriate high-level configuration languages, and a lack of practical concepts of neuronal computation.*

Although still a modest-sized community, research continues in this area, the largest group being that at ETH in Zurich. The commercialization of this technology has been limited, however, with the most notable success to date being that of Synaptics, Inc. This company created several products which used the basic aVLSI technology, the most successful being the first laptop touch pads.

10.2.3
Intel's Analog Neural Network Chip and Digital Neural Network Chip

During the "heyday" of neural network silicon, between 1986 and 1996, a major semiconductor vendor, Intel, produced two neural network chips. The first, the ETANN [24] (Intel part number 80170NX), was completely analog, but it was designed as a general-purpose chip for non-linear feed-forward ANN operation. There were two grids of analog "inner product" networks, each with 80 inputs and 64 outputs, and a total of 10 K (5 K for each grid) weights. The chip computed the two inner products simultaneously, taking about 5 μs for the entire operation. This resulted in a total performance (feed-forward only, no learning) of over two billion connections computed per second, where a connection is a single multiply-accumulate of an input-weight pair. All inputs and outputs were in analog. The weights were analog voltages stored on floating gates – with the chip being developed and manufactured by the flash memory group at Intel. Complementary signals for each input provided positive and negative inputs. An analog multiplier was used to multiply each input by a weight, current summation of multiplier outputs provided the accumulation, with the output being sent through a non-linear amplifier (giving roughly a sigmoid function) to the output pins.

Although not designed specifically to do real-time learning, it was possible to carry out "chip in the loop" learning where incremental modification of the weights was performed in an approximately stochastic fashion. Learning could also be done off-line and the weights then downloaded to the chip.

The ETANN chip had very impressive computational density, although the awkward learning and total analog design made it somewhat difficult to use. The multipliers were non-linear, which made the computation sensitive to temperature and voltage fluctuations. Ultimately, Intel retired the chip and moved to a significantly more powerful and robust all digital chip, the Ni1000.

The Ni1000 [25, 26] implemented a family of algorithms based on radial basis function networks (RBF [26]). This family included a variation of a proprietary algorithm created by Nestor, Inc., a neural network algorithm and software company.

Rather than doing incremental gradient descent learning, as can be seen with the BP algorithm, the Ni1000 used more of a template approach where each node represented a "basis" vector in the input space. The width of these regions, which was controlled by varying the node threshold, was reduced incrementally when errors were made, allowing the chip to start with crude over-generalizations of an input to output space mapping, and then fine-tune the mapping to capture more complex variations as more data are input. An input vector would then be compared to all the basis vectors, with the closest basis vector being the winner. The chip performed a number of concurrent basis computations simultaneously, and then, also concurrently, determined the classification of the winning output, both functions were performed by specialized hardware.

The Ni1000 was a two-layer architecture. All arithmetic was digital and the network parameters/weights were stored in Flash EEPROM. The first or hidden layer had 256 inputs of 16 bits each with 16 bit weights. The hidden layer had 1024 nodes and the second or output layer 64 nodes (classes). The hidden layer precision was 5 to 8 bits for input and weight precision. The output layer used a special 16-bit floating point format. One usage model was that of Bayesian classification, where the hidden layer learns an estimate of a probability density function (PDF) and the output layer classifies certain regions of that PDF into up to 64 different classes. At 40 MHz the chip was capable of over 10 billion connection computations per second, evaluating the entire network 40 K times per second with roughly 4 W peak power dissipation.

The Ni1000 used a powerful, compact, specialized architecture (unfortunately, space limitations prevent a more detailed description here, but the interested reader is referred to Refs. [25, 26]). The Ni1000 was much easier to use than the ETANN and provided very fast dedicated functionality. However, referring back to Figure 10.5, this chip was a specific family of algorithms wired into silicon. Having a narrower functionality it was much more at risk from Amdahl's law, as it was speeding up an even smaller part of the total problem. Like CNAPS (the Connected Network of Adapted Processors), it too was ultimately "run over by the silicon steam roller".

10.2.4
Cellular Neural Networks

Cellular neural networks (CNN) constitute another family of analog VLSI neural networks. This was proposed by Leon Chua in 1988 [27], who called it the "Cellular Neural Network", although now it is known as the "Cellular Non-Linear Network". Like aVLSI, CNN has a dedicated following, the most well-known being the Analogic and Neural Computing Laboratory of the Computer and Automation Research Institute of the Hungarian Academy of Sciences under the leadership of Tamas Roska. CNN-based chips have been used to implement vision systems, and complex image processing similar to that of the retina has been investigated by a number of groups [28].

Although there are variations, the basic architecture is a 2-D rectangular grid of processing nodes. Although the model allows arbitrary inter-node connectivity, most CNN implementations have only nearest-neighbor connections. Each cell computes its state based on the values of its four immediate neighbors, where the neighbor's

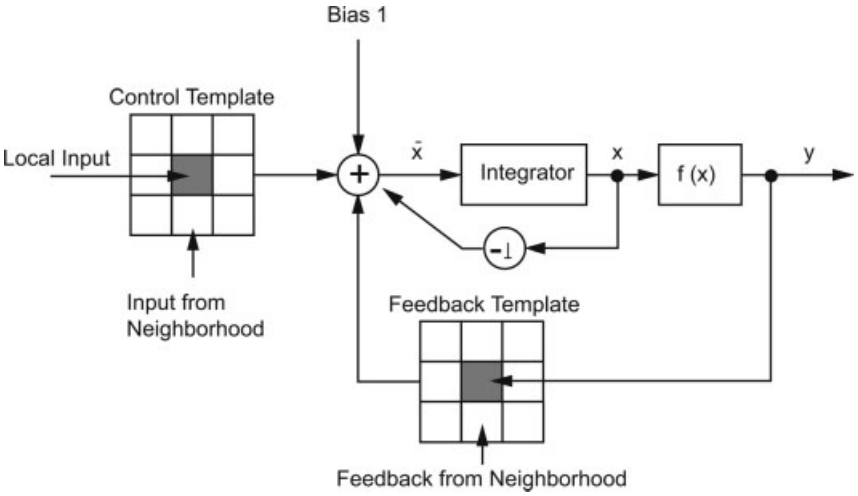


Figure 10.7 Basic cellular neural network (CNN) operation [72].

voltage and the derivative of this voltage are each multiplied by constants and summed. Each node then takes its new value and the process continues for another clock. This computation is generally specified as a type of filter, and is done entirely in the analog domain. However, as the algorithm steps are programmable one of the real strengths of CNN is that the inter-node functions and data transfer is programmable, with the entire array appearing as a digitally programmed array of analog-based processing elements. This is an example of a Single Instruction, Multiple Data (SIMD) architecture, which consists of an array of computation units, where each unit performs the same operation, but each on its own data. CNN programming can be complex and requires an intimate understanding of the basic analog circuits involved. The limited inter-node connectivity also restricts the chip to mostly "front-end" types of processing, primarily of images. A schematic of the basic CNN cell is shown in Figure 10.7.

Whereas, research and development continue, the technology has had only limited commercial success. As with aVLSI, it is a fascinating and technically challenging system, but in real applications it tends to be used for front-end problems and consequently is subject to Amdahl's law.

10.2.5
Other Analog/Mixed Signal Work

It is difficult to do justice to the large and rich area of biologically inspired analog design that has developed over the years. Other investigations include those of Murray [29], the former neural networks group at AT&T Bell Labs [30], Ettienne-Cummings [31], Principe [32], and many more that cannot be mentioned due to limited space. And today, some workers, such as Boahan, are beginning to move the processing further into the back end [33] by looking at cortical structures for early vision.

On returning to Figure 10.4, it can be seen that the first few boxes of processing require the type of massively parallel, locally connected feature extraction that CNN, aVLSI and other analog techniques provide. With regards to sensors, these can perform enhanced signal processing, and demonstrate better signal-to-noise ratios than more traditional implementations, providing such capabilities in compact, low-power implementations.

Although further studies are needed, there is a concern that the limited connectivity and computational flexibility make it difficult to apply these technologies to the back end. Although not a universally held opinion, the author feels that these higher-level association areas require a different approach to implementation. This general idea will be presented in more detail below, but first, it is important to examine another major family of neural network chips, the massively parallel digital processors.

10.2.6 Digital SIMD Parallel Processing

Concurrent with the development of analog neural chips, parallel effort was devoted to architecting and building digital neural chips. Although these could have dealt with a larger subset of pattern-recognition solutions they, like the analog chips, were mostly focused on neural network solutions to simple classification. A common design style that was well matched to the basic ANN algorithms was that of SIMD processor arrays. One chip that embodied that architecture was CNAPS, developed by Adaptive Solutions [34, 35].

The world of digital silicon has always flirted with specialized processors. During the early days of microprocessors, silicon limitations restricted the available functionality and as a result many specialized computations were provided by coprocessor chips. Early examples of this were specialized floating point chips, as well as graphics and signal processing. Following Moore's law, the chip vendors found that they could add increasing amounts of function and so began to pull some of these capabilities into the processor.

Interestingly, graphics and signal processing have managed to maintain some independence, and remain as external coprocessors in many systems. Some of the reasons for this were the significant complexity in the tasks performed, the software inertia that had built up around these functions, and the potential for very low power dissipation which is required for embedded signal processing applications, such as cell phones, PDAs, and MP3 players.

During the early 1990s it was clear that there was an opportunity to provide a significant speed-up of basic neural network algorithms because of their natural parallelism. This was particularly true in situations involving complex, incremental, gradient descent adaptation, as can be seen in many learning models. As a result, a number of digital chips were produced that aimed squarely at supporting both learning and non-learning network emulation.

It was clear from Moore's law that performance improvements and enhanced functionality continued for mainline microprocessors. This relentless march of the

desktop processors was referred to as the "silicon steam roller" where, as the 1990s continued, it became increasingly difficult for the developers of specialized silicon to stay ahead of. At Adaptive Solutions, the goal was to avoid the steam roller by steering between having enough flexibility to solve most of the problem, to avoid Amdahl's law, and yet to have a sufficiently specialized function that allowed enough performance to make the chip cost-effective – essentially sitting somewhere in the middle of the line in Figure 10.5. This balancing act became increasingly difficult until eventually the chip did not offer enough cost-performance improvement in its target applications to justify the expense of a specialized coprocessor chip and board.

The CNAPS architecture consisted of a one-dimensional (1-D) processor node (PN) array in an SIMD parallel architecture [36]. To allow as much performance-price as possible, modest fixed point precision was used to keep the PNs simple. With small PNs the chip could leverage a specialized redundancy technique developed by Adaptive Solution's silicon partner, Inova Corporation. During chip testing, each PN could be added to the 1-D processor chain, or bypassed. In addition, each PN had a large power transistor (with a width of 20 000 λ) connecting the PN to ground. Laser fuses on the 1-D interconnect and the power transistor were used to disconnect and power down defective PNs. The testing of the individual PNs was done at wafer sort, after which an additional lasing stage (before packaging and assembly) would configure the dice, fusing in the good PNs and fusing out and powering down the bad PNs. The first CNAPS chip had an array of 8×10 (80) PNs fabricated, of which only 64 needed to work to form a fully functional die. The system architecture and internal PN architecture is shown in Figure 10.8.

Simulation and analysis was used to determine approximately the optimal PN size (the "unit of redundancy") and the optimal number of PNs. Ultimately, the die was almost exactly 2.5 cm (1 inch) on a side with 14 million transistors fabricated; this led to 12 dice per 15-cm (6-inch) wafer which, until recently, made it physically the largest processor chip ever made (see Figure 10.9).

The large version of the chip, called the CNAPS-1064, had 64 operational PNs and operated at 25 MHz with 6 W worst-case power consumption. Each PN was a complete 16-bit DSP with its own memory. Neural network algorithms tend to be vector/matrix-based and map fairly cleanly to a 1-D grid, so it was easy to have all PNs performing useful work simultaneously. The maximum compute rate then was 1.2 billion multiply-accumulates per second per chip, which was about 1000-fold faster than the fastest workstation at that time. Part of this speed up was due to the fact that each PN did several things in one clock to realize a single clock multiply-accumulate: input data to the PN, perform a multiply, perform an accumulate, perform a memory fetch, and compute the next memory address. During the late 1980s, DSP chips were able to perform such a single multiply-accumulate in one clock, but it was not until the Pentium Pro that desktop microprocessors reached a point where they performed most of these operations simultaneously.

When developing the CNAPS architecture, a number of key decisions were made, including the use of limited precision and local memories, architecture support for the C programming language, and I/O bandwidth.

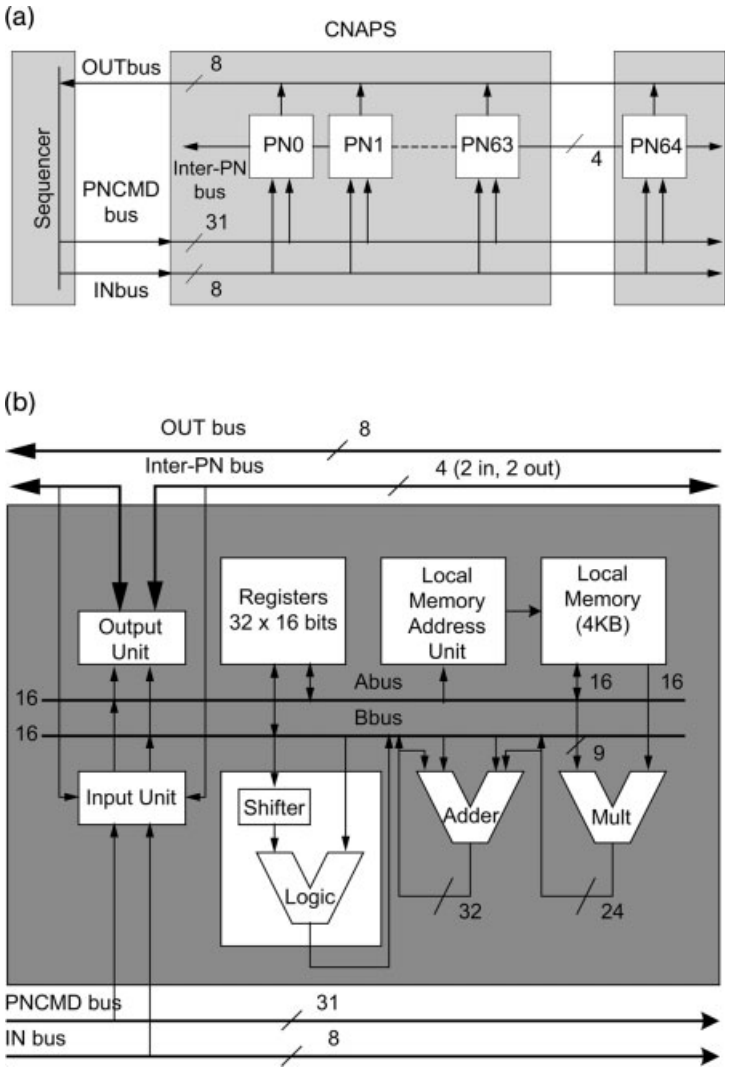


Figure 10.8 Connected Network of Adaptive Processors (CNAPS) architecture. (a) System architecture; (b) PN architecture.

At a time when the computing community was moving to floating-point computation, and the microprocessor vendors were pulling floating processing onto the processor chips and optimizing its performance, the CNAPS used limited precision, fixed point arithmetic. The primary reason for this decision was based on yield calculations, which indicated that a floating-point PN was too large to take advantage of PN redundancy. This redundancy bought an approximately twofold cost-performance improvement. Since a floating-point PN would have been two to three times larger than a fixed-point PN, the use of modest precision fixed-point arithmetic meant an almost sixfold difference in cost-performance. Likewise, simulation showed that

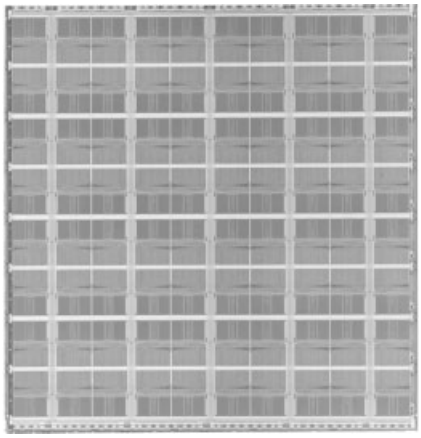


Figure 10.9 A photograph of the CNAPS die.

most of the intended algorithms could get by with limited precision fixed-point arithmetic, and this proved to be, in general, a good decision, as problems were rarely encountered with the limited precision. In fact, the major disadvantage was that it made programming more difficult, although DSP programmers had been effectively using fixed-point precision for many years.

The second decision was to use local, per PN, memory (4 KB of SRAM per PN). Although this significantly constrained the set of applications that could leverage the chip, it was absolutely necessary in achieving the performance goals. The reality was that it was unlikely that the CNAPS chip would have ever been built had performance been reduced enough to allow the use of off-chip memory. As with DSP applications, almost all memory access was in the form of long arrays that can benefit from some pre-fetching, but not much from caching.

The last two decisions – architecture and I/O bandwidth limitations – were driven by performance-price and design time limitations. One objective of the architecture was that it be possible for two integrated circuit (IC) engineers to create the circuits, logic schematics and layout (with some additional layout help) in one year. As a result, the architecture was very simple, which in turn made the design simpler and the PNs smaller, but programming was more difficult. One result of this strategy was that the architecture did not support the C language efficiently. Although there were some fabrication delays, the architecture, board, and software rolled out simultaneously and worked very well, and the first system was shipped in December 1991 and quickly ramped to a modest volume. One of the biggest selling products was an accelerator card for Adobe Photoshop which, in spite of Amdahl problems (poor I/O bandwidth), offered unprecedented performance.

By 1996, desktop processors had increased their performance significantly, and Intel was on the verge of providing the MMX SIMD coprocessor instructions. Although this first version of an SIMD coprocessor was not complete and was not particularly easy to use, the performance of a desktop processor with MMX reduced

the advantages of the CNAPS chipset even further in the eyes of their customers, and people stopped buying.

Everybody knew that the silicon steam roller was coming, but it was moving much faster (and perhaps even accelerating, as some had suggested) than expected. In addition, Intel quickly enhanced the MMx coprocessor to the now current SSE3, which is a complete and highly functional capability. DSPs were also vulnerable to the microprocessor steam roller, but managed, primarily through software inertia and very low power dissipation, to hold their own.

Although there were other digital neural network processors, none of them achieved any significant level of success, and basically for the same reasons. Although at this point the discussion of all but a few of these others is limited by space, two in particular deserve mention.

10.2.7 Other Digital Architectures

One important digital neural network architecture was the Siemens SYNAPSE-1 processor developed by Ramacher and colleagues at Siemens Research in Munich [37]. The chip was similar to CNAPS in terms of precision, fixed-point arithmetic, and basic processor node architecture, but differed by using off-chip memory to store the weight matrices.

A SYNAPSE-1 chip contained a 2×4 array of MA16 "neural signal processors," each with a 16-bit multiplier and 48-bit accumulator. The chip frequency was 40 MHz, and one chip could compute about five billion connections per second with feedforward (non-learning) execution.

Recall that, in architecting the CNAPS, one of the most important decisions was whether to use on-chip per PN memory, or off-chip shared memory for storing the primary weight matrices. For a number of reasons, including the targeted problems space and the availability of a state-of-the-art SRAM process, Adaptive Solutions chose to use on-chip memory for CNAPS. However, for performance reasons this decision limited the algorithm and application space to those whose parameters fit into the on-chip memories. Although optimized for matrix-vector operations, CNAPS was designed to perform efficiently over a fairly wide range of computations.

The SYNAPSE-1 processor was much more of a matrix–matrix multiplication algorithm mapped into silicon. In particular, Ramacher and colleagues were able to take advantage of a very clever insight – the fact that in any matrix multiplication, the individual elements of the matrix are used multiple times. The SYNAPSE-1 broke all matrices into 4×4 chunks. Then, while the elements of one matrix were broadcast to the array, 4×4 chunks of the other matrix would be read from external memory into the array. In a 4×4-matrix by 4×4-matrix multiplication, each element in the matrix was actually used four times, which allowed the processor chip to support four times as many processor units for a given memory bandwidth than a processor not using this optimization.

On returning to Figure 10.5, it can be seen that the SYNAPSE-1 architecture increased performance by specializing the architecture to matrix–matrix multiplications.

Fortunately, most neural network computations can be cast in a matrix form, though it did restrict maximum machine performance to algorithms that performed matrix–matrix multiples. However, like the other digital neural network chips, the SYNAPSE-1 eventually lost out to high-performance microprocessor and DSP hardware.

10.2.8 General Vision

A similar chip to the Ni1000 was the ZISC (Zero Instruction Set Computer) developed by Paillet and colleagues at IBM in Paris. The ZISC chip was digital, employed basically a "vector template" approach, and was simpler and cheaper than the Ni1000 but implemented approximately the same algorithms. Today, the ZISC chip survives as the primary product of General Vision, Petaluma, California.

In addition to the CNAPS, SYNAPSE-1, ZISC, and Ni1000, several other digital chips have been developed either specifically or in part to emulate neural networks. HNC developed the SNAP, a floating-point SIMD standard cell-based architecture [38]. One excellent architecture is the SPERT [39], which was developed by groups at the University of California Berkeley and the International Computer Science Institute (ICSI) in Berkeley. SPERT was designed to perform efficient integer vector arithmetic and to be configured into large parallel arrays. A similar parallel processor array that was created from field-programmable gate arrays (FPGAs) and suited to neural network emulation was REMAP [40].

10.3 Current Directions in Neuro-Inspired Hardware

One limitation of traditional ANN algorithms is that they did not scale particularly well to very large configurations. As a result, commercial silicon was generally fast enough to emulate these models, thus reducing the need for specialized hardware. Consequently, with the exception of on-going studies in aVLSI and CNN, general research in neural inspired hardware has languished.

Today, however, activity in this area is picking up again, for two main reasons. The first reason is that computational neuroscience is beginning to yield algorithms that can scale to large configurations and have the potential for solving large, very complex problems. The second reason is the excitement of using molecular-scale electronics, which makes possible comparably scalable hardware. As will be seen, at least one of the projected nanoelectronic technologies is a complementary match to biologically inspired algorithms.

Today, a number of challenges face the semiconductor industry, including power density, interconnect reverse scaling, device defects and variability, memory bandwidth limitations, performance overkill, density overkill, and increasing design complexity. *Performance overkill* is where the highest-volume segments of the market are no longer performance/clock frequency-driven. *Density overkill* is where it is

difficult for a design team to effectively design and verify all the transistors available to them on a single die. Although neither of these is a potential show-stopper, taken together they do create some significant challenges.

Another challenge is the growing reliance on parallelism for performance improvements. In general purpose applications, the primary source of parallelism has been within a single instruction stream, where many instructions can be executed simultaneously, sometimes even out of order. However, this instruction level parallelism (ILP) has its limits and becomes exponentially expensive to capture. Microprocessor manufacturers are now developing "multiple core" architectures, the goal of which is to execute multiple threads efficiently. As multiple core machines become more commonplace, software and application vendors will struggle to create parallel variations of their software.

Due to very small, high-resistance wires, many nano-scale circuits will be slow, and power density will be a problem because of high electric fields. Consequently, performance improvements at the nano-scale will also need to come almost exclusively from parallelism and to an even greater extent than traditional architectures.

When considering these various challenges, it is unclear which ones are addressed by nanoelectronics. In fact, nanoelectronics only addresses the end of Moore's law, and perhaps also the memory bandwidth problem. However, it also aggravates most other existing problems, notably signal/clock delay, device variability, manufacturing defects, and design complexity.

In proceeding down the path of creating nanoscale electronics, by far the biggest question is, how exactly will this technology be used? Can it be assumed that computation, algorithms, and applications will continue more or less as they have in the past? What should the research agenda be? Will the nanoscale processor of the future consist of thousands of $\times 86$ cores with a handful of application-specific coprocessors? The effective use of nanoelectronics will require solutions to more than just an increased density; rather, total system solutions will need to be considered. Today, computing structures cannot be created in the absence of some sense of how they will be used and what applications they will enable. Any paradigm shift in applications and architecture will have a profound impact on the entire design process and the tools required, as well as the requirements placed on the circuits and devices themselves.

As discussed above, algorithms inspired by neuroscience have a number of interesting and potentially useful properties, including fine-grain and massive parallelism. These are constructed from slow, low-power, unreliable components, are tolerant of manufacturing defects, and are robust in the presence of faulty and failing hardware. They adapt rather than be programmed, they are asynchronous, compute with low precision, and degrade gracefully in the presence of faults. Most importantly, they are adaptive, *self-organizing* structures which promise some degree of design error tolerance, and solve problems dealing with the interaction of an organism/system with the real world. The functional characteristics of neurons, such such as analog operation, fault tolerance, slow, massive parallelism, are radically different from those of typical digital electronics. Yet, some of these characteristics match very well the basic characteristics such as large numbers of faults and defects,

low speed, and massive parallelism that many research groups feel will characterize nanoelectronics systems.

Self-organization involves a system adapting (usually increasing in complexity) in response to an external stimulus. In this context, a system will learn about its environment and adjust itself accordingly, without any additional intervention. In order to achieve some level of self-organization, a few fundamental operating principles are required. Self-organizing systems are those that have been built with these principles in mind.

Recently, Professor Christoph von der Malsburg has defined a new form of computing science – "organic computing" – which deals with a variety of computations that are performed by biology. Organic computations are massively parallel, low precision, distributed, adaptive, and self-organizing. The neural algorithms discussed in this chapter form an important subset of this area (the interested reader is referred to the web site: www.organic-computing.org).

Several very important points should be made about biologically inspired models. The first point concerns the computational models and the applications they support. Biologically inspired computing uses a very different set of computational models than have traditionally been used. And subsequently they are aimed at a fairly specialized set of applications. Consequently, for the most part biological models are not a replacement for existing computation, but rather they are an enhancement to what is available now. Specialized hardware for implementing these models needs to be evaluated accordingly, and in the next few sections some of these models will be explored at different levels.

10.3.1
Moving to a More Sophisticated Neuro-Inspired Hardware

As mentioned above, it is the back end where the struggle with algorithms and implementation continues, and it is also the back end where potential strategic inflection points lie. Hence, the remainder of the chapter will focus on back-end algorithms and hardware.

The ultimate cognitive processor is the *cerebral cortex*. The cortex is remarkably uniform, not only across its different parts, but also across almost all mammalian species. Although current knowledge is far from providing an understanding of how the cerebral cortex does what it does, some of the basic computations are beginning to take shape. Nature has, it appears, produced a general-purpose computational device that is a fundamental component of higher level intelligence in mammals.

Some generally accepted notions about the cerebral cortex are that it represents knowledge in a sparse, distributed, hierarchical manner, and that it performs a type of Bayesian inference over this knowledge base, which it does with remarkable efficiency. This knowledge is added to the cortical data base by a complex process of adaptive learning.

One of the fundamental requirements of intelligent computing, the need to capture higher-order relationships. The problem with Bayesian inference is that it

is an exponentially increasing computation in the number of variables (it has been shown to be NP-Hard, which means that the number of computational steps increases exponentially with the size of the problem); in other words, as order increases the computational overhead increases even more rapidly. Consequently, to use Bayesian inference in real problems, order is reduced to make them computationally tractable.

One common way to do this is to create a Bayesian network, which is a graph structure where the nodes represent variables and the arcs connecting the nodes represent dependencies. If there is reasonable independence between many of the variables then the network itself will be sparsely connected. Bayesian networks "factorize" the inference computation by taking advantage of the independence between different variables. Factorization does reduce the computational load, but at the cost of limiting the knowledge represented in the network. A custom network is also required for each problem.

Cortical networks appear to use sparse distributed data representations, where each neuron participates in a number of specific data representations. Distributed representations also diffuse information, topologically localizing it to the areas where it is needed and reducing global connectivity. Computing with distributed representations can be thought of as the hardware equivalent of spread spectrum communication, where pseudo-random sequences of bits are used to spread a signal across time and frequency. In addition to spreading inter-node communication, distributed representations also spread the computation itself. One hypothesis of cortex operation is that distributed representations of information are a form of extreme factorization, allowing efficient, massively parallel Bayesian inference.

Mountcastle [41, 42] conducted many pioneering studies in understanding the structural architecture of the cortex, including proposing the columnar organization. The fundamental unit of computation appears to be the *cortical minicolumn*, a vertically organized group of about 80 to 100 neurons which traverses the thickness of the gray matter ($\sim$3 mm) and is about 50 μm in diameter. Neurons in a column tend to communicate vertically with other neurons on different layers in the same column. These are subsequently organized into larger columns variously called just "columns", "cortical columns", "hypercolumns", or "modules." Braitenberg [43] postulates two general connectivity systems in cortex: "metric" (high-density connections to physically local cells, based on actual 2-D layout); and "ametric" (low-density point-to-point connections to all large groups of densely connected cells). Connectivity is significantly denser in the metric system, but to a limited extent.

One approach to creating cortical-like algorithms is to model each column as an auto-associative network, such as the Palm model discussed above. The columns are then sparsely connected to each other, but the specifics of the inter-column connections are still not certain and different research groups have expressed different ideas about this [44]. In addition, the neocortex has a definite hierarchical organization where there are as many feedback paths as feed-forward paths.

Among other things, the massive scale is probably one of the more important advantages of biological computation. Consequently, it is likely that useful versions of

these algorithms will require networks with a million or more nodes. Back-end processing, because of a need to store large amounts of unique synaptic information, will most likely have simpler processing than is seen at the front end, albeit on a much larger scale.

Hecht-Nielsen [45] bases the inter-column (which he calls "regions") connections on conditional probabilities, which capture higher-order relationships. He also uses abstraction columns to represent groups of lower-level columns. He has demonstrated networks that perform a remarkable job of capturing aspects of English, as these networks consist of several billion connections and require a large computer cluster to execute.

Granger [46, 47] leverages nested distributed representations in a way that adds the temporal dimension, creating hierarchical networks that learn sequences of sequences. George and Hawkins [48] use model likelihood information ascending a hierarchy with model confidence information being fed back. Other researchers are also contributing to these ideas include Grossberg [49], Lansner [50, 51], Arbib [52], Roudi and Treves [53], Renart *et al.* [54], Levy *et al.* [55], and Anderson [56]. Clearly, this remains a dynamic area of research, and at this point there is no clear "winning" approach.

Another key feature of some of these algorithms is that there is an oscillatory sliding threshold that causes the more "confident" columns to converge to their attractors more quickly, the less-confident more slowly, while those of low confidence do not converge at all, taking a "NULL" state. This process is remarkably similar to the electromagnetic waves that flow through the cortex when it is processing data.

Connectivity remains one of the most important problems when first considering scaling to very large models. Axons are very fine and can be packed very densely in a three-dimensional (3-D) mesh. Interconnect in silicon generally operates in a 2-D plane, although with several levels, nine or more with today's semiconductor technologies. Most importantly, silicon communication is not continuously amplifying, as can be seen in axonal and some dendritic processes. The following result [57–59] demonstrates this particular problem.

> *Theorem*: Assume an unbounded or large rectangular array of silicon neurons where each neuron receives input from its N nearest neighbors; that is, the fan-out (divergence) and fan-in (convergence) is N. Each such connection consists of a single metal line, and the number of 2-D metal layers is much less than N. Then, the area required by the metal interconnect is $O(N^3)$.

So, if the fan-in per node is doubled from 100 to 200, the silicon area required for the metal interconnect increases by a factor of 8. This result means that, even for modest local connectivity, that portion of silicon area devoted to the metal interconnect will dominate. It has been shown that for some models even moderate multiplexing of interconnect would significantly decrease the silicon area requirements, without any real loss in performance [60]. Carver Mead's group at Caltech, and others, developed the address event representation (AER), a technique for multiplexing a number of pulse streams on a single metal wire [61, 62]. When analog computation is

used, signals can be represented by the time between action-potential-like "spikes". These signal "packets" or "pulses" are transmitted asynchronously the moment they occur, with the originating unit's address, on a single shared bus. This "pseudo-digital" representation allows multiplexing of the bus and the retention of analog and temporal information, without expensive conversions.

In studying the potential implementations of cortical structures, an efficient connection multiplexing architecture was developed where data transfer occurs via overlapping, hierarchical buses [63]. This structure, *The Broadcast Hierarchy* (TBH), allows simultaneous high-bandwidth local connectivity and long-range connectivity, thereby providing a reasonable match to many biological connectivity patterns.

The details of a relevant proposed hybrid CMOS/nanoelectronic technology, CMOL, are presented in the next section.

10.3.2 CMOL

Likharev has proposed CMOL (Cmos/MOLecular hybrid) as an implementation strategy for charge-based[5)] nanoelectronic devices. Likharev's group has analyzed a number of examples of CMOL configurations, including memory, reconfigurable logic, and neuromorphic CrossNets [64–66]. In addition, nanogrids are most likely to be the first commercial deployment of nanoelectronic circuits [67].

CMOL consists of a set of nanogrids fabricated on top of traditional CMOS, with the CMOS being used, among other things, for signal restoration and current drive, nanogrid addressing, and to communicate signals into and out of the nanogrids. The nanogrids themselves will generally have more specialized computation such as a memory, which augments the computation being performed by the CMOS.

A *nanogrid* consists of a set of parallel wires, with another set of parallel wires fabricated on top of and orthogonal to the first set. Likharev has shown that such grids need not be laid out in perfect dimensions or alignment, but can be reproduced by using nanoimprint templates. Sandwiched between the two grids is a planar material made from a specific molecular structure such that, where the horizontal and vertical wires cross, a molecular switch is created. Several mechanisms have been identified to effect the desirable electrical properties where two nanowires cross.

The most important property is of a binary "latching switch" with two metastable internal states [68]. This nanoscale device can be programmed to either an "on" or an "off" state by using two sets of voltages. The lower set is used to read out the device to determine its state, while the higher set is used to change the state of the device. The programming voltages are used to switch the device between high- and low-resistance states. The lower read-out voltages are used to determine the resistance or

5) Researchers are investigating a number of molecular technologies based on computational paradigms other than charge, such as spintronics, quantum cellular automata, and DNA computing. However, as neural circuits operate on the principle of charge accumulation, charge-based computation seems a better match, although further study of these other technologies is required.

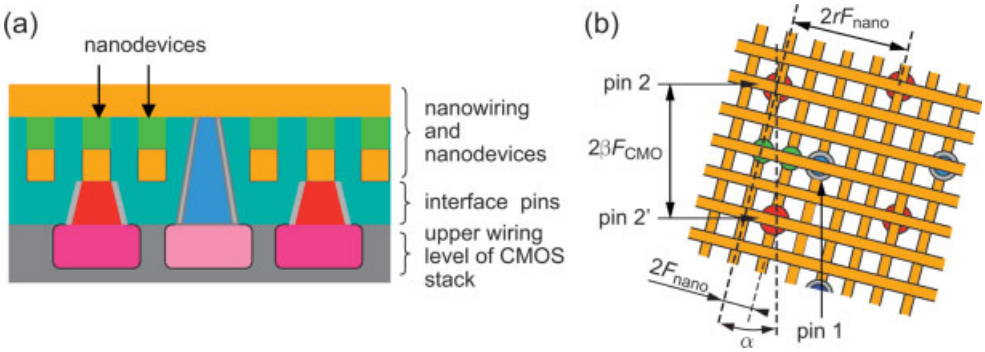


Figure 10.10 CMOL [1]. (a) A schematic side view. (b) Top view showing that any nanodevice may be addressed via the appropriate pin pair (e.g. pins 1 and 2 for the leftmost of the two shown devices, and pins 1 and 2′ for the rightmost device). Panel (b) shows only two devices; in reality, similar nanodevices are formed at all nanowire crosspoints. Also not seen on panel (b) are CMOS cells and wiring.

"state" of the molecule. Another required characteristic of these devices is rectification, where current flow is allowed only in one direction.

One of the most important characteristics of CMOL is the unique way in which the grids are laid at an angle with respect to the CMOS grid. Each nanowire is connected to the upper metal layers of the CMOS circuitry through a pin. In order for the CMOS to address each nanowire in the denser nanowire crossbar arrays, when it is fabricated, the nanowire crossbar is turned by an angle α, where α is the tangent of the ratio of the CMOS inter-pin distance to the nanogrid inter-pin distance. This technique allows the grid to be aligned arbitrarily with the CMOS and still have most nanowires addressable by some selection of CMOS cells. A nanowire is contacted by two CMOS cells, both of which are required to input a signal or read a signal. This basic connectivity structure is shown in Figure 10.10.

Although CMOL is not necessarily biologically inspired, it represents a promising technology for implementing such algorithms, as will be seen in the next section. CMOL uses charge-accumulation as its basic computational paradigm, which is also used by neural structures. Other nanoscale devices such as spin technologies do not implement a charge accumulation model, so such structures would have to emulate a charge-accumulation model, probably in digital.

10.3.3
An Example: CMOL Nano-Cortex

A high-level analysis has been performed of the implementation of a cortical column in CMOL. It is assumed that column operation is based on the Palm model discussed above. For this analysis, multiple column communication will be ignored and, for the sake of simplicity, a non-learning model will be assumed where the weights are computed off-line and downloaded into the nanogrid. Some typical values for the

Table 10.1 Typical values of parameters used for a cortical column analysis.

Parameter	Range	Typical Value I	Typical Value II
Hypercolumn node size	$128 \sim 128$ K	1 K	16 K
Weight matrix size (single-weight-bit)	$2^{14} \sim 2^{34}$ bits	2^{20} bits	2^{28} bits
Weight matrix size (multi-weight-bit)	$2^{18} \sim 2^{51}$ bits	2^{30} bits	2^{42} bits
Multi-weight-bits	$7 \sim 17$ bits	10 bits	14 bits
No. of active nodes in hypercolumn	$7 \sim 17$	16	16
Inner-product result bits (single-weight-bit)	$3 \sim 5$ bits	5 bits	5 bits
Inner-product result bits (multi-weight-bit)	$11 \sim 21$ bits	14 bits	18 bits

parameters used in the cortical column architectural analysis are listed in Table 10.1. These values represent typical numbers used by several different simulation models, in particular, Lansner and his group at KTH [69]. Related investigations have been conducted at IBM [70], where a mouse cortex-sized model has been simulated on a 32 K processor IBM BlueGene/L.

Four basic designs have been analyzed, as shown in Figure 10.11:

- All-digital CMOS
- Mixed-signal CMOS
- All-digital hybrid CMOS/CMOL
- Mixed-signal hybrid CMOS/CMOL.

For the CMOS designs and the CMOS portion of CMOL, a 22-nm process was assumed as a "maximally" scaled CMOS. To approximate the features for this process, a simple linear scaling of a 250-nm process was made. The results of this

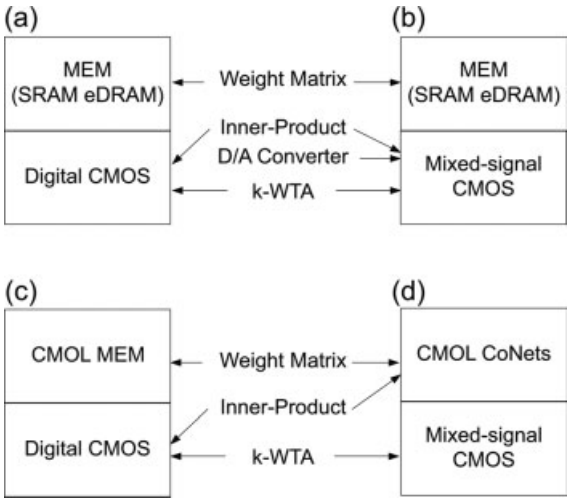


Figure 10.11 Architecture space - biologically inspired models. (a) All-digital CMOS; (b) mixed-signal CMOS; (c) all-digital hybrid CMOS/CMOL; (d) mixed-signal hybrid CMOS/CMOL.

Table 10.2 Analysis results.

Design		No. of column processors	Power (W)	Update rate (G nodes s^{-1})	Memory (%)
CMOS All Digital	1-bit eDRAM	6600	528	3072	2.9
CMOS Mixed-Signal	1-bit eDRAM	19 500	487	22 187	9.0
CMOL All Digital	1-bit CMOL Mm	4 042 752	317	4492	40
CMOL MS	1-bit CoNets	10 093 568	165	11 216	100

analysis, where the cost-performance for the four systems with the assumption of an 858 mm^2 die size (the maximum lithographic field size expected for a 22-nm process), are presented in Table 10.2.

With regards to Table 10.2, with the mixed signal CMOL it was possible to implement approximately 10 M columns, each having 1 K nodes, with 1 K connections each, for a total of 10 Tera-connections. In addition, this entire network can be updated once every millisecond – which is approaching biological densities and speeds, although of course with less functionality. Such technology could be built into a portable platform, with the biggest constraint being the high power requirements. Current studies include investigations into spike-based models [71] that should allow a significant lowering of the duty cycle and the power consumed.

Although real-time learning/adaptation was not included in the circuits analyzed here, deployed systems will need to be capable of real-time adaptation. It is expected that additional learning circuitry will reduce density by about two- to threefold. Neither has the issue of fault tolerance been addressed, although the Palm model has been found to tolerate errors, single 1 bits set to 0, in the weight matrix of up to 10%. For this reason, and given the excellent results of Likharev and Strukov [66] on the fault tolerance of CMOL arrays used as memory, it is expected that some additional hardware will be required to complement algorithmic fault tolerance, although this should not reduce the density in any significant way.

10.4 Summary and Conclusions

In this chapter, a brief and somewhat superficial survey has been provided of the specialized hardware developed over the past 20 years to support neurobiological models of computation. A brief examination was made of the current efforts and speculation made on how such hardware, especially when implemented in nanoscale electronics, could offer unprecedented compute density, possibly leading to new capabilities in computational intelligence. Biologically inspired models seem to be a better match to nanoscale circuits.

The mix of continued Moore's law scaling, models from computational neuroscience and molecular-scale technology portends a potential paradigm shift in how computing is carried out. Among other points, the future of computing is most likely not about discrete logic but rather about encoding, learning, and performing

inference over stochastic variables. There may be a wide range of applications for such devices in robotics, in the reduction and compression of widely distributed sensor data, and power management.

One of the leading lights of the first computer revolution saw this clearly. At the IEEE Centenary in 1984 (*The Next 100 Years*; IEEE Technical Convocation), Dr. Robert Noyce, the co-founder of Intel and co-inventor of the Integrated Circuit, noted that:

> *Until now we have been going the other way; that is, in order to understand the brain we have used the computer as a model for it. Perhaps it is time to reverse this reasoning: to understand where we should go with the computer, we should look to the brain for some clues.*

References

1 K. Likharev, CMOL: Freeing advanced lithography from the alignment accuracy burden, in: The International Conference on Electron, Ion, and Photon Beam Technology and Nanofabrication '07, Denver, 2007.

2 W. S. McCulloch, W. H. Pitts, A logical calculus of the ideas immanent in nervous activity, *Bull. Mathematical Biophys.* 1943, **5**, 115–133.

3 K. Steinbuch, Die Lernmatrix, *Kybernetik* 1961, 1.

4 F. Rosenblat, *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*, Spartan, New York, 1962.

5 L. Minsky, M. A. S. Papert, *Perceptrons: An introduction to computational geometry*, MIT Press, Cambridge MA, 1988.

6 SOAR.Web Page. http://sitemaker.umich.edu/soar.

7 J. Hopfield, Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. USA* 1982, **79**, 2554–2558.

8 D. Rumelhart, G. Hinton, R. Williams, Learning internal representations by error propagation, *Nature* 1986, **323**, 533–536.

9 P. J. Werbos, *The Roots of Backpropagation: From Ordered Derivatives to Neural Networks and Political Forecasting*, Wiley-Interscience, 1994.

10 G. Palm, F. Schwenker, F. T. Sommer, A. Strey, Neural associative memories, in: A. Krikelis, C. C. Weems (Eds.), *Associative Processing and Processors*, IEEE Computer Society, Los Alamitos, CA, 1997, pp. 284–306.

11 A. Sandberg, A. Lansner, K.-M. Petersson, Ö. Ekeberg, Bayesian attractor networks with incremental learning, *Network: Computation in Neural Systems* 2002, **13** (2), 179–194.

12 S. Zhu, Associative memory as a primary component in cognition, PhD Dissertation (in preparation) CSEE Department, School of Science and Engineering, Oregon Health & Science University, Portland, OR.

13 D. Hammerstrom, Neural networks at work, *IEEE Spectrum* 1993, 26–32.

14 D. Hammerstrom, Working with neural networks, *IEEE Spectrum* 1993, 46–53.

15 J. Lazzaro, J. Wawrzynek, Speech recognition experiments with silicon auditory models, *Analog Integrated Circ. Signal Proc.* 1997, **13**, 37–51.

16 S. Haykin, B. Kosko (Eds.), Intelligent signal processing. *Proceedings of the IEEE*, Volume 86, IEEE Press 1989.

17 L. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, *Proceedings IEEE* 1989, **77** (2), 257–286.

18 H. Bourlard, N. Morgan, *Connectionist Speech Recognition – A Hybrid Approach*, Kluwer Academic Publishers, Boston, MA, 1994.

19 J. L. Hennessy, D. A. Patterson, *Computer Architecture: A Quantitative Approach*, Morgan Kaufmann, Palo Alto, CA, 1991.

20 Intel. *IA-32 Intel Architecture Software Developer's Manual, Volume 1: Basic Architecture*. 2001 (cited 2001; Available from: http://developer.intel.com/design/pentium4/manuals/245470.htm)

21 C. Mead, *Analog VLSI and Neural Systems*, Addison-Wesley, Reading, Massachusetts, 1989.

22 M. A. Mahowald, *Computation and Neural Systems*, California Institute of Technology, 1992.

23 R. Douglas, Fifteen years of neuromorphic engineering: Progress, problems, and prospects, in: Proceedings, Brain Inspired Cognitive Systems – BICS2004, University of Stirling, Scotland, UK, 2006.

24 M. Holler, *et al.*, An electrically trainable artificial neural network (ETANN) with 10240 'floating gate' synapses, in: *Proceedings, International Joint Conference on Neural Networks*, IEEE, Washington DC, 1989.

25 I. Nestor, Ni1000 Recognition Accelerator – Data Sheet, 1996, 1–7. Available at: http://www.warthman.com/projects-intel-ni1000-TS.htm.

26 M. J. L. Orr, *Introduction to Radial Basis Function Networks*, Centre for Cognitive Science, University of Edinburgh, Edinburgh, 1996.

27 L. O. Chua, T. Roska, *Cellular Neural Networks and Visual Computing*, Cambridge University Press, 2002.

28 D. Balya, B. Roska, T. Roska, F. S. Werblin, A CNN framework for modeling parallel processing in a mammalian retina, *Int. J. Circuit Theory Applications* 2002, **30**, 363–393.

29 A. F. Murray, The future of analogue neural VLSI, in: *Proceedings Second International ICSC Symposium on Intelligent Systems for Industry*, The International ICSC Congress, June 26–29, Paisley, UK, 2001.

30 H. P. Graf, L. D. Jackel, W. E. Hubbard, VLSI implementation of a neural network model. *IEEE Computer* 1988, **21** (3), 41–49.

31 E. Culurciello, R. Etienne-Cummings, K. Boahen, An address event digital imager. *IEEE J. Solid-State Circuits* 2003, **38** (2), 505–508.

32 Y. N. Rao, D. Erdogmus, G. Y. Rao, J. C. Principe, Stochastic error whitening algorithm for linear filter estimation with noisy data, *Neural Networks Archive: Special issue: Advances in Neural Networks Research* 2003, **16** (5–6), 873–880.

33 T. Y. W. Choi, *et al.*, Neuromorphic implementation of orientation hypercolumns. *IEEE Trans. Circuits Systems II: Analog Digital Signal Proc.* 2005, **52** (6), 1049–1060.

34 D. Hammerstrom, A VLSI architecture for high-performance, low-cost, on-chip learning, in: *International Joint Conference on Neural Networks*, IEEE Press, San Diego, 1990.

35 D. Hammerstrom, A digital VLSI architecture for real-world applications, in: S. F. Zornetzer,*et al.* (Eds.), *An Introduction to Neural and Electronic Networks*, Academic Press, San Diego, CA, 1995, pp. 335–358.

36 J. L. Hennessy, D. A. Patterson, *Computer Architecture: A Quantitative Approach*, 3rd edn, Morgan Kaufmann, Palo Alto, CA, 2002.

37 U. Ramacher, W. Raab, J. Anlauf, U. Hachmann, J. Beichter, N. Brüls, M. Weißling, E. Schneider, R. Männer, J. Gläß, Multiprocessor and memory architecture of the neurocomputer SYNAPSE-1, in: *Proceedings, World Congress on Neural Networks*, INNS Press, Portland, Oregon, Volume 4, pp. 775–778, 1993.

38 R. Means, L. Lisenbee, Extensible linear floating point SIMD neurocomputer array processor, in: *Proceedings, International*

Joint Conference on Neural Networks, IEEE Press, Seattle, Washington, 1991.

39 J. Wawrzynek, K. Asanovic, B. Kingsbury, J. Beck, D. Johnson, N. Morgan, Spert-II: A vector microprocessor system. *IEEE Computer* 1996, 79–86.

40 L. Bengtsson, *et al*. The REMAP Reconfigurable Architecture: a Retrospective, in: A. R. Omondi, J. C. Rajapakse (Eds.), *FPGA Implementations of Neural Networks*, Springer-Verlag, 2006.

41 V. Mountcastle, *Perceptual Neuroscience – The Cerebral Cortex*, Harvard University Press, Cambridge, MA, 1998.

42 V. B. Mountcastle, An organizing principle for cerebral function: the unit model and the distributed system, in: G. M. Edelman, V. B. Mountcastle (Eds.), *The Mindful Brain*, MIT Press, Cambridge, MA, 1978.

43 V. Braitenberg, A. Schüz, *Cortex: Statistics and Geometry of Neuronal Connectivity*, Springer-Verlag, Berlin, 1998.

44 C. Johansson, M. Rehn, A. Lansner, Attractor neural networks with patchy connectivity, *Neurocomputing* 2006, **69**, 627–633.

45 R. Hecht-Nielsen, A theory of thalamocortex, in: R. Hecht-Nielsen, T. McKenna (Eds.), *Computational Models for Neuroscience – Human Cortical Information Processing*, Springer, London, 2003.

46 R. R. Granger, Engines of the brain: The computational instruction set of human cognition. *AI Magazine* 2006, **27** (2), 15–32.

47 R. Granger, *et al*., Non-Hebbian properties of LTP enable high-capacity encoding of temporal sequences. *Proc. Natl. Acad. Sci. USA* 1994, **91**, 10104–10108.

48 D. George, J. Hawkins, Invariant pattern recognition using Bayesian inference on hierarchical sequences, in: *Proceedings, 2005 IEEE International Joint Conference on Neural Networks*, Volume 3, pp. 1812–1817, 2005.

49 S. Grossberg, Adaptive resonance theory, in: *The Encyclopedia of Cognitive Science*, Macmillan Reference Ltd, London, 2003.

50 A. Lansner, A. Holst, A higher order Bayesian neural network with spiking units. *Int. J. Neural Systems* 1996, **7** (2), 115–128.

51 C. Johansson, A. Lansner, Towards cortex-sized artificial nervous systems, in: *Knowledge-Based Intelligent Information and Engineering Systems KES'04*, WelTec-Springer, Wellington, New Zealand, 2004.

52 M. Arbib, Towards a neurally-inspired computer architecture. *Natural Computing* 2003, **2** (1), 1–46.

53 Y. Roudi, A. Treves, An associative network with spatially organized connectivity, *J. Stat. Mech.: Theor. Exp.* 2004, **2004**, P07010.

54 A. Renart, N. Parga, E. T. Rolls, Associative memory properties of multiple cortical modules, *Network: Comput. Neural Syst.* 1999, **10**, 237–255.

55 N. Levy, D. Horn, E. Ruppin, Associative memory in a multimodular network, *Neural Computation* 1999, **11**, 1717–1737.

56 J. A. Anderson, P. Allopenna, G. S. Guralnik, D. Scheinberg, J. A. Santini, S. Dimitriadis, B. B. Machta, B. T. Merritt, Programming a parallel computer: The Ersatz Brain Project, in: W. Duch, J. Mandziuk, J. M. Zurada (Eds.), *Challenges to Computational Intelligence*, Springer, Berlin, 2006 (in press).

57 J. Bailey, A VLSI interconnect strategy for biologically inspired artificial neural networks, PhD Thesis, Computer Science/ Engineering Department, Oregon Graduate Institute, Beaverton, OR, 1993.

58 J. Bailey, D. Hammerstrom, Why VLSI implementations of associative VLCNs require connection multiplexing, *Proceedings, International Conference on Neural Network*, 1988, pp. 173–180.

59 D. Hammerstrom, The connectivity requirements of simple association, or How many connections do you need? in: *IEEE Conference on Neural Network Information Processing*, IEEE Press, 1987.

60 E. Means, D. Hammerstrom, Piriform model execution on a neurocomputer, in:

International Joint Conference on Neural Networks, Seattle, WA, 1991.

61 K. A. Boahen, Point-to-point connectivity between neuromorphic chips using address events. *IEEE Trans. Circuits Systems II - Analog Digital Signal Proc.* 2000, **47** (5), 416–434.

62 J. P. Lazzaro, J. Wawrzynek, A multi-sender asynchronous extension to the address-event protocol, in: *Proceedings, 16th Conference on Advanced Research in VLSI*. IEEE Computer Society, Washington, DC, USA, p. 158, 1995.

63 D. Hammerstrom, J. Bailey, Neural-model, computational architecture employing broadcast hierarchy and hypergrid, point-to-point communication. US Patent No. 4,983,962, issued January 8, 1991.

64 J. H. Lee, K. K. Likharev, *CMOL CrossNets as Pattern Classifiers*, Stony Brook University, Stony Brook, 2005.

65 K. K. Likharev, D. V. Strukov, CMOL: Devices, circuits, and architectures, in: G. Cuniberti, *et al.* (Eds.), *Introducing Molecular Electronics*, Springer, Berlin, 2004.

66 D. B. Strukov, K. K. Likharev, Defect-tolerant architectures for nanoelectronic crossbar memories. *J. Nanosci. Nanotechnol.* 2007, **7**, 151–167.

67 G. S. Snider, R. S. Williams, Nano/CMOS architectures using a field-programmable nanowire interconnect, *Nanotechnology* 2007, **18**, 035204.

68 K. K. Likharev, D. B. Strukov, CMOL: Devices, Circuits, and Architectures, in: G. Cuniberti, *et al.* (Eds.), *Introduction to Molecular Electronics*, Springer, Berlin, pp. 447–477, 2004.

69 A. Lansner, *et al.* Detailed simulation of large scale neural networks, in: J. M. Bower (Ed.), *Computational Neuroscience: Trends in Research 1997*, Plenum Press, Boston, MA, 1997, pp. 931–935.

70 R. Ananthanarayanan, D. S. Modha, *Anatomy of a Cortical Simulator,* in: *Super Computing 2007 (SC07)*, EEE Press, Reno, Nevada, 2007.

71 W. Maass, Computing with spiking neurons, in: W. Maass, C. M. Bishop (Eds.), *Pulsed Neural Networks*, MIT Press, A Bradford Book, Cambridge, MA, 1999.

72 M. Hänggi, Available from: http://www.ce.unipr.it/pardis/CNN/cnn.html.

11
Nanowire-Based Programmable Architectures

André DeHon

11.1
Introduction

Today, chemists are demonstrating bottom-up synthesis techniques which can construct atomic-scale switches, field-effect devices, and wires (see Section 11.2). While these are key components of a computing system, it must also be understood if these can be assembled and organized into useful computing devices. That is, can arbitrary logic be built from nanowire building blocks and atomic-scale switches?

- Do we have an adequate set of capabilities to build logic?
- How do we cope with the regularity demanded by bottom-up assembly?
- How do we accommodate the high defect rates and statistical assembly which accompany bottom-up assembly techniques?
- How do we organize and interconnect these atomic-scale building blocks?
- How do we address nanowires from the lithographic scale for testing, configuration, and IO?
- How do we get logic restoration and inversion?
- What net benefit do these building blocks offer us?

The regular synthesis techniques can be used to assemble tight-pitch, parallel nanowires; this immediately suggests that programmable crossbar arrays (Section 11.4.1) are built as the key building blocks in these architectures. These crossbar arrays can be used as memory cores (Section 11.5), wired-OR logic arrays (Section 11.6.1), and programmable interconnect (Section 11.6.3) – memory, logic, and interconnect – all of which are the key components needed for computation.

The length of the nanowires must be limited for yield, performance, and logical efficiency. Consequently, the nanowires are organized into a collection of

Nanotechnology. Volume 4: Information Technology II. Edited by Rainer Waser

ISBN: 978-3-527-31737-0

modest-sized, interconnected crossbar arrays (Section 11.6.3). A reliable, lithographic-scale support structure provides power, clocking, control, and bootstrap testing for the nanowire crossbar arrays. Each nanowire is coded so that it can be uniquely addressed from the lithographic support wires (Section 11.4.2). With the ability to address individual nanowires, individual crosspoints can be programmed (Section 11.8) to personalize the logic function and routing of each array and to avoid defective nanowires and switches (Section 11.7).

As specific nanowires cannot, deterministically, be placed in precise locations using these bottom-up techniques, stochastic assembly is exploited to achieve unique addressability (Section 11.4.2). Stochastic assembly is further exploited to provide signal restoration and inversion at the nanoscale (Section 11.4.3). Remarkably, starting from regular arrays of programmable diode switches and stochastic assembly of non-programmable field-effect controlled nanowires, it is possible to build fully programmable architectures with all logic and restoration occurring at the nanoscale.

The resulting architectures (Section 11.6) provide a high-level view similar to island-style field-programmable gate arrays (FPGAs), and conventional logic mapping tools can be adapted to compile logic to these arrays. Owing to the high defect rates likely to be associated with *any* atomic-scale manufacturing technology, all viable architectures at this scale are likely to be post-fabrication configurable (Section 11.7). That is, while nanowire architectures can be customized for various application domains by tuning their gross architecture (e.g. ratio of logic and memory), there will be no separate notion of custom atomic-scale logic.

Even after accounting for the required, regular structure, high defect rates, stochastic assembly, and the lithographic support structure, a net benefit is seen from being able to build with nanowires which are just a few atoms in diameter and programmable crosspoints that fit in the space of a nanowire junction. Mapping conventional FPGA benchmarks from the Toronto20 benchmark set [1], the designs presented here should achieve one to two orders of magnitude greater density than FPGAs in 22 nm CMOS lithography, even if the 22 nm lithography delivers defect-free components (Section 11.10).

The design approach taken here represents a significant shift in design styles compared to conventional lithographic fabrication. In the past, reliance has been placed on virtually perfect and deterministic construction and complete control of features down to a minimum technology feature size. Here, it is possible to exploit very small feature sizes, although there is no complete control of device location in all dimensions. Instead, it is necessary to rely on the statistical properties of large ensembles of wires and devices to achieve the desired, aggregate component features. Further, post-fabrication configuration becomes essential to device yield and personalization.

This chapter describes a complete assembly of a set of complementary technologies and architectural building blocks. The particular ensemble presented is one of several architectural proposals which have a similar flavor (Section 11.11) based on these types of technologies and building blocks.

11.2 Technology

11.2.1 Nanowires

Atomic-scale nanowires (NWs) can be engineered to have a variety of conduction properties, from insulating to semiconducting to metallic. The composition of a NW can be varied along its axis and along its radius, offering powerful heterostructures to provide both controllable devices and interconnect integrated into a single structure.

Seed catalysts are used to control the diameter of a NW during the composition process and constrain the growth to a small region [2, 3].

In addition, semiconducting can be doped during the growth process by controlling the mix of elements in the ambient environment [4]. This can produce conducting NWs with heavy doping and field effect controllable NWs with a suitably light doping [5].

The doping profile or material composition can change along the length of a NW by controlling the ambient process environment over the time [6–8]. This leads to properties such as gateable and not-gateable regions within a single NW (Figure 11.1). After the axial growth, the NW's surface can be used as a substrate for atomic layer growth to produce a radial material composition (Figure 11.2), for example with SiO_2 as an insulator and spacer [9–11].

In order to increase the conductivity of NWs beyond heavily doped semiconductors, nickel silicide (NiSi) can be generated by coating selected regions with nickel and subsequently annealing the area [12].

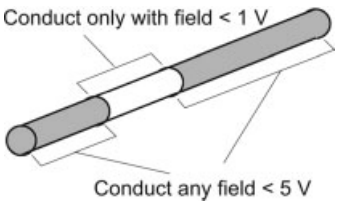


Figure 11.1 Axial doping profile places selective gateable regions in a nanowire.

Figure 11.2 Radial doping profile.

11.2.2
Assembly

Langmuir–Blodgett (LB) flow techniques can be used to align a set of NWs into a single orientation, close pack them, and transfer them onto a surface [11, 13]. The resulting wires are all parallel with nematic alignment. By using wires with an oxide sheath around the conducting core, the wires can be packed tightly. The oxide sheath defines the spacing between conductors and can, optionally, be etched away after assembly. The LB step can be rotated and repeated so that multiple layers of NWs are obtained [11, 13] such as crossed NWs for building a crossbar array or memory core (see Section 11.4.1).

11.2.3
Crosspoints

Many technologies have been demonstrated for non-volatile, switched crosspoints. Common features include:

- resistance which changes significantly between on and off states;
- the ability to be made rectifying;
- the ability to turn the device on or off by applying a large voltage differential across the junction;
- the ability to operate at a low voltage differential without switching the device state; and
- the ability to be placed within the area of a crossed NW junction.

Chen *et al*. [14] demonstrated a nanoscale Ti/Pt-[2]rotaxane-Ti/Pt sandwich which exhibits hysteresis and non-volatile state storage showing an order of magnitude resistance difference between on and off states for several write cycles. With 1600 nm^2 junctions, the on resistance ($R_{ondiode}$) was roughly 500 KΩ, and the off resistance ($R_{offdiode}$) 9 MΩ. After an initial burn-in step, the state of these devices can be switched at ±2 V and read at ±0.2 V. The basic hysteretic molecular memory effect is not unique to the [2]rotaxane, and the junction resistance is continuously tunable [15]. The exact nature of the physical phenomena involved is the subject of active investigation.

In conventional very large-scale integration (VLSI), the area of an SRAM-based programmable crosspoint switch is much larger than the area of a wire crossing. A typical, CMOS switch might be $2500\lambda^2$ [16], compared to a $5\lambda \times 5\lambda$ bottom-level metal wire crossing, making the crosspoint 100-times the area of the wire crossing. Consequently, the nanoscale crosspoints offer an additional device size reduction beyond that implied by the smaller NW feature sizes. This particular device size benefit reduces the overhead for configurability associated with programmable architectures [e.g. FPGAs, programmable logic arrays (PLAs)] in this technology compared to conventional CMOS.

11.2.4
Technology Roundup

It is possible to create wires which are nanometers in diameter and which can be arranged into crossbar arrays with nanometer pitch. Crosspoints which both switch conduction between the crossed wires and store their own state can be placed at every wire crossing without increasing the pitch of the crossbar array. NWs can be controlled in FET-like manner, and can be designed with selectively gateable regions. This can all be done without relying on ultrafine lithography to create the nanoscale feature sizes. Consequently, these techniques promise smaller feature sizes and an alternate – perhaps more economical – path to atomic-scale computing structures than top-down lithography. Each of the capabilities previously described has been demonstrated in a laboratory setting as detailed in the reports cited. It is assumed that, in future, it will be possible to combine these capabilities and to scale them into a repeatable manufacturing process.

11.3
Challenges

In the top-down lithographic model, a minimum, lithographically imageable feature size is defined, and devices are built that are multiples of this imageable feature size (e.g. half-pitch). Within the limits of this feature size, the size of features and their relative location to each other in three dimensions – both in the two-dimensional (2-D) plane of each lithographic layer and with adequate registration between layers – could be perfectly specified. This provided complete flexibility in the design of circuit structures as long as the minimum imageable and repeatable feature size rules were adhered to.

When approaching the atomic-scale, it becomes increasingly difficult to maintain this model. The precise location of atoms becomes relevant, and the discreteness of the underlying atoms begins to show up as a significant fraction of feature size. Variations occur due to statistical doping and dopant placement and interferometric mask patterning. Perfect repeatability may be extremely difficult or infeasible for these feature sizes.

These bottom-up approaches, in contrast, promise finer feature sizes that are controlled by physical phenomena but do not promise perfect, deterministic alignment in three dimensions. It may be possible to achieve good repeatability of certain types of small feature sizes (e.g. NW diameters) and correlation of tiny features within a single NW using axial and radial composition, but there may be little correlation from NW to NW in the plane or between NW planes. This may prompt the question of whether it would be reasonable to forego the perfect correlation and complete design freedom in three dimensions in order to exploit smaller feature sizes. The techniques summarized here suggest that this is a viable alternative.

11.3.1 Regular Assembly

The assembly techniques described above (see Sections 11.2.2 and 11.2.3) suggest that regular arrays can be built at tight pitch with both NW trace width and trace spacing using controlled NW diameters. While this provides nanometer pitches and crosspoints that are tens of nanometers in area, it is impossible to differentiate deterministically between features at this scale; that is, one particular crosspoint cannot be made different in some way from the other crosspoints in the array.

11.3.2 Nanowire Lengths

Nanowire lengths can be grown to hundreds of microns [17] or perhaps millimeters [18] in length. However, at this high length to diameter ratio, they become highly susceptible to bending and ultimately breaking. Assembly puts stresses along the NW axis which can break excessively long NWs. Consequently, a modest limit must be placed on the NW lengths (tens of microns) in order to yield a large fraction of the NWs in a given array. Gudiksen *et al.* [19] reported the reliable growth of Si NWs which are over 9 μm long, while Whang et al. [11, 20] demonstrated collections of arrays of NWs of size 10 μm × 10 μm. Even if it was possible physically to build longer NWs, the high resistivity of small-diameter NWs would force the lengths to be kept down to the tens of microns range.

11.3.3 Defective Wires and Crosspoints

At this scale, wires and crosspoints are expected to be defective in the 1 to 10% range:

- NWs may break along their axis during assembly as suggested earlier, and the integrity of each NW depends on the ~100 atoms in each radial cross-section.
- NW to microwire junctions depend on a small number of atomic scale bounds which are statistical in nature and subject to variation in NW properties.
- Junctions between crossed NWs will be composed of only tens of atoms or molecules, and individual bond formation is statistical in nature.
- Statistical doping of NWs may lead to high variation among NWs.

For example, Huang *et al.* [13] reports that 95% of the wires measured had good contacts, while Chen *et al.* [21] reported that 85% of crosspoint junctions measured were usable. Both of these were early experiments, however, and the yield rates would be expected to improve. Nonetheless, based on the physical phenomena involved it is anticipated that the defect rates will be closer to the few percent range than the minuscule rates frequently seen with conven-

tional, lithographic processing. Consequently, two main defect types may be considered:

- *Wire defects:* a wire is either functional or defective. A functional wire has good contacts on both ends, conducts current with a resistance within a designated range, and is not shorted to any other NWs. Broken wires will not conduct current. Poor contacts will increase the resistance of the wire, leaving it outside of the designated resistance range. Excessive variation in NW doping from the engineered target can also leave the wire out of the specified resistance range. It can be determined if a wire is in the appropriate resistance range during testing (see Section 11.8.1) and arranged not to use those which are defective (see Section 11.7.1).
- *Non-programmable crosspoint defects:* a crosspoint is programmable, non-programmable, or shorted into the on state. A programmable junction can be switched between the resistance range associated with the on-state and the resistance range associates with the off-state. A non-programmable junction can be turned off, but cannot be programmed into the on-state; a non-programmable junction could result from the statistical assembly of too few molecules in the junction, or from poor contacts between some of the molecules in the junction and either of the attached conductors. A shorted junction cannot be programmed into the off-state. Based on the physical phenomena involved, non-programmable junctions are considered to be much more common than shorted junctions. Further, it is expected that fabrication can be tuned to guarantee that this is the case. Consequently, shorted junctions will be treated like a pair of defective wires, and both wires associated with the short will be avoided.

Currently, the bridging of adjacent NWs is considered NOT to be a major defect source. Radial shells around the (semi) conducting NW cores prevent the shorting of adjacent NWs. At present, there is insufficient experience to determine if variations in core shell thickness, imperfect planar NW alignment, or other effects may, nonetheless, lead to bridging defects between adjacent NWs. If such bridging were to occur, it could make a pair of NWs indistinguishable, perhaps effectively giving two addresses to the NW pair. These bridged NW pairs could be detected and avoided but their occurrence would necessitate slightly more complicated testing and verification algorithms than those detailed in Section 11.8.

After describing the building blocks and architecture, the way in which the two main defect types within the architecture are accommodated is described in Section 11.7.

11.4 Building Blocks

By working from the technological capabilities and within the regular assembly requirements, it is possible to construct a few building blocks which enable the creation of a wide range of interesting programmable architectures.

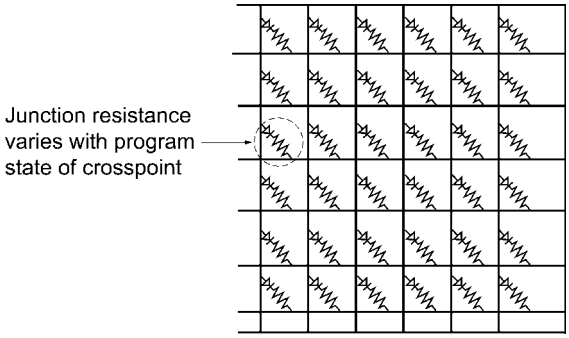


Figure 11.3 Logical diode crossbar formed by crossed nanowires.

11.4.1 Crosspoint Arrays

As suggested in Section 11.2.2 and demonstrated by Chen *et al.* [21] and Wu *et al.* [22], assembly processes allow the creation of tight-pitch arrays of crossed NWs with switchable diodes at the crosspoints (see Figure 11.3). Assuming for the moment that contact can be made to individual NWs in these tight-pitch arrays (see Section 11.4.2), these arrays can serve as:

- memory cores
- programmable, wired-OR planes
- programmable crossbar interconnect arrays.

11.4.1.1 Memory Core

As noted in Section 11.2.3, by applying a large voltage across a crosspoint junction, the crosspoint can be switched into a high or low resistance state. Consequently, if the voltage on a single row and a single column line can be set to desired voltages, each of the crosspoints can be set to a particular conduction state. It is further noted that the system can operate at a lower voltage without resetting the crosspoint. Consequently, a crosspoint's state can be read back by applying a small, test voltage to a column input and observing the current flow, or rate of charging, of a row line to tell if the crosspoint has been set into a high or low resistance state.

11.4.1.2 Programmable, Wired-OR Plane

When a method of programming the crosspoints into high or low resistance states has been effected, the OR logic can be programmed into a crosspoint array. Each row output NW serves as a wired-OR for all of the inputs programmed into the low resistance state. Consider a single row NW, and assume for the moment that the means is available to pull a non-driven NW down to ground. Now, if any of the column NWs which cross this row NW are connected with low resistance crosspoint junctions and are driven to a high voltage level, the current into the column NW will be able to flow into the row NW and charge the row NW up to a higher voltage value (see O1, O3, O4, and O5 in Figure 11.4). However, if none of the connected column NWs is high,

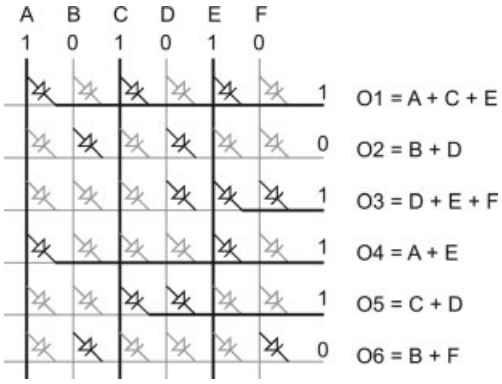


Figure 11.4 Wired-OR plane operation. Programmed on crosspoints are shown in black; off crosspoints are shown in gray. Dark lines represent a nanowire (NW) pulled high, while light lines remain low. Output NWs are marked dark, starting at the diode that pulls them high, in order to illustrate current flow; the entire output NW would be pulled high in actual operation.

the row NW will remain low (see O2 and O6 in Figure 11.4). Consequently, the row NW effectively computes the OR of its programmed inputs.

The output NWs do pull their current directly off the inputs and may not be driven as high as the input voltage. Hence, these outputs will need restoration (see Section 11.4.3).

11.4.1.3 **Programmable Crossbar Interconnect Arrays**

A special use of the Wired-OR programmable array is for interconnect. That is, if a restriction is introduced to connecting a *single* row wire to each column wire, the crosspoint array can serve as a crossbar switch. This allows any input (column) to be routed to any output (row) (e.g. see Figure 11.5). This structure is useful for

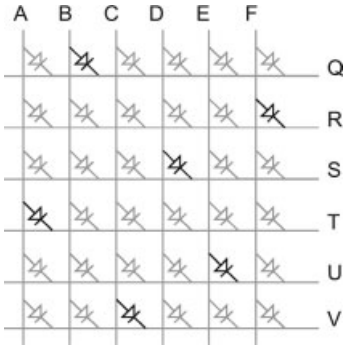


Figure 11.5 Example of crossbar routing configuration. Programmed on crosspoints are shown in black; off crosspoints are shown in gray. Here, the crossbar is shown programmed to connect A → T, B → Q, C → V, D → S, E → U, and F → R.

post-fabrication programmable routing to define a logic function and to avoid defective resources (see Section 11.3.3).

11.4.2 Decoders

A key challenge is bridging the length scale between the lithographic-scale wires that can be created using conventional top-down lithography and the small-diameter NWs that can be grown and assembled into tight-pitch arrays. As noted above, it must be possible to establish a voltage differential across a single row and column NW to write a bit in the tight-pitch NW array. It must also be possible to drive and sense individual NWs to read back the memory bit. By building a decoder between the coarse-pitch lithographic wires and the tight-pitch NWs, it is possible to bridge this length scale and to address a single NW at this tight pitch [23–26].

11.4.2.1 NW Coding

One way to build such a decoder is to place an address on each NW using the axial doping or material composition profile described previously. In order to interface with lithographic-scale wires, address bit regions are marked off at the lithographic pitch. Each such region is then either doped heavily so that it is oblivious to the field applied by a crossed lithographic-scale wire, or is doped lightly so that it can be controlled by a crossed lithographic scale wire. In this way, the NW will only conduct if all of the lithographic-scale wires crossing its lightly doped, controllable regions have a suitable voltage to allow conduction. If any of the lithographic-scale wires crossing controllable regions provide a suitable voltage to turn off conduction, then the NW will not be able to conduct.

It should be noted that each bit position can only be made controllable or non-controllable with respect to the lithographic-scale control wire; different bit positions cannot be made sensitive to different polarities of the input. Consequently, the addresses must be encoded differently from the dense, binary address normally used for memories. One simple way to generate addresses is to use a dual-rail binary code. That is, for each logical input bit, the value and its complement are provided. This results in two bit positions on the NW for each logical input address bit – one for the true sense and one for the false sense. To code a NW with an address, either bit position is simply coded to be sensitive to exactly one sense of each of the bit positions (see Figure 11.6). This results in a decoder which requires $2\log_2(N)$ address bits to address N NWs.

Denser addressing can be achieved by using $N_a/2$-hot codes; that is, rather than forcing one bit of each pair of address lines to be off and one to be on, it is simply required that half of the address bits, N_a, be set to a voltage which allows conduction, and half to be set to a voltage that prevents conduction. This scheme requires only $1.1\log_2(N) + 3$ address bits [24].

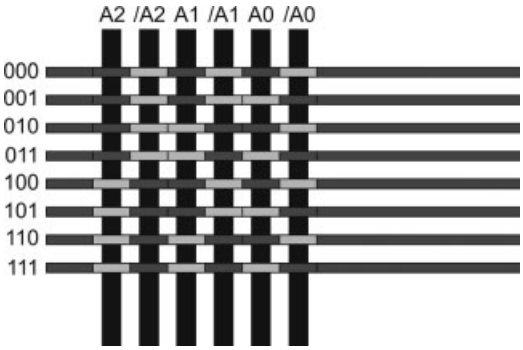


Figure 11.6 Dual-rail address coding.

11.4.2.2 **Decoder Assembly**

If each NW in the array has a unique address in the selected coding scheme, then each individual NW in the array can be uniquely addressed. However, the NW assembly techniques do not allow particular NWs to be placed in particular locations – it can only be arranged to create a tight-pitch parallel ensemble of a collection of NWs.

Instead, it appears that if the code space for the NWs is made large compared to the size of the array to be addressed, it can be guaranteed statistically (with arbitrarily high probability) that every NW in an array has a unique address. Starting with a very large number of NW codes, the NWs can be mixed up before assembly so that a random selection occurs of which NW codes go into each of the crosspoint arrays being assembled. As long as the array formed is sufficiently small compared to the code space for the NWs, strong guarantees can be provide that each array contains NWs with unique codes [24]. It transpires that there is no need for a large number of address bits in order to guarantee this uniqueness. For example, the $N_a/2$-hot codes need a total of only $\lceil 2.2 \log_2(N) \rceil + 11$ bits to achieve over a 99% probability that all NWs in an array will have unique addresses. If a few duplicates are tolerable, the codes can be much tighter [25, 26].

Wires can be coded to tolerate misalignment during assembly. Hybrid addressing schemes which segment the collection of NWs in an array into lithographic-scale contact groups can be used to reduce the size of the NW codespace needed (for further details see Ref. [24]).

11.4.2.3 **Decoder and Multiplexer Operation**

Now it is known that uniquely coded NWs can be assembled into an array, it can be seen how the decoder operates. First, assume that all the NWs are either precharged or weakly pulled to a nominal voltage. The desired NW address is then applied to the lithographic-scale address lines. The desired drive voltage is also applied to a common line attached to all the NWs. If the selected address is present in the array, it will allow conduction from the common line into the array charging up the selected NW. All other NWs will differ in at least one bit position, and will thus be disabled by the address lines. Consequently, only the selected NW is charged strongly to the

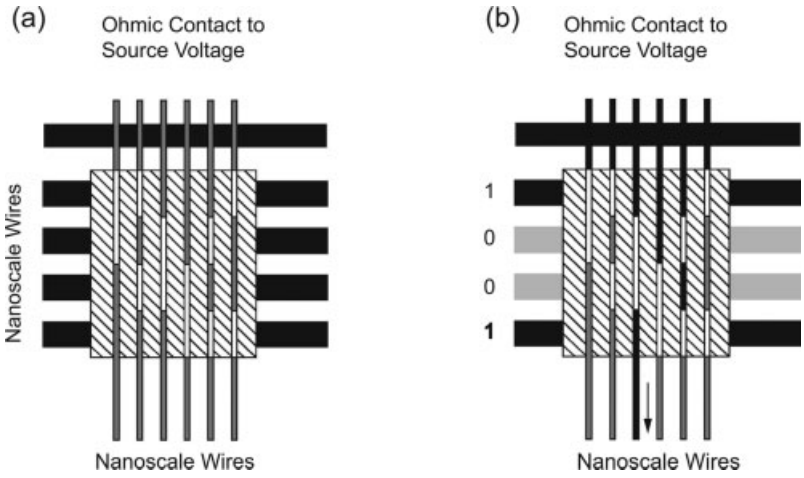


Figure 11.7 Coded NW decoder. (a) Decoder configuration: white NW regions are coded and controllable, while gray regions are not controllable and acts as wires. (b) Dark represents lines driven high; light gray shows lines low or undriven. Only the coding on the third line matches the applied address (1 0 0 1) and allows conduction. All other cases have a high address voltage crossing a lightly doped region, which prevents conduction.

voltage driven on the common line, and all other NWs are held at the nominal voltage (see Figure 11.7).

It should be noted that there is no directionality to the decoder, and consequently this same unit can serve equally well as a multiplexer. That is, when an address is applied to the lithographic-scale wires it allows conduction through the addressing region for only one of the NWs. Consequently, the voltage on the common line can be sensed rather than driven. Now, the one line which is allowed to conduct through the array can potentially pull the common line high or low. All other lines have a high resistance path across the lithographic-scale address wires and will not be able to strongly effect the common line. This allows a single NW to be sensed at a time (see Figure 11.8) as there is a need to read out the crosspoint state, as described in Section 11.4.1.1.

11.4.3 Restoration and Inversion

As noted in Section 11.4.1.2, the programmable, wired-OR logic is passive and non-restoring, drawing current from the input. Further, OR logic is not universal, and to build a good composable logic family an ability will be required to isolate inputs from output loads, restore signal strength and current drive, and invert signals.

Fortunately, NWs can be field-effect controlled, and this provides the potential to build FET-like gates for restoration. However, in order to realize these ways must be found to create the appropriate gate topology within the regular assembly constraints (see Section 11.3.1).

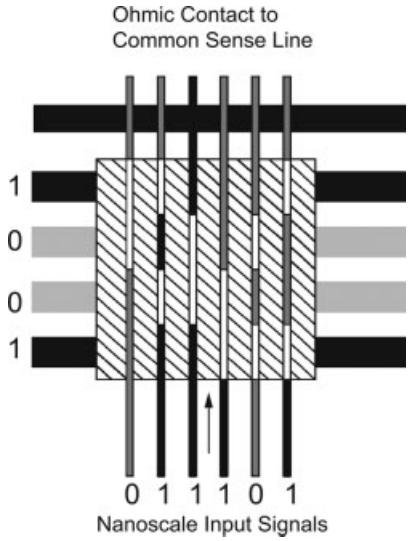


Figure 11.8 Coded NW multiplexer operation.

11.4.3.1 NW Inverter and Buffer

If two NWs are separated by an insulator, perhaps using an oxide core shell (see Section 11.2.1), then the field from one NW can potentially be used to control the other NW. Figure 11.9 shows an inverter which has been built using this basic idea. The horizontal NW serves as the input and the vertical NW as the output. This

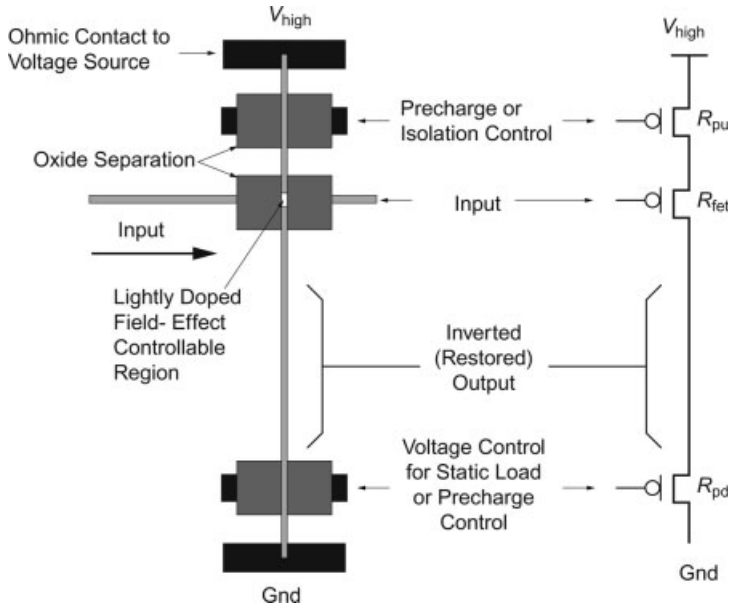


Figure 11.9 NW inverter.

gives a voltage transfer equation:

$$V_{\text{out}} = V_{\text{high}}\left(\frac{R_{\text{pd}}}{R_{\text{pd}} + R_{\text{fet}}(\text{input}) + R_{\text{pu}}}\right) \tag{11.1}$$

For the sake of illustration, the vertical NW has a lightly doped p-type depletion mode region at the input crossing forming a FET controlled by the input voltage (R_{fet}(Input)). Consequently, a low voltage on the input NW will allow conduction through the vertical NW ($R_{\text{fet}} = R_{\text{onfet}}$ is small), and a high input will deplete the carriers from the vertical NW and prevent conduction ($R_{\text{fet}} = R_{\text{offfet}}$ is large). As a result, a low input allows the NW to conduct and pull the output region of the vertical NW up to a high voltage. A high input prevents conduction and the output region remains low. A second crossed region on the NW is used for the pull down (R_{pd}); this region can be used as a gate for predischarge, so the inverter is pulled low before the input is applied, then left high to disconnect the pulldown voltage during evaluation. Alternately, it can be used as a static load for PMOS-like ratioed logic. By swapping the location of the high- and low-power supplies, this same arrangement can be used to buffer rather than invert the input.

Note that the gate only loads the input capacitively, and consequently current isolation is achieved at this inverter or buffer. Further, NW field-effect gating has sufficient non-linearity so that this gate provides gain to restore logic signal levels [27].

11.4.3.2 Ideal Restoration Array

In many scenarios, there is a need to restore a set of tight-pitch NWs such as the outputs of a programmable, wired-OR array. To do this, the approach would be to build a restoration array as shown in Figure 11.10a. This array is a set of crossed NWs

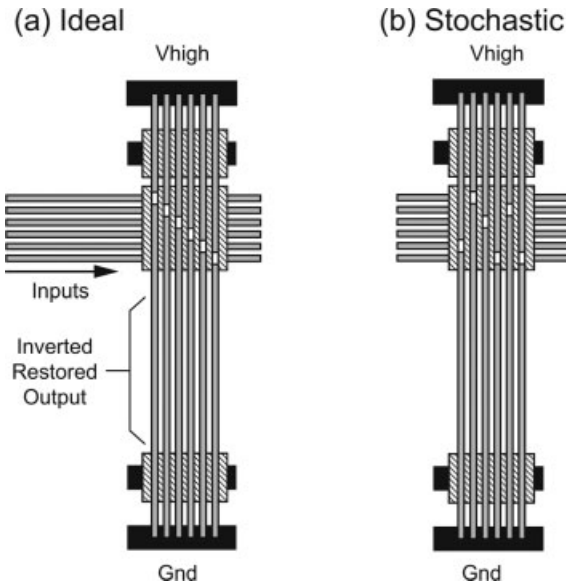


Figure 11.10 Restoration array.

which can be assembled using NW assembly techniques. If each of the NWs was sensitive to all of the crossed inputs, the result would be that all of the outputs would actually compute the NOR of the same set of inputs. To avoid computing a redundant set of NORs and instead simply to invert each of the inputs independently, these NWs are coded using an axial doping or material composition profile. In this way, each NW is field-effect sensitive to only a single NW, and hence provides the NW inversion described for a single one of the crossed NWs and is oblivious to the inputs of the other NWs.

The only problem here is that there is no way to align and place axially doped NWs so that they provide exactly this pattern, as the assembly treats all NWs as identical.

11.4.3.3 **Restoration Array Construction**

Although the region for active FETs is a nanoscale feature, it does not require small pitch or tight alignment. As such, there may be ways to mask and provide material differentiation along a diagonal as required to build this decoder.

Nonetheless, it is also possible to stochastically construct this restoration array in a manner similar to the construction of the address decoder. That is, an assembly is provided with a set of NWs with their restoration regions in various locations. The restoration array will be built by randomly selecting a set of restoration NWs for each array (see Figure 11.10b).

Two points differ compared to the address decoder case.

- The code space will be the same size as the desired restoration population.
- Duplication is allowed.

The question then is how large a fraction of the inputs will be successfully restored for a given number of randomly selected restoration NWs? This is an instance of the Coupon Collector Problem [28]. If the restoration array is populated with the same number of NWs as inputs, the array will typically contain restoration wires for 50–60% of the NW inputs. One way to consider this is that the array must be populated with 1.7- to 2-fold as many wires as would be hoped to yield due to these

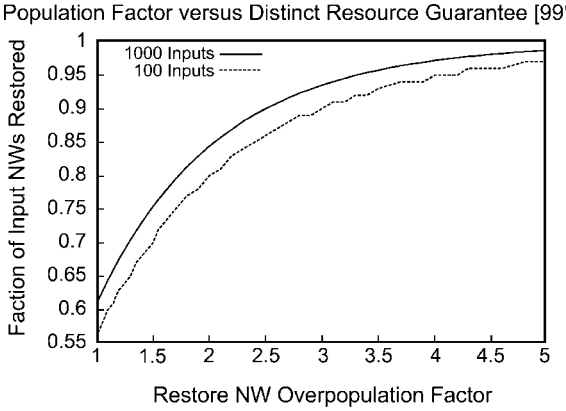


Figure 11.11 Fraction of input NWs restored as a function of restoration overpopulation.

stochastic assembly effects. If the number of restoration wires is increased relative to the number of input NWs, then a higher fraction of the inputs can be restored (as shown in Figure 11.11). For further details on these yield calculations, see Refs. [26, 29].

11.5 Memory Array

By combining the crosspoint memory cores with a pair of decoders, it is possible to build a tight-pitch, NW-based memory array [30]. Figure 11.12 shows how these elements come together in a small memory array, which is formed using crossed, tight-pitch NWs. Programmable diode crosspoints are assembled in the NW–NW crossings, while lithographic-scale address wires form row and column addresses. Write operations into the memory array can be performed by driving the appropriate write voltages onto a single row and column line. Read operations occur by driving a reference voltage onto the common column line, setting the row and column addresses, and sensing the voltage on the common row read line.

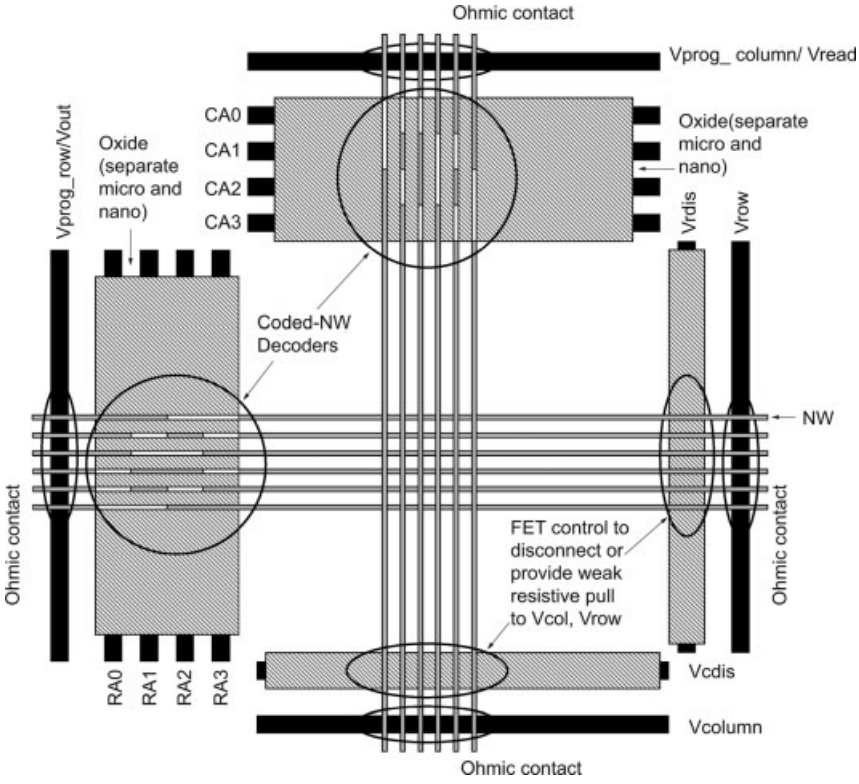


Figure 11.12 Memory array built from coded NW decoder and crosspoint memory core.

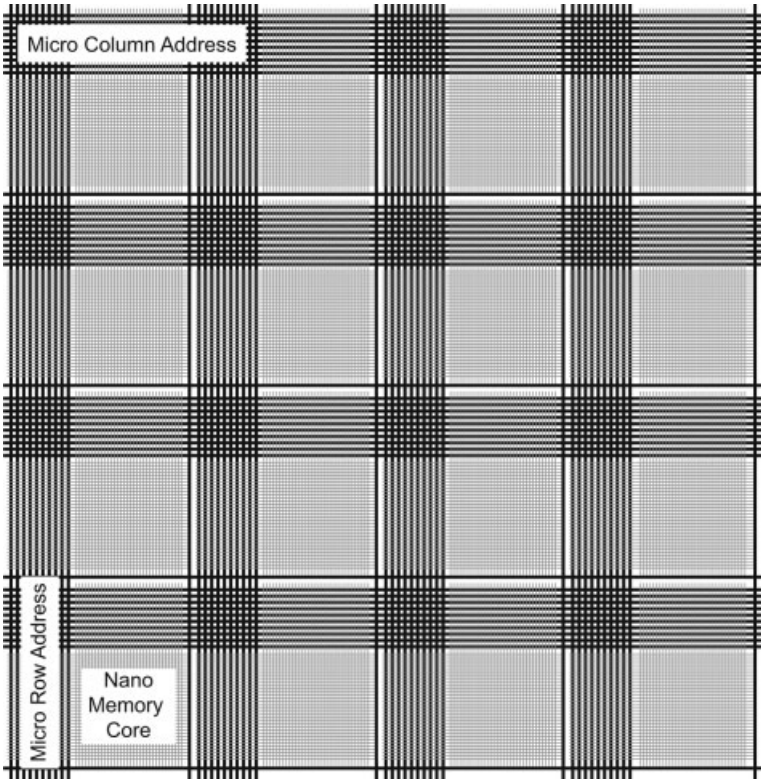


Figure 11.13 Tile of NW-based memory banks to construct large-scale memory.

Limitations on reliable NW length and the capacitance and resistance of long NWs prevent the building of arbitrarily large memory arrays. Instead, the large NW memories are broken up into banks similar to the banking used in conventional DRAMs (see Figure 11.13). Reliable, lithographic-scale wires provide address and control inputs and data inputs and outputs to each of the NW-based memory banks. The expected yield would be only a fraction of the NWs in the array due to wire defects. Error-correcting codes (ECC) can be used to tolerate non-programmable crosspoint defects. After accounting for defects, ECC overhead, and lithographic control overhead, net densities on the order of 10^{11} bits cm^{-2} appear achievable, using NW pitches of about 10 nm [29].

11.6 Logic Architecture

By combining the building blocks introduced in Section 11.4 it is possible to construct complete, programmable logic architectures with all logic, interconnect, and restoration occurring in the atomic-scale NWs. Diode crosspoints organized into Wired-OR logic arrays provide programmable logic, field-effect restoration arrays

provide gain and signal inversion, and the NWs themselves provide interconnect among arrays. Lithographic scale wires provide a reliable support infrastructure which allows device testing and programming (see Section 11.8), addressing individual NWs using the decoders introduced in Section 11.4.2. Lithographic-scale wires also provide power and control logic evaluation.

11.6.1 Logic

Figure 11.14 shows a simple PLA created using the building blocks from Section 11.4 and first introduced by DeHon and Wilson [31]. The design includes two interconnected logic planes, each of which is composed of a programmable Wired-OR array, followed by a restoration array. It should be noted here that two restoration arrays are actually used – one providing the inverted sense of the OR-term logic and one providing the non-inverted buffered sense. This arrangement is similar to conventional PLAs where the true and complement sense of each input is provided in each PLA plane. Since Wired-OR logic NWs can be inverted in this nanoPLA, each plane effectively serves as a programmable NOR plane. The combination of the two coupled NOR–NOR planes can be viewed as an AND–OR PLA with suitable application of DeMorgan's laws and signal complementation.

11.6.1.1 Construction

The entire construction is simply a set of crossed NWs as allowed by the regular assembly constraints (see Section 11.3.1). Lithographic-scale etches are used to differentiate regions (e.g. programmable-diode regions for the Wired-OR). The topology allows the same NWs that perform logic or restoration to carry their outputs across as inputs to the array that follows it.

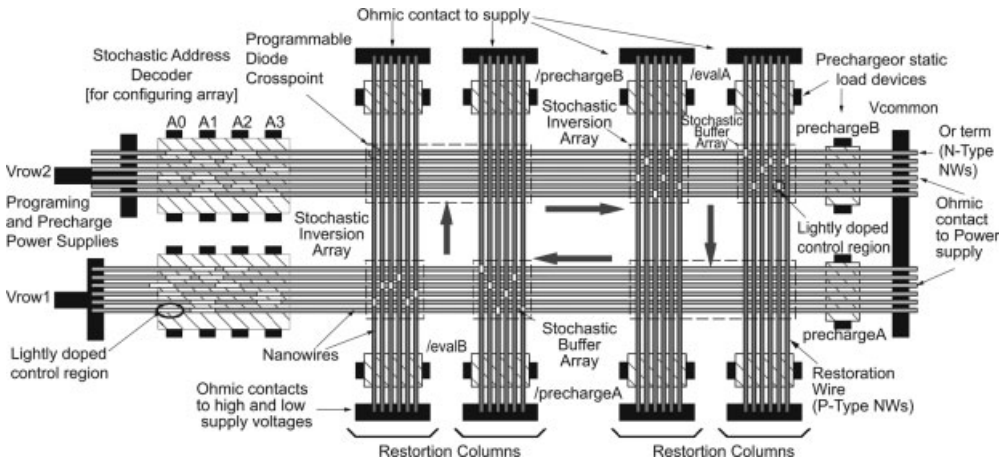


Figure 11.14 Simple nanoPLA block.

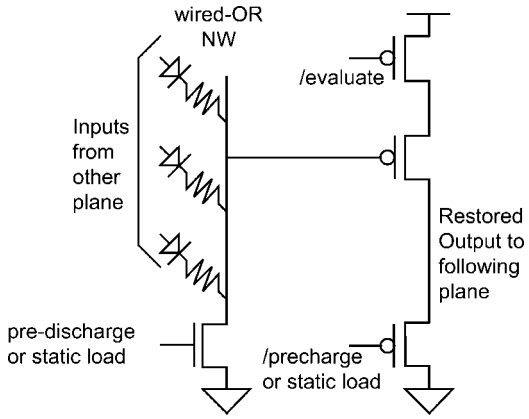


Figure 11.15 Rough circuit equivalent for each nanoPLA plane.

11.6.1.2 **Logic Circuit**

The logic gates in each PLA plane are composed of a diode-programmable Wired-OR NW, followed by a field-effect buffer or inverter NW (see Figure 11.15). The field-effect stage provides isolation as there is no current flow between the diode stage and the field-effect stage output. That is, the entire OR stage is capacitively loaded rather than resistively loaded. The OR stage simply needs to charge up its output which provides the field for the field-effect-based restoration stage. When the field is high enough (low enough for P-type NWs) to enable conduction in the field-effect stage, the NW will allow the source voltage to drive its output.

11.6.1.3 **Programming**

At the left-hand side of Figure 11.14 a decoder is formed (as introduced in Section 11.4.2) using the vertical microscale wires A0 to A3. These lithographic-scale wires allow the selection of individual NWs for programming. Each usable vertical restoration NW is driven by a horizontal NW. Consequently, decoders are only needed to address the horizontal NWs (see Section 11.8).

11.6.2
Registers and Sequential Logic

With slight modification as to how the control signals on the identified logic stages are driven, this can be turned into a clocked logic scheme. An immediate benefit is the ability to create a finite-state machine out of a single pair of PLA planes. A second benefit is the ability to use precharge logic evaluation for inverting restoration stages.

11.6.2.1 **Basic Clocking**

The basic nanoPLA cycle shown in Figure 11.14 is simply two restoring logic stages back-to-back (see Figure 11.16). For the present clocking scheme, the two stages are evaluated at altering times.

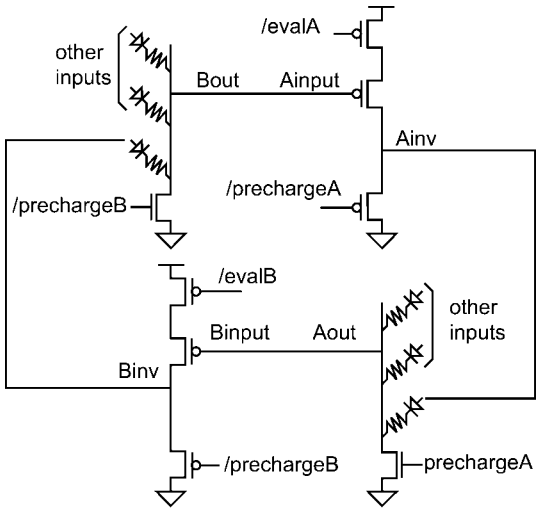


Figure 11.16 Precharge clocked INV-OR-INV-OR (NOR-NOR, AND-OR) cycle.

First, it should be noted that if all three of the control transistors in the restoring stages (restoring precharge and evaluate and diode precharge; e.g. evalA and prechargeA in Figure 11.16) are turned off, there is no current path from the input to the diode output stage. Hence, the input is effectively isolated from the output. As the output stage is capacitively loaded, the output will hold its value. As with any dynamic scheme, eventually leakage on the output will be an issue which will set a lower bound on the clock frequency.

With a stage isolated and holding its output, the following stage can be evaluated. It computes its value from its input, the output of the previous stage, and produces its result by suitably charging its output line. When this is done, this stage can be isolated and the succeeding stage (which in this simple case is also its predecessor) can be evaluated. This is the same strategy as two-phase clocking in conventional VLSI (e.g. Refs. [32, 33]).

In this manner, there is never an open current path all the way around the PLA (see Figures 11.16 and 11.17). In the two phases of operation, there is effectively a single register on any PLA outputs which feed back to PLA inputs.

11.6.2.2 **Precharge Evaluation**

For the inverting stage, the pulldown gate is driven hard during precharge and turned off during evaluation. In this manner, the line (A_{inv}) is precharged low and pulled it up only if the input (A_{input}) is low. This works conveniently in this case because the output will also be precharged low. If the input is high, then there is no need to pullup the output and it is simply left low. If the input is low, the current path is allowed to pullup the output. The net benefit is that inverter pulldown and pullup are both controlled by strongly driven gates and can be fast, whereas in a static logic scheme, the pulldown transistor must be weak, making pulldown slow compared to pullup. Typically, the weak pulldown transistor would be set to have an order of magnitude

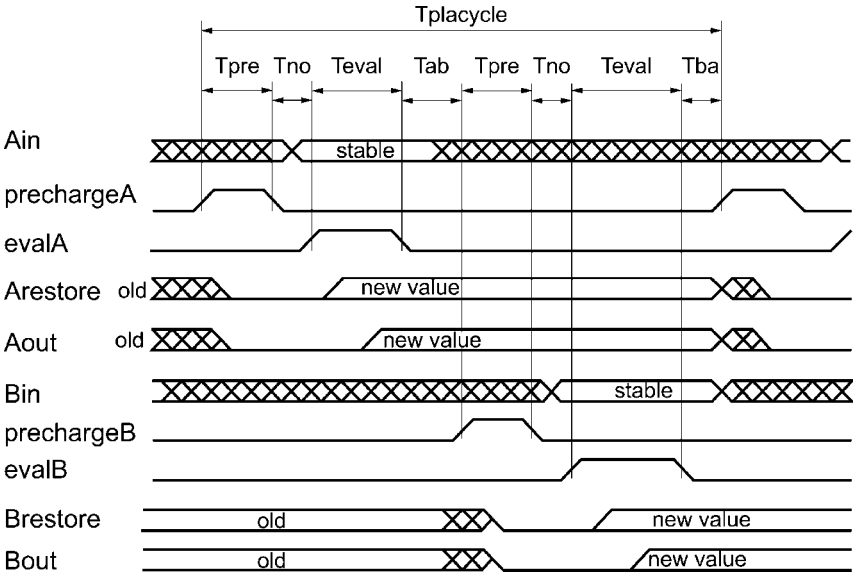


Figure 11.17 Clocking/precharge timing diagram.

higher resistance than the pullup transistor so this can be a significant reduction in worst-case gate evaluation latency.

Unfortunately, in the buffer case the weak pullup resistor can neither be precharged to high nor turned off, and so there are no comparable benefits there. It is possible that new devices or circuit organizations will eventually allow precharge buffer stages to be built.

11.6.3 Interconnect

It is known from VLSI that large PLAs do not always allow the structure which exists in logic to be exploited. For example, an *n*-input XOR requires an exponential number of product terms to construct in the two-level logic of a single PLA. Further, the limitation on NW length (see Section 11.3.2) bounds the size of the PLAs that can reasonably be built. Consequently, in order to scale up to large-capacity logic devices, modest size nanoPLA blocks must be interconnected; these nanoPLA blocks are extended to include input and output to other nanoPLA blocks and then assembled into a large array (see Figure 11.18), as first introduced by DeHon [34].

11.6.3.1 Basic Idea

The key idea for interconnecting nanoPLA blocks is to overlap the restored output NWs from each such block with the wired-OR input region of adjacent nanoPLA blocks (see Figure 11.18). In turn, this means that each nanoPLA block receives inputs from a number of different nanoPLA blocks. With multiple input sources and outputs routed in multiple directions, this allows the nanoPLA block also to serve as a

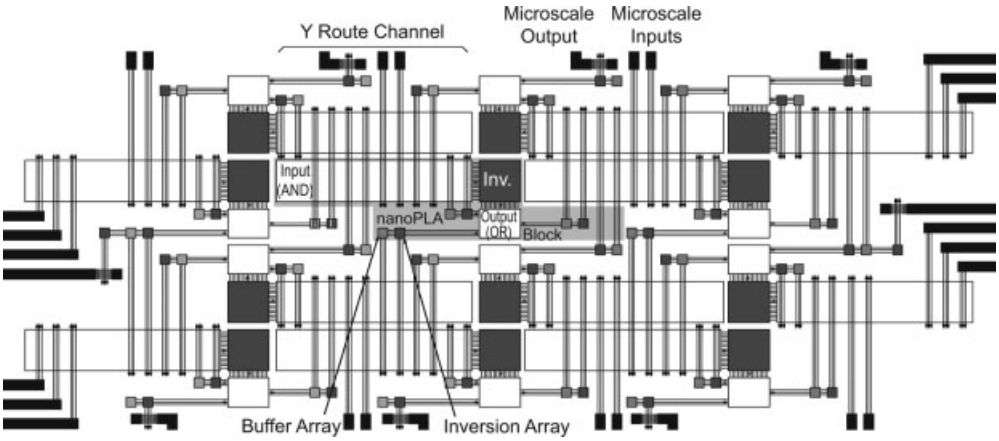


Figure 11.18 nanoPLA block tiling with edge IO to lithographic scale.

switching block. By arranging the overlap appropriately, Manhattan routing can be supported, thereby allowing the array of nanoPLA blocks to be configured to route signals between any of the blocks in the array.

11.6.3.2 **NanoPLA Block**

- *Input wired-OR region*. One or more regions of programmable crosspoints serves as the input to the nanoPLA block. Figures 11.18 and 11.19 show a nanoPLA block design with a single such input region. The inputs to this region are restored output NWs from a number of different nanoPLA blocks. The programmable crosspoints

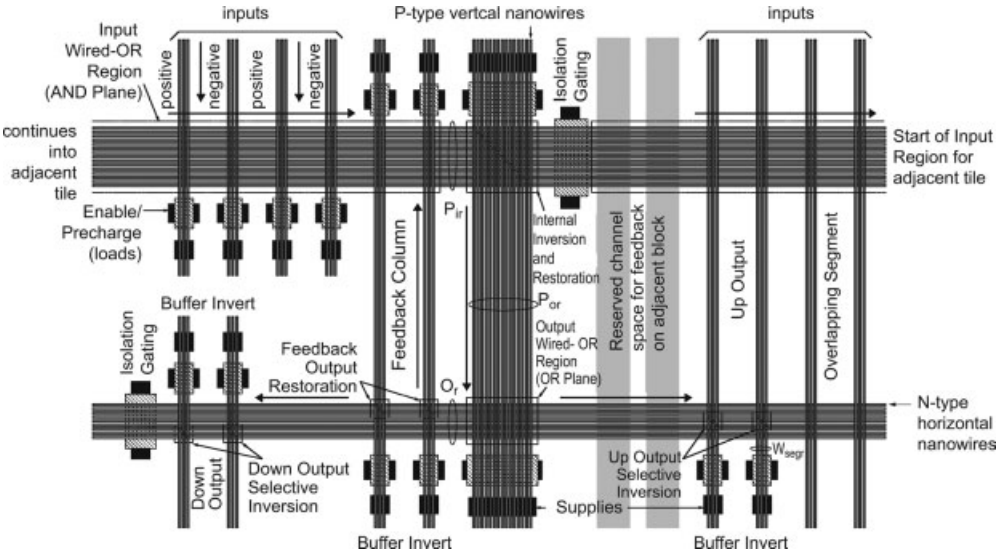


Figure 11.19 nanoPLA block tile.

allow those inputs to be selected which participate in each logical product term (PTERM) building a wired-OR array, as in the base nanoPLA (see Section 11.6.1).

- *Internal inversion and restoration array.* The NW outputs from the input block are restored by a restoration array. The restoration logic is arranged at this stage to be inverting, thus providing the logical NOR of the selected input signals into the second plane of the nanoPLA.
- *Output OR plane.* The restored outputs from the internal inversion plane become inputs to a second programmable crosspoint region. Physically, this region is the same as the input plane. Each NW in this plane computes the wired-OR of one or more of the restored PTERMs computed by the input plane.
- *Selective output inversion.* The outputs of the output OR plane are then restored in the same way as the internal restoration plane. On this output, however, the selective inversion scheme introduced in Section 11.6.1 is used. This provides both polarities of each output, and these can then be provided to the succeeding input planes. This selective inversion plays the same role as a local inverter on the inputs of conventional, VLSI PLA; here it is placed with the output to avoid introducing an additional logic plane into the design. As with the nanoPLA block, these two planes provide NOR–NOR logic. With suitable application of DeMorgan's laws, these can be viewed as a conventional AND–OR PLA.
- *Feedback.* As shown in Figures 11.18 and 11.19, one set of outputs from each nanoPLA block feeds back to its own input region. This completes a PLA cycle similar to the nanoPLA design (see Section 11.6.1). These feedback paths serve the role of intracluster routing similar to internal feedback in conventional Island-style [35] FPGAs. The nanoPLA block implements registers by routing signals around the feedback path (Section 11.6.2.1). The signals can be routed around this feedback path multiple times to form long register delay chains for data retiming.

11.6.3.3 **Interconnect**

- *Block outputs.* In addition to self feedback, output groups are placed on either side of the nanoPLA block and can be arranged so they cross input blocks of nanoPLA blocks above or below the source nanoPLA block (see Figure 11.18). Like segmented FPGAs [36, 37], output groups can run across multiple nanoPLA block inputs (i.e. Connection Boxes) in a given direction. The nanoPLA block shown in Figure 11.19 has a single output group on each side, one routing up and the other routing down. It will be seen that the design shown is sufficient to construct a minimally complete topology.
- Since the output NWs are directly the outputs of gated fields: (i) an output wire can be driven from only one source; and (ii) it can only drive in one direction. Consequently, unlike segmented FPGA wire runs, directional wires must be present that are dedicated to a single producer. If multiple control regions were coded into the NW runs, conduction would be the AND of the producers crossing the coded regions. Single direction drive arises from the fact that one side of the

gate must be the source logic signal being gated so the logical output is only available on the opposite side of the controllable region. Interestingly, the results of recent studies have suggested that conventional, VLSI-based FPGA designs also benefit from directional wires [38].

- *Y route channels.* With each nanoPLA block producing output groups which run one or more nanoPLA block heights above or below the array, the result is vertical (Y) routing channels between the logic cores of the nanoPLA blocks (see Figure 11.18). The segmented, NW output groups allow a signal to pass a number of nanoPLA blocks. For longer routes, the signal may be switched and rebuffered through a nanoPLA block (see Figure 11.20). Because of the output directionality, the result is separate sets of wires for routing up and routing down in each channel.
- *X routing.* While Y route channels are immediately obvious in Figure 11.18, the X route channels are less apparent. All X routing occurs through the nanoPLA block. As shown in Figure 11.19, one output group is placed on the opposite side of the nanoPLA block from the input. In this way, it is possible to route in the X direction by going through a logic block and configuring the signal to drive a NW in the

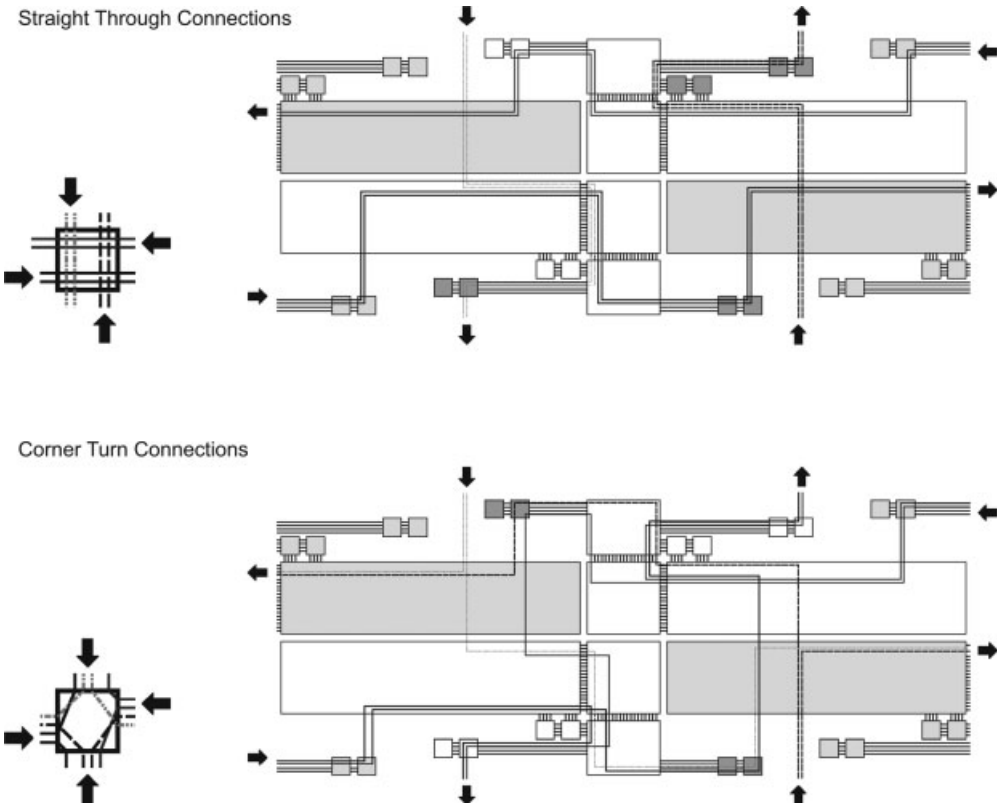


Figure 11.20 Routing view of nanoPLA logic block.

output group on the opposite side of the input. If all X routing blocks had their inputs on the left, then it would be possible only to route from left to right. To allow both left-to-right and right-to-left routing, the orientation of the inputs is alternated in alternate rows of the nanoPLA array (see Figures 11.18 and 11.20). In this manner, even rows provide left-to-right routing, while odd rows allow right-to-left routing.

- *Relation to Island-style Manhattan design.* Logically viewed, this interconnected nanoPLA block is very similar to conventional, Island-style FPGA designs, especially when the Island-style designs use directional routing [38]. As shown in Figure 11.20, there are X and Y routing channels, with switching to provide X–X, Y–Y, and X–Y routing.

11.6.4 CMOS IO

These nanoPLAs will be built on top of a lithographic substrate. The lithographic circuitry and wiring provides a reliable structure from which to probe the NWs to map their defects and to configure the logic (see Section 11.8).

For input and output to the lithographic scale during operation, IO blocks can be provided to connect the nanoscale logic to lithographic-scale wires, in much the same way that lithographic-scale wires are connected to bond pads on FPGAs. The simplest arrangement resembles the traditional, edge IO form of a symmetric FPGA with inputs and outputs attached to NWs at the edges of the routing channels (see Figure 11.18).

NW inputs can easily be driven directly by lithographic-scale wires. As the lithographic-scale wires are wider pitch, a single lithographic wire will connect to a number of NWs. With the lithographic wire connected to the NWs, the NW crosspoints in the nanoPLA block inputs can be programmed in the same way they are for NW inputs.

It is possible to connect outputs in a similar manner. Such a direct arrangement could be particularly slow, as the small NWs must drive the capacitance of a large, lithographic-scale wire. Alternately, the NWs can be used as gates on a lithographic-scale field-effect transistor (FET) (see Figure 11.21). In this manner, the NWs are only

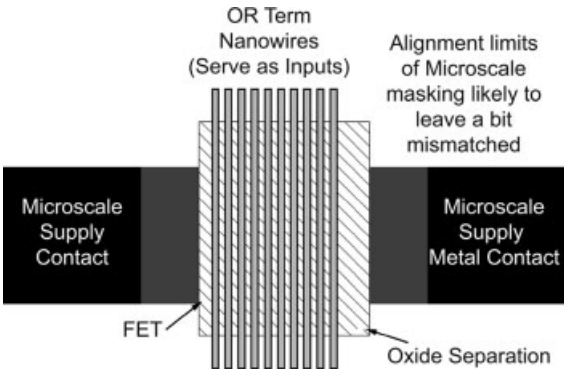


Figure 11.21 Nanoscale to lithographic-scale FET output structure.

loaded capacitively by the lithographic-scale output, and only for a short distance. The NW thresholds and lithographic FET thresholds can be tuned into comparable voltage regions so that the NWs can drive the lithographic FET at adequate voltages for switching. As shown, multiple NWs will cross the lithographic-scale gate. The OR-terms driving these outputs are all programmed identically, allowing the multiple-gate configuration to provide strong switching for the lithographic-scale FET.

11.6.5 Parameters

The key parameters in the design of the nanoPLA block are shown in Figure 11.22, where:

- W_{seg} is the number of NWs in each output group.
- L_{seg} is the number of nanoPLA block heights up or down which each output crosses; equivalently, the number of parallel wire groups across each Y route channel in each direction. In Figure 11.1 $L_{seg} = 2$, and this is maintained throughout the chapter.
- F is the number of NWs in the feedback group; for simplicity, $F = W_{seg}$ is maintained throughout the chapter.
- P is the number of logical PTERMs in the input (AND) plane of the nanoPLA logic block.
- O_p is the number of totals outputs in the OR plane. As each output is driven by a separate wired-OR NW, $O_p = 2 \times W_{seg} + F$ for the nanoPLA block focused on in this chapter, with two routing output groups and a feedback output group.
- P_p is the number of total PTERMs in the input (AND) plane. As these are also used for route-through connections, this is larger than the number of logical PTERMs in each logic block.

$$P_p \leq P + 2 \times W_{seg} + F \tag{11.2}$$

That is, in addition to the P logical PTERMs, one physical wire may be needed for each signal that routes through the array for buffering; there will be at most O_p of these.

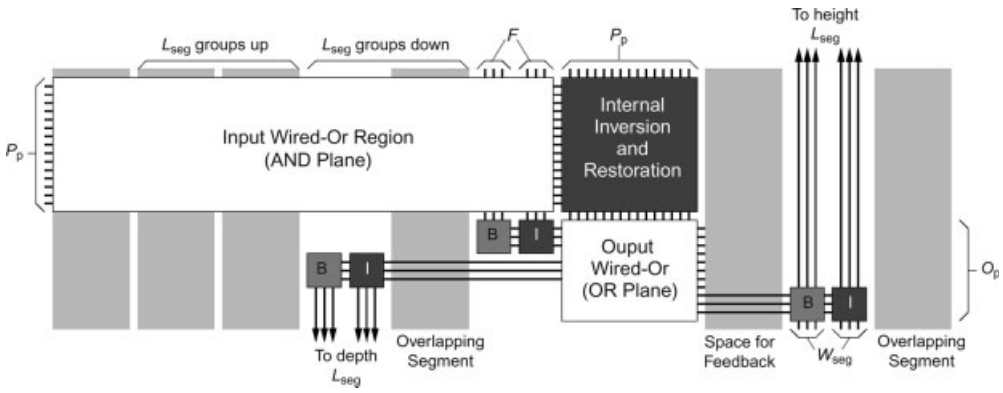


Figure 11.22 nanoPLA block parameters.

Additionally, the number and distribution of inputs [e.g. one side (as shown in Figure 11.22), from both sides, subsets of PTERMs from each side], the output topology (e.g. route both up and down on each side of the array), and segment length distributions could be parameterized. However, in this chapter attention is focused on this simple topology, with $L_{seg} = 2$. Consequently, the main physical parameters determining nanoPLA array size are W_{seg} and P_p.

11.7 Defect Tolerance

As noted in Section 11.3.3, it is likely that a small percentage of wires are defective and crosspoints are non-programmable. Furthermore, stochastic assembly (see Sections 11.4.2.2 and 11.4.3.3) and misalignment will also result in a percentage of NWs which are unusable. Fortunately, NWs are interchangeable and the crosspoints are small. Consequently, spare NWs can be provisioned into an array (e.g. overpopulate compared to the desired P_p and W_{seg}), NWs can be tested for usability (see Section 11.8.1), and the array configured using only the non-defective NWs. Further, a NW need not have a perfect set of junctions to be usable (see Section 11.7.4).

11.7.1 NW Sparing

Tolerating wire defects is a simple matter of provisioning adequate spares, separating the good wires from the bad, and configuring the nanoPLA blocks accordingly. For a given PLA design, each block should have a minimum number of usable wires (P_p and W_{seg}). As there will then be wire losses, the physical array is designed to include a larger number of physical wires to ensure that the yield of usable wires is sufficient to meet the logical requirements.

Using the restoration scheme described in Section 11.4.3, wires work in pairs. A horizontal OR-term wire provides the programmable computation or programmable interconnect, and a vertical restoration wire provides signal restoration and perhaps inversion. A defect in either wire will result in an unusable pair. Consequently, each logical OR-term or output will yield only when both wires yield. Let P_{wire} be the probability that a wire is not defective; then, the probability of yielding each OR-term is:

$$P_{OR} = (P_{input\text{-}wire} \times P_{restore\text{-}wire}). \tag{11.3}$$

An M-choose-N calculation can then be performed to determine the number of wires that must physically populate (N) to achieve a given number of functional wires (M) in the array. The probability of yielding exactly i restored OR-terms is:

$$P_{yield}(N, i) = \left(\binom{N}{i} (P_{OR})^i (1 - P_{OR})^{N-i} \right). \tag{11.4}$$

That is, there are $\binom{N}{i}$ ways to select i functional OR-terms from N total wires, and the yield probability of each case is: $(P_{OR})^i(1 - P_{OR})^{N-i}$. An ensemble is yielded with M items whenever M or more items yield, so the system yield is actually the cumulative distribution function:

$$P_{\text{M of N}} = \sum_{M \le i \le N} \left(\binom{N}{i} (P_{OR})^i (1 - P_{OR})^{N-i} \right) \tag{11.5}$$

Given the desired probability for yielding at least M functional OR-terms, $P_{\text{M of N}}$, Eq. (11.5) provides a way of finding the number of physical wires, N, that must be populated to achieve this. For the interconnected nanoPLA blocks, the product terms (P_p) and interconnect wires (W_{seg}) will be the Ms in Eq. (11.5), and a corresponding pair of raw numbers N will be calculated to determine the number of physical wires that must be placed in the fabricated nanoPLA block. Here, P_r will be used to refer to the number of raw product term NWs needed to assemble, and W_{segr} to the number of raw interconnect NWs. Figure 11.23 illustrates how much larger N needs to be than $M = 100$ for various defect rates and yield targets.

11.7.2 NW Defect Modeling

A NW could fail to be usable for several reasons:

- The NW may make poor electrical contact to microwires on either end (let P_c be the probability the NW makes a good connection on one end).
- The NW may be broken along its length (let P_j be the probability that there is no break in a NW in a segment of length L_{unit}).
- The NW may be poorly aligned with address region (wired-OR NWs) or restoration region (restoration NW) (let P_{ctrl} be the probability that a the NW is aligned adequately for use).

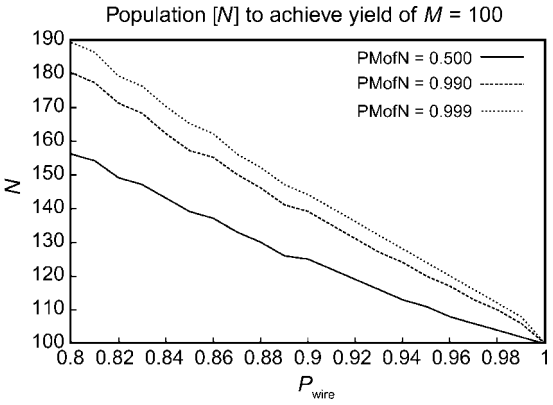


Figure 11.23 Physical population (N) of wires to achieve 100 restored OR-terms (M).

Consequently, the base NW yield looks like:

$$P_{\text{wire}} = (P_{\text{C}})^2 \times (P_{\text{j}})^{L_{\text{wire}}/L_{\text{unit}}} \times P_{\text{ctrl}} \tag{11.6}$$

Typically, $P_{\text{c}} = 0.95$ (after Ref. [5] and $P_{\text{j}} = 0.9999$ with $L_{\text{unit}} = 10\,\text{nm}$ (after Ref. [19]; see also Refs. [27, 34]). P_{ctrl} can be calculated from the geometry of the doped regions [24]. P_{wire} is typically about 0.8.

11.7.3
Net NW Yield Calculation

A detailed calculation for NW population includes both wire defect effects and stochastic population effects. Starting with a raw population number for the NWs in each piece of the array, it is possible to:

- calculate the number of non-defective wired-OR wires within the confidence bound [Eqs. (11.6) and (11.5)];
- calculate the number of those which can be uniquely addressed using the following recurrence:

$$\begin{aligned} P_{\text{different}}(T, N, u) = &\left(\frac{T-(u-1)}{T}\right) \times P_{\text{different}}(T, N-1, u-1) \\ &+ \left(\frac{u}{T}\right) \times P_{\text{different}}(T, N-1, u) \end{aligned} \tag{11.7}$$

where T is the number of different wire types (i.e. the size of the address space), N is the raw number of nanowires populated in the array, and u is the number of unique NWs in the array.

- calculate the number of net non-defective restored wire pairs within the confidence bound [Eqs. (11.3), (11.6), and (11.5)];
- calculate the number of uniquely restored OR terms using Eq. (11.7); in this case, T is the number of possible restoration wires rather than the number of different NW addresses.

These calculations indicate how to obtain P_{r} and W_{segr} to achieve a target P_{p} and W_{seg}.

11.7.4
Tolerating Non-Programmable Crosspoints

As will be seen in Table 11.1, PLA crosspoint arrays are typically built with approximately 100 net junctions. If were demanded that all 100 crosspoint junctions on a NW were programmable in order for the NW to yield, then an unreasonably high yield rate per crosspoint would be required. That is, assuming a crosspoint is programmable with probability P_{pgm} and a NW has N_{junc} input NWs – and hence crosspoint junctions

Table 11.1 Area minimizing nanoPLA design points (Ideal Restoration, with $W_{litho} = 105$ nm, $W_{fnano} = W_{dnano} = 10$ nm); area ratios estimate how much larger 22 nm lithographic FPGAs would be compared to the mapped nanoPLA designs.

Design	P_p	W_{seg}	Area ratio
alu4	60	8	340
apex2	54	15	39
apex4	62	7	210
bigkey	44	13	69
clma	104	28	30
des	78	25	26
diffeq	86	21	32
deip	58	18	59
elliptic	78	27	27
ex1010	66	9	290
ex5p	67	18	390
frisc	92	34	17
misex3	64	8	150
pdc	74	13	360
s298	79	15	110
s38417	76	22	32
seq	72	18	69
spla	68	12	630
tseng	78	25	20

– then the probability that all junctions on a NW are programmable is

$$P_{\text{pgmwire}} = (P_{\text{pgm}})^{N_{\text{junc}}} \tag{11.8}$$

To have $P_{\text{pgmwire}} \geq 0.5$, P_{pgm} would need to be >0.993. However, as was noted in Section 11.3.3, the non-programmable crosspoint defect rates would be expected to be in the range of 1 to 10% ($0.9 \leq P_{\text{pgm}} \leq 0.99$).

As introduced by Naeimi and DeHon [39], it is apparent that a NW with non-programmable crosspoints can still be used to implement a particular OR-term, as long as it has programmable crosspoints where the OR-term needs on-programmed junctions. Furthermore, as the array has a large number of otherwise interchangeable NWs (e.g. 100), it is possible to search through the array for NWs that can implement each particular OR-term.

For example, if a logic array (AND or OR plane) of a nanoPLA has defective junctions (as marked in Figure 11.24), the OR-term $f = A + B + C + E$ can be assigned to NW $W3$, despite the fact that it has a defective (non-programmable) junction at ($W3$, D); that is, the OR-term f is compatible with the defect pattern of NW $W3$.

As the number of programmed junctions needed for a given OR-term is usually small (e.g. 8–20) compared to the number of inputs in an array (e.g. 100), the probability that a NW can support a given OR-term is much larger than the probability that it has no junction defects. Assuming that C is the fan-in to the OR-term, and assuming random

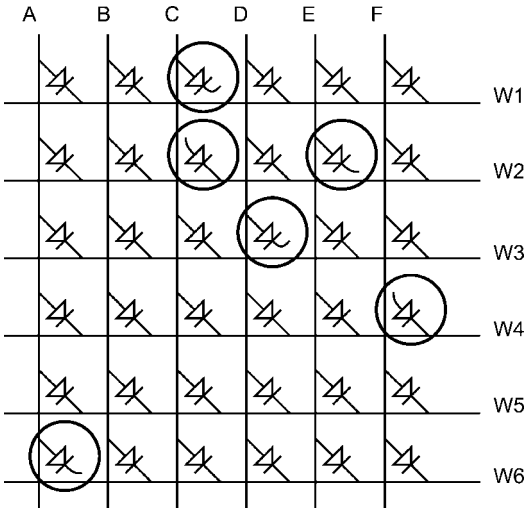


Figure 11.24 OR array with defective junctions.

junction defects, the probability that the NW can support the OR-term is

$$P_{\text{support}}(C) = (P_{\text{pgm}})^{C}. \tag{11.9}$$

For example, in a 100 NW array, if $P_{\text{pgm}} = 0.95$, $P_{\text{support}}(13) \approx 0.51$, and $P_{\text{pgmwire}} \approx 0.006$. Furthermore, as multiple NWs can be used in an array to find a compatible match, failure to map a NW will only occur if there are no compatible NWs in the array.

$$P_{\text{match}} = (C,\ N_{\text{wire}}) = (1 - (1 - P_{\text{support}}(C)^{N_{\text{wire}}}). \tag{11.10}$$

Hence, the probability of failing to find a match for the $C = 13$ OR-term in a 100 NW array is $[1 - P_{\text{match}}(13, 100) \leq 10^{-31}]$. Alternately, this means we have a 99% chance of finding a match after checking only 8 NWs ($P_{\text{match}}(13, 8) > 0.99$).

Naeimi and DeHon [39] developed the analysis and mapping strategy in greater detail for tolerating non-programmable crosspoints. DeHon and Naeimi [30] further expanded the mapping strategy to the interconnected nanoPLAs described in Section 11.6.3, and showed that non-programmable defect rates of up to 5% could be accommodated, with no additional overhead.

11.8 Bootstrap Testing

11.8.1 Discovery

Since addressing and restoration is stochastic, there is a need to discover the live addresses and their restoration polarity. Further, as the NWs will be defective it is vital

to identify those NWs which are usable and those which are not. Here, the restoration columns (see Figures 11.14 and 11.19) are used to help identify useful addresses. The gate side supply (e.g. the top set of lithographic wire contacts in Figure 11.10) can be driven to a high value, after which a voltage is sought on the opposite supply line (e.g. the bottom set of lithographic wire contacts in Figure 11.10; these contacts are marked V_{high} and Gnd in Figure 11.10, but will be controlled independently as described here during discovery). There will be current flow into the bottom supply only if the control associated with the p-type restoration wire can be driven to a sufficiently low voltage. The process is started by driving all the row lines high, using the row precharge path. A test address is then applied and the supply (V_{row} in Figure 11.14) is driven low. If a NW with the test address is present, only that line will now be strongly pulled low. If the associated row line can control one or more wires in the restoration plane, the selected wires will now see a low voltage on their field-effect control regions and enable conduction from the top supply to the bottom supply. By sensing the voltage change on the bottom supply, the presence of a restored address can be deduced. Broken NWs will not be able to effect the bottom supply. NWs with excessively high resistance due to doping variations or poor contacts will not be able to pull the bottom supply contact up quickly enough. As the buffering and inverting column supplies are sensed separately it will be known whether the line is buffering, inverting, or binate.

No more than $O((P_p)^2)$ unique addresses are needed to achieve virtually unique row addressing [24], so the search will require at most $O((P_p)^2)$ such probes. A typical address width for the nanoPLA blocks is $N_a = 14$, which provides 3432 distinct 7-hot codes, and a typical number of OR-terms might be 90 (see Table 11.1). Hence, 3432 addresses may need to be probed to find 90 live row wires.

When all the present addresses in an array and the restoration status associated with each address are known, logic can be assigned to logical addresses within each plane, based on the required restoration for the output. With logic assigned to live addresses in each row, the address of the producing and consuming row wires can now be used to select and program a single junction in a diode-programmable OR plane.

11.8.2 Programming

In order to program any diode crosspoint in the OR planes (e.g. Figure 11.14), one address is driven into the top address decoder, and the second address into the bottom. The stochastic restoration performs the corner turn, so that the desired programming voltage differential is effectively placed across a single crosspoint. The voltages and control gating on the restoration columns are then set to define which programmable diode array is actually programmed during a programming operation. For example, in Figure 11.14 the ohmic supply contacts at the top and bottom are the control voltages; the signals used for control gating are labeled with precharge and eval signal names. To illustrate the discovery and programming process, DeHon [29] presents the steps involved in discovering and programming an exemplary PLA.

11.8.3 Scaling

It should be noted that each nanoPLA array is addressed separately from its set of microscale wires ($A0$, $A1$, ... and V_{row}, V_{bot}, and V_{top}; see Figure 11.14). Consequently, the programming task is localized to each nanoPLA plane, and the work required to program a collection of planes (e.g. Figure 11.18) only scales linearly with the number of planes.

11.9 Area, Delay, and Energy

11.9.1 Area

From Figures 11.19 and 11.22 the basic area composition of each tile can be seen. For this, the following feature size parameters are used:

- W_{litho} is the lithographic interconnect pitch; for example, for the 45-nm node, $W_{litho} = 105$ nm [40].
- W_{dnano} is the NW pitch for NWs which are inputs to diodes (i.e. Y route channel segments and restored PTERM outputs).
- W_{fnano} is the NW pitch for NWs which are inputs to field-effect gated NWs; this may be larger than W_{dnano} in order to prevent inputs from activating adjacent gates and to avoid short-channel FET limitations.

The tile area is computed by first determining the tile width, TW, and tile height, TH:

$$TW = (3 + 4(L_{seg} + 1)) \times W_{litho} + (P_{or} + 4(L_{seg} + 1)W_{segr}) \times W_{dnano} \quad (11.11)$$

$$TH = 12 \times W_{litho} + (O_r + P_{ir}) \times W_{fnano} \quad (11.12)$$

$$AW = (N_a + 2) \times W_{litho} \quad (11.13)$$

$$Area = (AW + TW) \times TH \quad (11.14)$$

where P_{or}, P_{ir}, O_r, and W_{segr} (shown in Figure 11.19) are the raw number of wires needed to populate in the array in order to yield P_p restored inputs, O_p restored outputs, and W_{seg} routing channels (see Section 11.7.3). The two 4s in TW arise from the fact that there are $L_{seg} + 1$ wire groups on each side of the array ($2\times$), and each of those is composed of a buffer/inverter selective inversion pair ($2\times$). A lithographic spacing is charged for each of these groups as they must be etched for isolation and controlled independently by lithographic scale wires. The 12 lithographic pitches in TH account for the three lithographic pitches needed on each side of a group of wires for the restoration supply and enable gating. As segmented wire runs end and begin

between the input and output horizontal wire runs, these three lithographic pitches are paid for four-fold in the height of a single nanoPLA block: once at the bottom top of the block (see Figure 11.19).

N_a is the number of microscale address wires needed to address individual, horizontal nanoscale wires [24]; for the nanoPLA blocks in these studies, N_a is typically 14 to 20. Two extra wire pitches in the AddressWidth (*AW*) are the two power supply contacts at either end of an address run.

11.9.2 Delay

Figures 11.16 and 11.17 show the basic nanoPLA clock cycle, $T_{placycle}$. The component delays shown in Figure 11.17 (e.g. nanowire precharge and evaluation times) are calculated based on the NW resistances and capacitances, the crosspoint resistances, and the nanowire FET resistances [29]. NW resistance and capacitance can be calculated based on geometry and material properties using the NW lengths, which are roughly multiples of the tile width, *TW*, and tile height, *TH*, identified in the previous section. If simply heavily doped silicon nanowires are used, the NW resistances can be close to 10 MΩ, and this results in nanoPLA clock cycle times in the tens of nanoseconds. However, if the regions of the NW which do not need to be semiconducting are converted selectively – that is, everything except the diode crosspoint region and the field-effect restoration region – into a nickel silicide (NiSi) [12], the NW resistances can be reduced to the 1 MΩ range. As a result, the nanoPLA clock cycle is brought down to the nanosecond region. This selective conversion can be performed as a lithographic-scale masking step and, with careful engineering, subnanosecond nanoPLA cycle times may be possible. As long as the NW resistance is in the 1 MΩ range, it will dominate the on-resistance of both the field-effect gating in the restoration NW (R_{onfet}) and diode on-resistances ($R_{ondiode}$) in the 100 KΩ range.

11.9.3 Energy and Power

The nanoPLAs will dissipate active energy, charging and discharging the functional and configured NWs.

$$E_{NW} = \frac{1}{2} C_{Wire} V^2. \tag{11.15}$$

As noted in the previous section, C_{wire} can be computed from the material properties and geometry. To tolerate variations in NW doping, it is likely the operating voltage will need to be 0.5 to 1 V.

The raw E_{NW} can be discounted by the fraction of NWs typically used in a routing channel or a logic array, *F*. This tends to be 70–80% with the current tools and designs. When using the selective inversion scheme, both polarities of most signals will typically be driven to guarantee a close to 50% activity factor, *A*.

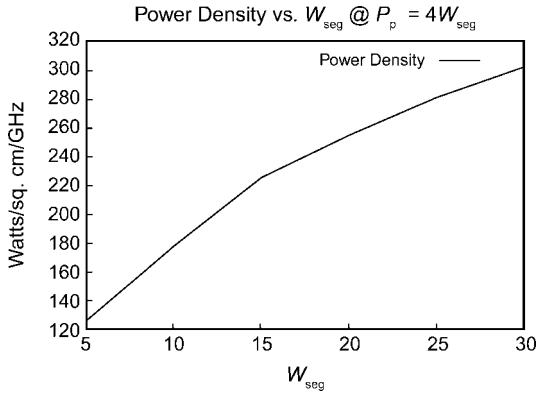


Figure 11.25 Power density as a function of W_{seg} for ideal restore, stochastic address case with $W_{litho} = 105$ nm, $W_{fnano} = 10$ nm, $W_{dnano} = 10$ nm.

Assuming an operating frequency of f, the power for a nanoPLA tile is

$$P_{array} = \sum_{\text{all NWs}} (A \times F \times E_{NW} \times f). \tag{11.16}$$

The power density is then

$$P_{density} = \frac{P_{array}}{Area}. \tag{11.17}$$

Here, *Area is* the area for the tile as calculated in Eq. (11.14).

Figure 11.25 shows the power density associated with interconnected nanoPLAs, and suggests that the designs may dissipate a few hundred Watts per cm^2. In typical designs, compute arrays would be interleaved with memory banks (see Section 11.5), which have much lower power densities. Nonetheless, this suggests that power management is as much an issue in these designs as it is in traditional, lithographic, designs.

11.10 Net Area Density

Recent developments in technology suggest that it is possible to build and assemble 10 nm-pitch NWs with crosspoints at every NW–NW crossing. To use these, it is necessary to pay for lithographic addressing overhead, to use regular architectures, and tolerate defects. In order to understand the net benefits, the characteristics of composite designs are analyzed. As an example, conventional FPGA benchmarks are mapped from the Toronto 20 benchmark suite [1] to NW logic with $W_{litho} = 105$ nm (45 nm roadmap node) and $W_{fnano} = W_{dnano} = 10$ nm. This provides a count of nanoPLA blocks and the logical P_p and W_{seg} parameters identified in Section 11.6.5, and these calculations can then be used for yield and statistical assembly

(see Section 11.7) to compute physical nanowire population, and the area equations in Section 11.9.1 to compute composite area. Subsequently, the resultant minimum area obtainable is compared with the nanoPLA designs to lithographic 4-LUT FPGAs at the 22 nm node [40]. As shown in Table 11.1, and further detailed in Ref. [29], the routed nanoPLA designs are one to two orders of magnitude smaller than 22 nm lithographic FPGAs, even after accounting for lithographic addressing overhead, defects, and statistical addressing.

The datapoints in Table 11.1 are based on a number of assumptions about lithographic and nanowire pitches and statistical assembly. DeHon [29] also examined the sensitivity to these various parameters, and showed that the statistical restoration assembly costs a factor of three in density for large arrays, while the cost of statistical addressing is negligible. If the diode pitch (W_{dnano}) could be reduced to 5 nm, another factor of almost two in area could be saved. Moreover, if the lithographic support were also reduced to the 22 nm node ($W_{\text{litho}} = 45$ nm), a further three-fold factor in density advantage would be gained compared to the data in Table 11.1.

11.11 Alternate Approaches

During recent years, several groups have been studying variants of these nanowire-based architectures (see Table 11.2). Heath *et al.* [41] articulated the first vision for constructing defect-tolerant architectures based on molecular switching and bottom-up construction. Luo *et al.* [42] elaborated the molecular details and diode-logic structure, while Williams and Kuekes [23] introduced a random particle decoder scheme for addressing individual NWs from lithographic-scale wires. These early designs assumed that diode logic was restored and inverted using lithographic scale CMOS buffers and inverters.

Goldstein and Budiu [43] described an interconnected set of these chemically-assembled diode-based devices, while Goldstein and Rosewater [44] used only two-terminal non-restoring devices in the array, but added latches based on resonant-tunneling diodes (RTDs) for clocking and restoration. Snider *et al.* [45] suggested nanoFET-based logic and also tolerated non-programmable crosspoint defects by matching logic to the programmability of the device.

Strukov and Likharev [46] also explored crosspoint-programmable nanowire-based programmable logic and used lithographic-scale buffers with an angled topology and nanovias so that each long NW could be directly attached to a CMOS-scale buffer. Later, Snider and Williams [47] built on the Strukov and Likharev interfacing concept and introduced a more modest design which used NWs and molecular-scale switches only for interconnect, performing all logic in CMOS.

These designs all share many high-level goals and strategies, as have been described in this chapter. They suggest a variety of solutions to the individual technical components including the crosspoint technologies, NW formation, lithographic-scale interfacing, and restoration (see Table 11.2). The wealth of technologies

Table 11.2 Comparison of NW-based logic designs.

Design source	Component					
	Crosspoint technology	**NW**	**Logic**	**Litho ↔ NW**	**Restoration**	**Reference(s)**
HP/UCLA	Molecular switch diode	Imprint lithography	Nanoscale wired-OR	Random particles	CMOS	22, 41
CMU nanoFabric	Molecular switch diode	NanoPore templates	Nanoscale wired-OR	–	RTD latch	43, 44
SUNY CMOL	Single-electron transistor	Interferometric lithography	Nanoscale wired-OR	Offset angles	CMOS	46
HP FPNI	Molecular switch diode	Imprint lithography	CMOS (N)AND	Offset angles	CMOS	47
This chapter	Switchable diode	Catalyst NWs	Nanoscale wired-OR	Coded NWs	NW FET	–

and construction alternatives identified by these and other research groups has increased the general confidence that there are options to bypass any of the challenges that might arise when realizing any single technique or feature in these designs.

11.12 Research Issues

While the key building blocks have been demonstrated as previously cited, considerable research and development remains in device synthesis, assembly, integration, and process development. At present, no complete fundamental understanding of the device physics at these scales is available, and a detailed and broader characterization of the devices, junctions, interconnects, and assemblies is necessary to refine the models, to better predict the system properties, and to drive architectural designs and optimization.

The mapping results outlined in Section 11.10 were both area- and defect-tolerance driven. For high-performance designs, additional techniques, design transformations, and optimizations will be needed, including interconnect pipelining (e.g. Ref. [48]) and fan-out management (e.g. Ref. [49]).

In Section 11.7 it was noted that high defect rates could be tolerated when the defects occurred before operation. However, new defects are likely to arise during operation, and additional techniques and mechanisms will be necessary to detect their occurrence, to guard the computation against corruption when they do occur, and rapidly to reconfigure around the new defects.

Further, it is expected that these small feature devices will encounter transient faults during operation. Although the exact fault rates are at present unknown, they are certainly expected to exceed those rates traditionally seen in lithographic silicon. This suggests the need for new lightweight techniques and architectures for fault identification and correction.

11.13 Conclusions

Bottom-up synthesis techniques can be used to produce single nanometer-scale feature sizes. By using decorated NWs – for example, by varying composition at the nanometer scale, both axially and radially – the key nanoscale features may be built into the NWs. Moreover, the NWs can be assembled at tight, nanoscale pitch into dense arrays, contacted to a reliable, lithographic-scale infrastructure, and individually addressed from the lithographic scale. The aggregate set of synthesis and assembly techniques appears adequate for the building of arbitrary logic at the nanoscale, even if the only programmable elements are non-restoring diodes.

Bottom-up self-assembly demands that highly regular structures are built that can be differentiated stochastically for addressing and restoration. NW field-effect gating provides signal restoration and inversion while keeping signals at the dense,

nanoscale pitch. Post-fabrication configuration allows the definition of deterministic computation on top of the regular array, despite random differentiation and high rates of randomly placed defects. When these NWs are assembled into modest-sized interconnected PLA arrays, it is estimated that the net density would be one-to-two orders of magnitude higher than for defect-free lithographic-scale FPGAs built in 22 nm CMOS. This should provide a pathway by which to exploit nanometer-pitch devices, interconnect, and systems without pushing lithography into providing these smallest feature sizes.

Acknowledgments

The architectural studies into devices and construction techniques which emerge from scientific research do so only after close and meaningful with the physical scientists. These studies have been enabled by collaboration with Charles M. Lieber and his students. The suite of solutions summarized here includes joint investigations with Helia Naeimi, Michael Wilson, John E. Savage, and Patrick Lincoln.

These research investigations were funded in part by National Science Foundation Grant CCF-0403674 and the Defense Advanced Research Projects Agency under ONR contracts N00014-01-0651 and N00014-04-1-0591.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors, and do not necessarily reflect the views of the National Science Foundation or the Office of Naval Research.

Christian Nauenheim and Rainer Waser helped to produce this brief chapter as a digested version of Ref. [29].

References

1 V. Betz, J. Rose, FPGA place-and-route challenge, 1999. Available at http://www.eecg.toronto.edu/~vaughn/challenge/challenge.html.

2 A. M. Morales, C. M. Lieber, A laser ablation method for synthesis of crystalline semiconductor nanowires. *Science* 1998, **279**, 208–211.

3 (a) Y. Cui, L. J. Lauhon, M. S. Gudiksen, J. Wang, C. M. Lieber, Diameter-controlled synthesis of single crystal silicon nanowires. *Appl. Phys. Lett.* 2001, **78** (15), 2214–2216. (b) Y. Cui, Z. Zhong, D. Wang, W. U. Wang, C. M. Lieber, High performance silicon nanowire field effect transistors. *Nano Lett.* 2003, **3** (2), 149–152.

4 Y. Cui, X. Duan, J. Hu, C. M. Lieber, Doping and electrical transport in silicon nanowires. *J. Phys. Chem. B* 2000, **104** (22), 5213–5216.

5 Y. Huang, X. Duan, Y. Cui, L. Lauhon, K. Kim, C. M. Lieber, Logic gates and computation from assembled nanowire building blocks. *Science* 2001, **294**, 1313–1317.

6 M. S. Gudiksen, L. J. Lauhon, J. Wang, D. C. Smith, C. M. Lieber, Growth of nanowire superlattice structures for nanoscale photonics and electronics. *Nature* 2002, **415**, 617–620.

7 Y. Wu, R. Fan, P. Yang, Block-by-block growth of single-crystalline Si/SiGe

superlattice nanowires. *Nano Lett.* 2002, **2** (2), 83–86.

8 M. T. Björk, B. J. Ohlsson, T. Sass, A. I. Persson, C. Thelander, M. H. Magnusson, K. Depper, L. R. Wallenberg, L. Samuelson, One-dimensional steeplechase for electrons realized, *Nano Lett.* 2002, **2** (2), 87–89.

9 L. J. Lauhon, M. S. Gudiksen, D. Wang, C. M. Lieber, Epitaxial core-shell and core-multi-shell nanowire heterostructures. *Nature* 2002, **420**, 57–61.

10 M. Law, J. Goldberger, P. Yang, Semiconductor nanowires and nanotubes. *Annu. Rev. Mater. Sci.* 2004, **34**, 83–122.

11 D. Whang, S. Jin, C. M. Lieber, Nanolithography using hierarchically assembled nanowire masks. *Nano Lett.* 2003, **3** (7), 951–954.

12 Y. Wu, J. Xiang, C. Yang, W. Lu, C. M. Lieber, Single-crystal metallic nanowires and metal/semiconductor nanowire heterostructures. *Nature* 2004, **430**, 61–64.

13 Y. Huang, X. Duan, Q. Wei, C. M. Lieber, Directed assembly of one-dimensional nanostructures into functional networks. *Science* 2001, **291**, 630–633.

14 (a) D. Chen, J. Cong, M. Ercegovac, Z. Huang, Performance-driven mapping for cpld architectures. *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.* 2003, **22** (10), 1424–1431. (b) Y. Chen, G.-Y. Jung, D. A. A Ohlberg, X. Li, D. R. Stewart, J. O. Jeppesen, K. A. Nielsen, J. F. Stoddart, R. S. Williams, Nanoscale molecular-switch crossbar circuits. *Nanotechnology* 2003, **14**, 462–468.

15 D. R. Stewart, D. A. A. Ohlberg, P. A. Beck, Y. Chen, R. S. Williams, J. O. Jeppesen, K. A. Nielsen, J. F. Stoddart, Molecule-independent electrical switching in pt/organic monolayer/ti devices. *Nano Lett.* 2004, **4** (1), 133–136.

16 A. DeHon, Reconfigurable architectures for general-purpose computing. AI Technical report 1586 (oct.), MIT Artificial Intelligence Laboratory, Cambridge, MA, 1996.

17 Y. Wu, P. Yang, Germanium nanowire growth via simple vapor transport. *Chem. Mater.* 2000, **12**, 605–607.

18 B. Zheng, Y. Wu, P. Yang, J. Liu, Synthesis of ultra-long and highly-oriented silicon oxide nanowires from alloy liquid. *Adv. Mater.* 2002, **14**, 122.

19 M. S. Gudiksen, J. Wang, C. M. Lieber, Synthetic control of the diameter and length of semiconductor nanowires. *J. Phys. Chem. B* 2001, **105**, 4062–4064.

20 D. Whang, S. Jin, Y. Wu, C. M. Lieber, Large-scale hierarchical organization of nanowire arrays for integrated nanosystems. *Nano Lett.* 2003, **3** (9), 1255–1259.

21 Y. Chen, D. A. A Ohlberg, X. Li, D. R. Stewart, R. S. Williams, J. O. Jeppesen, K. A. Nielsen, J. F. Stoddart, D. L. Olynick, E. Anderson, Nanoscale molecular-switch devices fabricated by imprint lithography. *Appl. Phys. Lett.* 2003, **82,** 10, 1610–1612.

22 W. Wu, G.-Y. Jung, D. Olynick, J. Straznicky, Z. Li, X. Li, D. Ohlberg, Y. Chen, S.-Y. Wang, J. Liddle, W. Tong, R. S. Williams, One-kilobit cross-bar molecular memory circuits at 30-nm half-pitch fabricated by nanoimprint lithography. *Appl. Physics A* 2005, **80**, 1173–1178.

23 S. Williams, P. Kuekes,Demultiplexer for a molecular wire crossbar network. United States Patent Number 6,256,767, 2001.

24 A. DeHon, P. Lincoln, J. Savage, Stochastic assembly of sublithographic nanoscale interfaces. *IEEE Trans. Nanotech.* 2003, **2** (3), 165–174.

25 B. Gojman, E. Rachlin, J. E. Savage, Decoding of stochastically assembled nanoarrays, in: *Proceedings of the International Symposium on VLSI*, Lafayette, USA, IEEE Computer Society, 2004.

26 A. DeHon, Law of large numbers system design, in: S. K. Shukla, R. I. Bahar (Eds.), *Nano, Quantum and Molecular Computing: Implications to High Level Design and Validation*, Kluwer Academic Publishers, Boston, MA, Chapter 7, pp. 213–241, 2004.

27 A. DeHon, Array-based architecture for FET-based, *nanoscale electronics*. *IEEE Trans. Nanotech.* 2003, **2** (1), 23–32.

28 F. G. Maunsell, A problem in cartophily. *The Math. Gazette* 1937, **22**, 328–331.

29 A. DeHon, Nanowire-based programmable architecture. *ACM J. Emerging Technol. Comput. Systems* 2005, **1** (2), 109–162

30 A. DeHon, H. Naeimi, Seven strategies for tolerating highly defective fabrication. *IEEE Design Test Comput.* 2005, **22** (4), 306–315.

31 A. DeHon, M. J. Wilson, Nanowire-based sublithographic programmable logic arrays, in: *Proceedings International Symposium on Field-Programmable Gate Arrays*, Napa Valley, CA, IEEE Publishers, pp. 123–132, 2004.

32 C. Mead, L. Conway, *Introduction to VLSI Systems*, Addison-Wesley, 1980.

33 N. H. E Weste, D. Harris, *CMOS VLSI Design: A Circuits and Systems Perspective*, 3rd edn., Addison-Wesley, 2005.

34 A. DeHon, Design of programmable interconnect for sublithographic programmable logic arrays, in: *Proceedings International Symposium on Field-Programmable Gate Arrays*, Monterey, CA, ACM Publishers, pp. 127–137, 2005.

35 V. Betz, J. Rose, A. Marquardt, *Architecture and CAD for Deep-Submicron FPGAs*, Kluwer Academic Publishers, Norwell, MA, 1999.

36 S. Brown, M. Khellah, Z. Vranesic, Minimizing FPGA interconnect delays. *IEEE Des. Test Comput.* 1996, **13** (4), 16–23.

37 V. Betz, J. Rose, FPGA routing architecture: Segmentation and buffering to optimize speed and density, in: *Proceedings International Symposium on Field-Programmable Gate Arrays*, Monterey, CA, ACM Publishers, pp. 59–68, 1999.

38 G. Lemieux, E. Lee, M. Tom, A. Yu, Directional and single-driver wires in FPGA interconnect, in: *Proceedings International Conference on Field-Programmable Technology*, Brisbane, Australia, IEEE Publishers, pp. 41–48, 2004.

39 H. Naeimi, A. DeHon, A greedy algorithm for tolerating defective crosspoints in NanoPLA design, in: *Proceedings IEEE International Conference on Field-Programmable Technology*, Brisbane, Australia, IEEE Publishers, pp. 49–56, 2004.

40 ITRS, International technology roadmap for semiconductors. http://public.itrs.net/Files/2001ITRS/, 2001.

41 J. R. Heath, P. J. Kuekes, G. S. Snider, R. S. Williams, A defect-tolerant computer architecture: Opportunities for nanotechnology. *Science* 1998, **280** (5370), 1716–1721.

42 Y. Luo, P. Collier, J. O. Jeppesen, K. A. Nielsen, E. Delonno, G. Ho, J. Perkins, H.-R. Tseng, T. Yamamoto, J. F. Stoddart, J. R. Heath, Two-dimensional molecular electronics circuits. *ChemPhysChem* 2002, **3** (6), 519–525.

43 S. C. Goldstein, M. Budiu, NanoFabrics: Spatial computing using molecular electronics, in: *Proceedings International Symposium on Computer Architecture*, Gothenburg, Sweden, ACM Publishers, pp. 178–189, 2001.

44 S. C. Goldstein, D. Rosewater, Digital logic using molecular electronics. *IEEE ISSCC Digest Tech. Papers* 2002, 204–205

45 G. Snider, P. Kuekes, R. S. Williams, CMOS-like logic in defective, nanoscale crossbars. *Nanotechnology* 2004, **15**, 881–891.

46 D. B. Strukov, K. K. Likharev, CMOL FPGA: a reconfigurable architecture for hybrid digital circuits with two-terminal nanodevices. *Nanotechnology* 2005, **16** (6), 888–900.

47 G. Snider, R. S. Williams, Nano/CMOS architectures using a field-programmable nanowire interconnect. *Nanotechnology* 2007, **18** (3).

48 W. Tsu, K. Macy, A. Joshi, R. Huang, N. Walker, T. Tung, O. Rowhani, V. George, J. Wawrzynek, A. DeHon, HSRA: High-

speed, hierarchical synchronous reconfigurable array, in: *Proceedings International Symposium on Field-Programmable Gate Arrays*, Monterey, CA, ACM Publishers, pp. 125–134, 1999.

49 H. J. Hoover, M. M. Klawe, N. J. Pippenger, Bounding fan-out in logical networks. *J. Assoc. Comput. Machinery* 1984, **31** (1), 13–18.

12
Quantum Cellular Automata

Massimo Macucci

12.1
Introduction

The concept of quantum cellular automata (QCA) was first proposed by Craig Lent and coworkers [1] at the University of Notre Dame in 1993, as an alternative based on bistable electrostatically coupled cells to traditional architectures for computation. Overall, the QCA architecture probably represents the proposal for an alternative computing paradigm that has been developed furthest, up to the experimental proof of principle [2]. As will be discussed in the following sections, its strengths are represented by the reduced complexity (in particular for the implementation based on ground-state relaxation), extremely low power consumption, and potential for ultimate miniaturization; its drawbacks are the extreme sensitivity to fabrication tolerances and stray charges, the difficulty in achieving operating temperatures reasonably close to room temperature, the undesired interaction among electrodes operating on different cells, and the very challenging control of dot occupancy.

The initial formulation of the QCA architecture relied on the relaxation to the ground state of an array of cells: computation was performed enforcing the polarization state of a number of "input" cells and then reading the state of a number of "output" cells, once the array had relaxed down to the ground state consistent with the input data. Such an approach is characterized (as will be discussed in the following) by a simple – at least in principle – layout, but suffers from the presence of many states very close in energy to the actual ground state. This leads to an extremely slow and stochastic relaxation, which may lead to unacceptable computational times.

The slow and unreliable evolution of the ground-state relaxation approach was addressed with the introduction of a modified QCA architecture based on clocked cells [3], which can exist in three different conditions, depending on the value of a clock signal:

- The "locked" condition corresponds to having tunneling between dots inhibited, and therefore the cell can be used to "drive" nearby cells.

Nanotechnology. Volume 4: Information Technology II. Edited by Rainer Waser

ISBN: 978-3-527-31737-0

- The "null" condition corresponds to having no electrons in the cell and therefore no polarization.
- The "active" condition is the one in which the cell adiabatically reaches the polarization condition resulting from that of the nearby cells.

The clocked QCA architecture solves the problem of unreliable evolution and allows data pipelining, but introduces a remarkable complication: the clock signal must be distributed, with proper phases, to all the cells in the array. Unless a "wireless" technique for clock distribution could be devised (some proposals have been made in this direction, but a definite solution is yet to be found), one of the most attractive features of QCA circuits – the lack of interconnections – would be lost.

Current research is focusing on the possibility of implementing QCA cells with molecules [4] or with nanomagnets [5], in order to explore the opportunities for further miniaturization (molecular cells) and for overcoming the limitations imposed by Coulomb coupling (nanomagnetic cells). However, these technologies do not seem to be suitable for fast operation: highly parallel approaches could make up for the reduced speed, but this would further complicate system design and the definition of the algorithms.

Although the basic principle of operation is sound, the above-mentioned technological difficulties and the reliability problems make practical application of QCA technology unlikely, at least in the near future. Nevertheless, the QCA concept remains of interest and the subject of lively research, because of its innovation potential and because it opens up a perspective beyond the so far unchallenged three-terminal device paradigm for computation.

In this chapter, an overview of the QCA architecture will be provided, with a discussion of its two main formulations: the ground-state relaxation approach and the clocked QCA approach. In Section 12.2 the issue of operation with cells with more than two electrons will also be addressed, as well as the details of intercell interaction. Section 12.3 will focus on the various techniques that have been developed to model QCA devices and circuits, while Section 12.4 will be devoted to the challenges facing the implementation of QCA technology. Actual physical implementations of QCA arrays will be addressed in Section 12.5, and an overlook for the future will be presented in Section 12.6.

12.2
The Quantum Cellular Automaton Concept

12.2.1
A New Architectural Paradigm for Computation

An early proposal for an architecture based on interacting quantum dots was formulated by Bakshi *et al.* in 1991 [6]: these authors considered parallel elongated quantum dots, defined "quantum dashes" (see Figure 12.1), each of which should have an occupancy of one electron. Their basic argument was that, once the electron

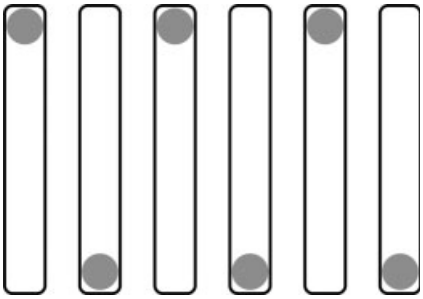


Figure 12.1 Series of elongated quantum dots (quantum dashes) with the hypothesized anti-ferroelectric ordering.

in the first dash was confined into one end of the dash, the electron in the next dash would be confined into the opposite end, as a result of electrostatic repulsion. This configuration would propagate along the line of dashes, leading to a sort of anti-ferroelectric ordering that could then be exploited for the implementation of more complex functions. This initial concept, however, had a serious problem, consisting in the fact that the localization of electrons along the chain of dashes would soon decay (See Figure 12.2), because the electrostatic repulsion due to an electron localized at the end of a dash is not sufficient to significantly localize the electron in the nearby dash, the probability density of which would just be slightly displaced. The Notre Dame group realized that this problem could be solved with the insertion of a barrier in the middle of the dash: in this way, the electron wave function must be localized on either side of the barrier and the electrostatic interaction from the electron in the nearby dash is sufficient to push the electron into the opposite half of the dash (Figure 12.3). This concept can be easily understood considering a two-level system subject to an external perturbing potential V [7]. The Hamiltonian of such a system in second quantization reads:

$$\hat{H} = \sum_{i=1,2} n_i E_i + t(b_1^\dagger b_2 + b_2^\dagger b_1) + \sum_{i=1,2} n_i q V_i, \tag{12.1}$$

where n_1 and n_2 are the occupation numbers of levels 1 and 2, respectively, $b_i^\dagger$ and b_j are the creation and annihilation operators for levels i and j, t is the coupling between

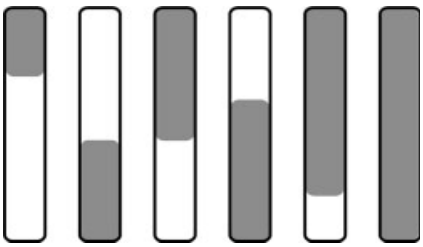


Figure 12.2 Sketch of the actual electron density within a chain of dashes.

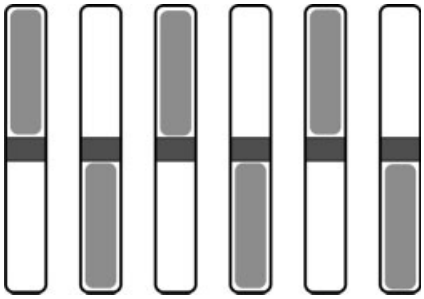


Figure 12.3 Chain of dashes with the inclusion of potential barriers: electrons are now localized on either side of the barriers and an anti-ferroelectric ordering is achieved.

the two levels and V_i is the external perturbing potential at the location of the i-th level. The creation operator $b_i^\dagger$ applied to a state with $(n-1)$ electrons yields a state with n electrons, thereby "creating" an electron in state i, while the annihilation operator b_j applied to a state with n electrons yields a state with $(n-1)$ electrons, thereby "destroying" an electron from state j. For example, the application of $b_i^\dagger b_2$ transfers an electron from level 2 to level 1. Each level can be associated with one of the sides into which the dash is divided by a potential barrier: if the barrier is exactly in the middle of the dash, $E_1 = E_2$ and the value of t depends on the height and thickness of the barrier; the higher and the thicker the barrier, the smaller t will be. A sketch of the potential profile is provided in Figure 12.4, where the dots represent the locations of the two levels 1 and 2. If $E_1 + V_1$ is chosen as the energy reference and ε is defined as $E_2 + V_2 - E_1 - V_1$, the Hamiltonian can be represented in the basis ($|0\rangle$ ($n_1 = 1$, $n_2 = 0$), $|1\rangle$ ($n_1 = 0$, $n_2 = 1$)) simply as (computing the matrix elements between the basis states)

$$\mathbf{H} = \begin{pmatrix} 0 & t \\ t & \varepsilon \end{pmatrix}. \tag{12.2}$$

The state $|0\rangle$ corresponds to having an electron in level 1 and no electron in level 2, while $|1\rangle$ corresponds to the situation with level 1 empty and an electron in level 2. The eigenvalues of this representation can be easily computed:

$$e_1 = \frac{1}{2}\left(\varepsilon - \sqrt{\varepsilon^2 + 4t^2}\right) \quad e_2 = \frac{1}{2}\left(\varepsilon + \sqrt{\varepsilon^2 + 4t^2}\right). \tag{12.3}$$

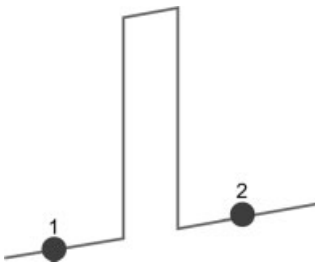


Figure 12.4 Sketch of the potential landscape defining the two levels, 1 and 2, separated by a barrier.

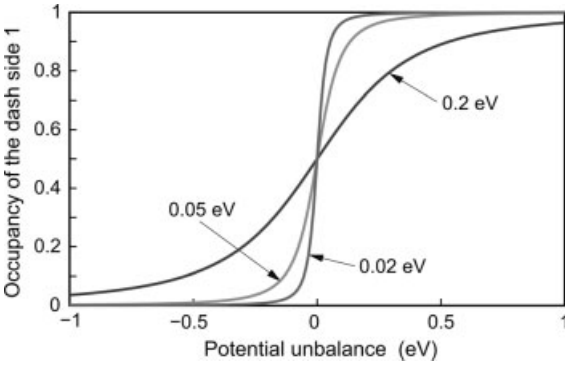


Figure 12.5 Occupancy of one of the levels of a two-level system, as a function of the potential unbalance resulting from an external applied electric field, for different values of the coupling parameter *t*.

The unbalance term ε in this case depends only on the external potential produced by the electron in the nearby dash: it will vary between a negative and a positive value, depending on the position of the electron.

The occupancy of the first level – that is, of the dash side labeled with 1 – will be given by the square modulus of the corresponding coefficient of the ground-state eigenvector, which can be computed along with the eigenvalues. Such a quantity is plotted in Figure 12.5 as a function of the unbalance ε for different values of the coupling energy *t*. It is apparent that, for low values of *t* (and therefore for an opaque barrier), the electron moves abruptly from one level to the other, as the external field is varied. Therefore, it is strongly localized even for very small values of such a field, while for high values of *t* (and therefore for a transparent or inexistent barrier) a very smooth transfer of the probability density from one level to the other are needed and large values of the perturbing field are required to achieve some degree of localization. Thus, the introduction of a barrier in the middle of the dash creates a sort of bistability – that is, a strongly non-linear response to external perturbations, which is at the heart of QCA operation and allows the regeneration of logic values. From the wire of dashes with barriers of Figure 12.3, the next step is represented by joining two adjacent dashes to form a square cell, which is the basis of the Notre Dame QCA architecture. The square cell allows the creation of two-dimensional (2-D) arrays, as shown in Figure 12.6, which can implement any combinatorial logic function. In an isolated cell the two configurations or polarization states, with the electrons along one or the other diagonal, are equally likely. However, if an external perturbation is applied, or in the presence of a nearby cell with a fixed polarization, one of them will be energetically favored. The two logic states, 1 and 0, can be associated with the two possible polarization configurations, as shown in Figure 12.7, where the solid dots represent the electrons. If a linear array of dots is created, enforcing the polarization corresponding to a given logic state on the first cell will lead to the propagation of the same polarization state along the whole chain, in a domino fashion. Such a linear array is usually defined a "binary wire", and can be used to propagate a logic variable

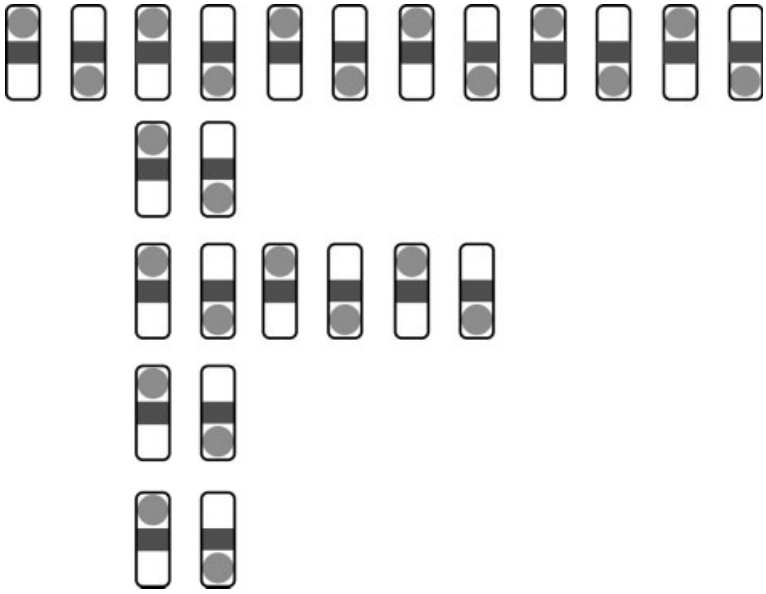


Figure 12.6 Two-dimensional array of cells made up of two adjacent dashes: this is the basis of the QCA architecture.

across a circuit: here, the strength and, at the same time, the weakness of the QCA architecture is noticed. Indeed, signal regeneration occurs during propagation along the chain, as a result of the non-linear response of the cells, but the transfer of a logic variable from one location in the circuit to a different location may require a relatively large number of cells. In other words, there are no interconnects, but the number of elementary devices needed to implement a given logic function may become much larger than in a traditional architecture.

The basic gate in QCA logic is represented by the majority voting gate, which is shown in Figure 12.8. Cells A, B and C are input cells, whose polarization state is enforced from the outside (here and in the following such "driver" cells are represented with a double boundary), while cell Y is the output cell, the polarization of which represents the result of the calculation. On the basis of a full quantum calculation or of simple considerations based on a classical electrostatic model, it is possible to show that the logic value at the output will correspond to the majority of the input values. Thus, for example, if $A = 1$, $B = 0$, $C = 0$, then $Y = 0$, or, if $A = 1$, $B = 0$, $C = 1$, the output will be $Y = 1$. From the majority voting gate it is

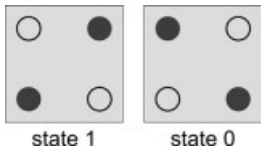


Figure 12.7 Basic configurations of a QCA cell with two electrons.

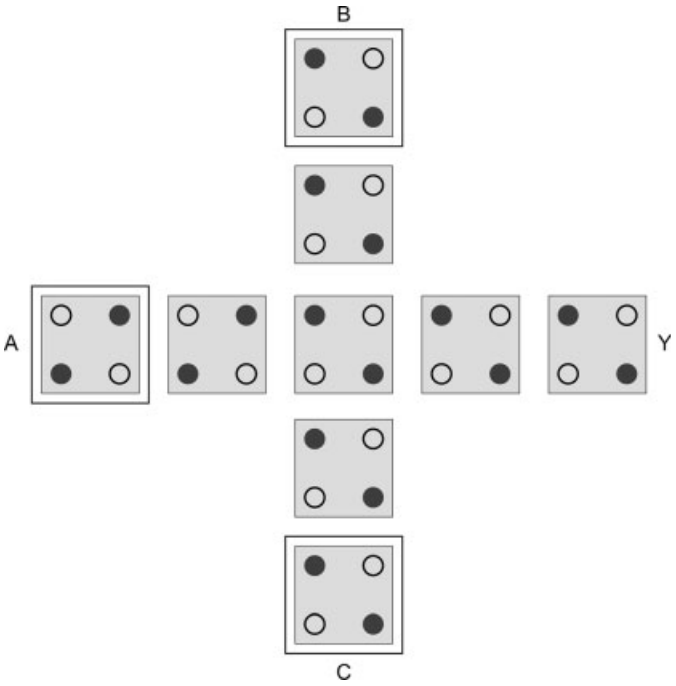


Figure 12.8 Layout of a majority voting gate.

straightforward to derive a two-input AND and OR gate: if A = 1, B and C will be the inputs of an OR gate, while if A = 0, B and C will represent the inputs of an AND gate. In order to be able to create an arbitrary combinatorial network, there is also a need for the NOT gate: this is just slightly more complex, and can be implemented with the layout shown in Figure 12.9 [15].

A generic logic function can thus be obtained with a 2-D array of cells: a number of cells at the perimeter of the array will be used as input cells, by enforcing their polarization condition with properly biased gates, and another group of perimetral

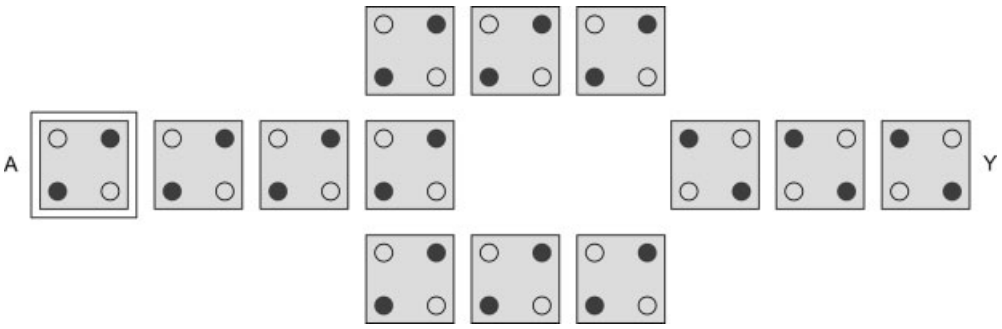


Figure 12.9 Layout for a NOT gate in the QCA architecture.

cells will act as outputs. Once the input values have been enforced, the array is allowed to evolve into the ground state and, when this has been reached, the state of the output cells corresponds to the result of the computation.

As the number of cells in the array increases, its energy spectrum – that is, the set of energies corresponding to all the possible configurations – becomes more complex, with a large number of configurations that have energies very close to the actual ground state. As a result, the evolution of the array may become stuck in one of these configurations for a long time, thus leading to a very slow computation. Furthermore, due to the appearance of states that are very close in energy to the ground state, the maximum operating temperature decreases as the number of cells is increased. In particular, it has been shown with entropy considerations [8] or by means of an analytical model [9] that, for the specific case of a binary wire, the maximum operating temperature falls logarithmically with the number of cells.

12.2.2
From the Ground-State Approach to the Clocked QCA Architecture

The above-mentioned problems severely limit the applicability of ground-state computation to real-life situations, and make the evolution of a large QCA system unreliable. To solve such problems, a modified architecture was proposed by Lent and coworkers [3], inspired, in its implementation with metal tunnel junctions, to the parametron concept introduced by Korotkov [10]. The so-called "clocked QCA architecture" derives from the concept of adiabatic switching [3]: based on the adiabatic theorem [11], it is possible to show that if the Hamiltonian of a system is made to evolve slowly from one initial form H_i to a final form H_f, a particle starting in the n-th eigenstate of H_i will be carried over into the n-th eigenstate of H_f. Thus, starting with particles in the ground state of the system, they will never leave the ground state during the evolution of the Hamiltonian, thereby preventing the previously mentioned problems of trapping into metastable states.

To implement this concept of adiabatic switching, the confinement potential defining the cell must be variable in time. In practice, the barriers separating the dots are modulated by an external potential, representing the clock signal. When the barriers are low, the cell is in the "null" state and has no polarization. In contrast, when the barriers reach their maximum height the cell is in the "locked" state and its polarization cannot be modified as a result of the action of the nearby cells (the electrons are prevented from tunneling between dots). It is only during the "active" phase, when the barriers have an intermediate height, that the polarization state can change according to the that of the nearby cells.

Attention should now be focused on the particular implementation of a clocked cell that has been experimentally realized [12]. This consists of a six-dot cell as proposed by Tóth and Lent [13], developed from the metal-island cells used in the first experimental demonstration of QCA action [2]. The barriers are represented by tunnel junctions obtained by including a very thin dielectric (usually created by oxidation) between two metallic electrodes (usually aluminum), and the quantum dots are replaced with metal islands. The six-island clocked cell is represented in

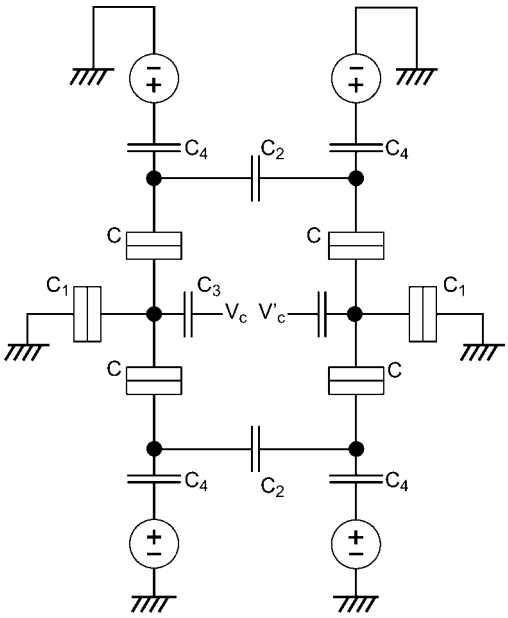


Figure 12.10 Clocked cell for implementation with tunnel junctions: tunneling between the dots of each half cell is controlled by the voltages applied to the middle dots through the C_3 capacitors.

Figure 12.10. Tunneling is possible only between the upper and the lower dot of each half of the cell (it can be shown that this does not limit in any way the logic operation of the cell) and the barrier height between the active islands is controlled by means of the potential applied to the central island. If the potential on the central island is low, then the electron will be trapped there (null state). As the potential on the central island is raised, a condition will be reached in which the electron can tunnel into one of the active dots, the one that is favored at the time by the potential created by the nearby cells (active state). Finally, as the potential on the central dot is further raised, the electron will be trapped in the dot into which it has previously tunneled, even if the polarization of the other cells is reversed (locked state).

Ideally, the computation should evolve with a cell in the locked state driving the next cell that moves from the null state to the locked state, going through the active state. When the state of a cell must be the result of that of more than one neighboring cell (as in the case of the central cell of a majority voting gate), all the cells acting on it should be at the same time in the locked state. The sequence of clock phases would allow the information to travel along the circuit in a controlled way, thus achieving a deterministic evolution and eliminating the uncertainty about the time when the calculation is actually completed that plagues the ground-state relaxation scheme. Furthermore, since the flux of data is steered by the clock, it would also possible to have data pipelining: new input data could be fed into the inputs of the array as soon as the previous data have left the first level of logic gates

and moved to the second. Ideally, within this scheme each cell should be fed a different clock phase with respect to its nearest neighbors, which would imply an extremely complex wiring structure. Such a solution has been adopted in the experiments performed so far to demonstrate the principle of operation of clocked QCA logic [12]. However, in large-scale circuits it would forfeit one of the main advantages of the QCA architecture – that is, the simplicity deriving from the lack of interconnections. In order to address this problem, it has been proposed to divide the overall QCA array into "clocking zones": such regions consist of a number of QCA cells and would be subject to the same clock phase and evolve together while in the active state (similarly to a small array operating according to the ground-state relaxation principle). They would then be locked all at the same time, in order to drive the following clocking zone. This would reduce the complexity of the required wiring, and has been proposed in particular for the implementation of QCA circuits at the molecular scale, where it is impossible to provide different clock phases to each molecule, as the wires needed for clocking would be much larger than the molecules themselves! There are many difficulties involved, however, because each clocking zone is affected by the problems typical of ground-state relaxation (although on a smaller scale), and the clock distribution is still extremely challenging. For example, conducting nanowires have been suggested as a possible solution to bring the clock signal to regions of a molecular QCA circuit, but achieving uniformity in the clocking action of a nanowire on many molecular cells is certainly a very challenging task.

Notwithstanding all of these difficulties, the clocked scheme appears to be the only one capable of yielding a reasonably reliable QCA operation in realistic circuits, as will be discussed in the following sections.

12.2.3 Cell Polarization

At this point it is necessary to provide a rigorous definition of cell polarization, in order to be able to describe quantitatively the interaction between neighboring cells and to determine whether cells with an occupancy of more than two electrons could possibly be used. Indeed, according to the initial definition of cell polarization given by the Notre Dame group, the operation of cells with more than two electrons would not be possible. Their original definition of cell polarization was

$$P = \frac{\rho_1 + \rho_3 - \rho_2 - \rho_4}{\rho_1 + \rho_2 + \rho_3 + \rho_4}, \tag{12.4}$$

where ρ_i is the charge in the i-th dot (dots are numbered counterclockwise starting from the one in the upper right quadrant). With such an expression, as soon as the number of electrons increases above two, full polarization can no longer be achieved, as the maximum possible value for the numerator is 2. There can be at most a difference of two electrons between the occupancy of one diagonal and that of the other, since configurations with a larger difference would require an extremely large external electric field (to overcome the electrostatic repulsion between electrons).

Starting from the observation that a QCA cell is overall electrically neutral, because of the presence of ionized donors, of the positive charge induced on the metal electrodes, and of the screening from surface charges, Girlanda *et al.* [14] proposed a different expression for the polarization of a cell, which is more representative of its action upon the neighboring cells. Indeed, neutralization occurs over an extended region of space; thus, although the global monopole component of the field due to a cell is zero, there can be some effect on the neighboring cells associated with the total number of electrons. However, in practical cases this turns out to be negligible compared to the dipole component associated with the charge unbalance between the two diagonals. The alternative expression for cell polarization introduced in Ref. [14] reads

$$P = \frac{\rho_1 + \rho_3 - \rho_2 - \rho_4}{2q}, \tag{12.5}$$

where q is the electron charge. Use of this expression is supported by semiclassical electrostatic considerations and by detailed quantum simulations [14], and leads to the conclusion that QCA action can be observed whenever the cell occupancy is of $4N + 2$ electrons, where N is the integer. This means that control of the occupancy of the dots is less stringent than previously expected, but is still quite difficult.

12.3 Approaches to QCA Modeling

12.3.1 Hubbard-Like Hamiltonian

The first approach to QCA simulation was developed by the Notre Dame group [15], based on an occupation number, Hubbard-like formalism. Within such an approach the details of the electronic structure of each quantum dot are neglected, and a few parameters are used to provide a description of the dots and their interaction. Although based on a few phenomenological parameters, this technique has been successful in providing a good basic understanding of the operation of QCA cells.

The occupation number Hamiltonian for a single, isolated cell reads

$$\begin{aligned} H_0 = &\sum_{i,\sigma} E_{0,i}, n_{i,\sigma} + \sum_{i>j,\sigma} t(b^{\dagger}_{i,\sigma} b_{i,\sigma} + b^{\dagger}_{j,\sigma} b_{i,\sigma}) \\ &+ \sum_{i} E_{Qi} n_{i,\uparrow} n_{i,\downarrow} + \sum_{i>j,\sigma,\sigma'} V_Q \frac{n_{i,\sigma} n_{j,\sigma'}}{|\vec{R_i} - \vec{R_j}|}, \end{aligned} \tag{12.6}$$

where $E_{0,i}$ is the ground-state energy of the i-th dot (assumed to be isolated), $b^{\dagger}_{j,\sigma}$ and $b_{j,\sigma}$ are the creation and annihilation operators, respectively, for an electron in the j-th dot with spin σ, $n_{j,\sigma}$ is the number operator for electrons in the i-th dot with spin σ, t is the tunneling energy between neighboring dots, V_Q is equal to $e^2/(4\pi\varepsilon)$ (e is the electron charge and ε is the dielectric permittivity), E_{Qi} is the on-site charging energy

for the i-th dot [16], and $\vec{R_i}$ is the position of the i-th dot center. The tunneling energy t cannot be computed directly, and must be evaluated with some approximation. A commonly used approximation consists in assuming t to be equal to half of the level-splitting resulting because of the presence of a barrier of finite height between the dots. In the presence of a driver cell in a given polarization state, the above-written Hamiltonian must be augmented with a term that expresses the electrostatic contribution from such a cell:

$$H_{int} = \sum_{i \in \text{cell1}} \sum_{j \in \text{cell2}} V_Q \frac{\rho_{j,2} - \bar{\rho}}{|\vec{R}_{j,2} - \vec{R}_{i,1}|}, \tag{12.7}$$

where $\rho_{j,2}$ is the number of electrons in the j-th dot of the driver cell, ρ is the average number of positive neutralizing charges per dot, $\vec{R}_{j,2}$ and $\vec{R}_{i,1}$ are the positions of the j-th dot of the driver cell (cell 2) and of the i-th dot of the driven cell (cell 1), respectively.

Diagonalization of the total Hamiltonian can be performed easily using an occupation number representation: $|n_{1,\uparrow}, n_{1,\downarrow}; n_{2,\uparrow}, n_{2,\downarrow}; n_{3,\uparrow}, n_{3,\downarrow}; n_{4,\uparrow}, n_{4,\downarrow}\rangle$.

The dimension of the basis, considering only two-electron states, would be 256 but, on the basis of spin considerations [17], the number of basis vectors required for the determination of the ground state is just 16.

A representation of the complete Hamiltonian on such a basis consists in a sparse matrix with only four non-zero off-diagonal elements in each row. Eigenvalues and eigenvectors can be obtained numerically, and the ground state of the driven cell will be

$$|\psi_0\rangle = \sum_{i=1}^{16} \alpha_i |i\rangle, \tag{12.8}$$

where α_i is the i-th element of the eigenvector, corresponding to the lowest eigenvalue, and $|i\rangle$ is the i-th element of the basis used for the representation of the Hamiltonian. The average charge in each dot is given by

$$\rho_i = \langle \psi_0 | n_{i\uparrow} + n_{i\downarrow} | \psi_0 \rangle, \tag{12.9}$$

from which the cell polarization can then be computed. In Figure 12.11 the polarization of the driven cell computed as a function of the polarization of the driver cell is reported; that is, the cell-to-cell response function. Calculations have been performed for a cell with four dots located at the vertices of a square with a 24 nm side. The dots have a diameter of 16 nm, except for one, the diameter of which varies between 15.94 and 16.06 nm, and the separation between the centers of the driver and driven cells is 32 nm. Material parameters for GaAs have been considered, with an effective mass $m^* = 0.067\ m_0$ and a relative permittivity $\varepsilon_r = 12.9$; furthermore the tunneling energy t has been assumed to be 0.1×10^{-3} eV.

In the case of identical dots (all with the same 16-nm diameter), the response function is symmetric, while just an extremely small variation in the diameter of a dot leads to strong asymmetry and eventually to failure of operation, with the driven cell always stuck in the same state for any value of the polarization of the driver cell. It appears that such a sensitivity to geometric tolerances is a very serious practical

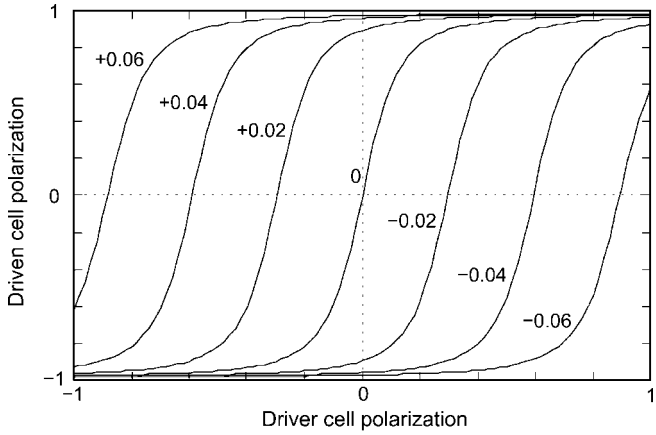


Figure 12.11 Cell-to-cell response function for cells with a separation of 32 nm, an interdot distance of 24 nm, and dots with a diameter of 16 nm. The different curves correspond to an error on the diameter of the lower left dot varying between − 0.06 nm and 0.06 nm.

problem, but cannot be fully gauged with the occupation-number Hamiltonian approach, because it is not possible to directly relate the diameter of idealized quantum dots to actual geometric quantities, such as the size of the gates defining the quantum dots.

In order to be able to provide reliable estimates of the acceptable errors on actual geometric parameters, a more realistic model is needed which takes into account the detailed structure of the cell. To this purpose, the approach described in the following subsection has been developed.

12.3.2 Configuration–Interaction

More traditional, iterative self-consistent approaches, such as the Hartree technique or techniques based on the local density functional approximation (LDA), perform very poorly in the simulation of an active QCA cell, in particular in the region around zero polarization. The main problem with iterative self-consistent methods is that, in the application to QCA problem, they tend to become unstable, as a result of the strong degeneracies and of the marked bistability of the system. One effective technique to treat a realistic model for a QCA cell consists in configuration–interaction. This method is very well known in the field of molecular chemistry [18], and has found significant application also for treating semiconductor quantum dots [19, 20].

While, for example, the Hartree–Fock method consists in finding an optimized Slater determinant representing the single-determinant solution for the many-body Schrödinger equation (i.e. the Slater determinant that minimizes the ground-state energy), in the configuration–interaction picture the many-particle wave function is

expressed as a linear combination of Slater determinants. In principle, if the basis of determinants were infinite, the solution would be exact; however, in practice a finite basis must be considered, which introduces some degree of approximation, depending on the number of elements and on how good their choice is in terms of the actual solution.

The application of configuration–interaction to the analysis of QCA cells is presented in Ref. [21]: the Hamiltonian for a cell is written as

$$\begin{aligned}\hat{H} = & -\frac{\hbar^2}{2m^*}\nabla_1^2 - \frac{\hbar^2}{2m^*}\nabla_2^2 + V_{\text{con}}(\vec{r_1}) + V_{\text{con}}(\vec{r_2}) + V_{\text{driv}}(\vec{r_1}) + V_{\text{driv}}(\vec{r_2}) \\ & + \frac{1}{4\pi\varepsilon}\frac{e^2}{|\vec{r_1}-\vec{r_2}|} - \frac{1}{4\pi\varepsilon}\frac{e^2}{\sqrt{|\vec{r_1}-\vec{r_2}|^2+(2z)^2}} - \frac{1}{4\pi\varepsilon}\frac{e^2}{2z},\end{aligned} \tag{12.10}$$

where $\hbar$ is the reduced Planck constant, m^* is the electron effective mass, V_{con} is the bare confinement potential (due to the electrodes, the ionized donors, the charged impurities, and the bandgap discontinuities), V_{driv} is the Coulomb potential due to the charge distribution in the neighboring driver cell, e is the electron charge, the last two terms include the effects of the image charges (since, for simplicity, a Dirichlet boundary condition is assumed at the surface and at an infinitely far away conducting substrate), and z is the distance of the 2DEG from the surface of the heterostructure where the boundary condition is enforced.

A matrix representation of this Hamiltonian is derived by computing the matrix elements of Eq. (12.10) between the elements of the basis of Slater determinants, and is then diagonalized, obtaining the ground-state energy as the lowest eigenvalue and the ground-state wave function as a linear combination of the basis elements with coefficients corresponding to the elements of the associated eigenvector.

This technique does not have convergence problems, as it is intrinsically a one-shot method and allows the consideration of a realistic confinement potential, obtained from a detailed numerical calculation. However, if the intention was to introduce more realistic boundary conditions for the semiconductor surface, or in general to provide a more refined treatment of the electrostatic problem, going beyond the method of images, the problem would, computationally, be very intensive. This is because, in order to compute the matrix elements of the Hamiltonian, the complete Green's function of the Poisson equation between each pair of points in the domain would be needed.

If an occupancy of only two electrons per cell were to be considered, the actual two-electron wave function is very close to the Slater determinant constructed from the one-electron wave functions of the isolated dots. Therefore, the size of the basis needed to obtain a good configuration–interaction solution is small, of the order of 100 determinants. Instead, if there are more than four electrons per cell (and thus more than one electron per dot), the strong electron–electron interaction determines a significant deformation of the wave functions and therefore a large number of basis elements constructed from the single-electron orbitals is needed. For example, in the case of a six-electron cell, more than 1000 determinants are necessary. As the number

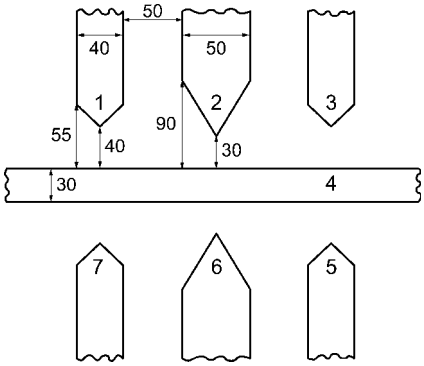


Figure 12.12 Gate layout for the definition of a working QCA cell (all measures are in nanometers).

of electrons is raised, there is a combinatorial increase in the size of the basis, and consequently the problem soon becomes intractable from a computational point of view.

Notwithstanding the above-mentioned limitations on the way that Coulomb that interaction can be included, and on the maximum number of electrons that can be considered, configuration–interaction has been very successfully applied to the simulation of QCA systems. In fact, it has allowed the demonstration that, for a semiconductor implementation, an array of holes (defining the quantum dots) in a depletion gate held at constant potential cannot possibly be fabricated with the required precision. Alternative gate arrangements, such as those shown in Figure 12.12, are possible [21], and have been used in the experimental demonstration of QCA action in GaAs/AlGaAs heterostructure-based devices [22]. However, they imply severe technological difficulties and the need for adjustment of individual gate voltages to correct for geometrical imperfections and for the unintentional asymmetries introduced by the presence of nearby cells [23].

12.3.3
Semi-Classical Models

Quantum models of QCA cells are needed to describe the bistable behavior of the single cell, and also to provide information on the technological requirements needed to obtain successful QCA operation. They are, however, too complex (from a computational point of view) to be applied to the analysis of complete QCA circuits consisting of a large number of cells. The time required to complete a simulation of a circuit made up of just a few tens of cells would be prohibitive. Therefore, a multiscale approach is needed, which is structured in a way similar to that of traditional microelectronics, where circuit portions of increasing complexity are treated with models based on progressively more simplified physical models.

It should be noted that the effect at the core of QCA action is purely classical – that is, it is the Coulomb interaction between electrons. As long as electrons are strongly localized, they behave substantially as classical particles, and a semi-classical model,

based on the minimization of the electrostatic energy, can capture most of the behavior of a QCA circuit.

If the only point of interest is to determine the ground-state configuration of an array of QCA cells and in computing the energy separation ΔE between the first excited state and the ground state, then a simple electrostatic model can be used. The quantity ΔE is essential to determine the maximum operating temperature of the circuit: it must be at least a few tens of kT (where k is the Boltzmann constant and T is the absolute temperature); otherwise, the system will not remain stably in the ground state. The basic electrostatic model developed in Ref. [9] relies on a cell model in which the charge of the two electrons is neutralized either by positive charges of value $e/2$ located in each dot, or by image charges located at a distance from the dots and representing the effect of metal electrodes or of Fermi level pinning at the semiconductor surface. Although a cell can be in the two configurations with the electrons aligned along one of the diagonals, other configurations are also possible. However, in most cases they are not energetically favored. A more complete model must also introduce such configurations, corresponding to the two electrons occupying the dots along one of the four sides of the cell. While the configurations with the electrons on the diagonals are associated with the logical values 1 and 0, the other configurations do not correspond to any logical value and are thus indicated with X in Figure 12.13, where all possible configurations are represented.

The total electrostatic energy is given by [24]:

$$E = \sum_{i \neq j} \frac{q_i q_j}{4\pi\varepsilon_0 \varepsilon_r r_{ij}} \tag{12.11}$$

If, for the sake of simplicity, the neutralizing charge is considered to be located directly in each dot (in an amount $e/2$), the total charge in each dot can assume only two values: either $+e/2$ or $-e/2$, thereby leading to

$$q_i q_j = \frac{1}{4} e^2 \mathrm{sgn}(q_i q_j) \tag{12.12}$$

If the distance between the dots is expressed in terms of the ratio $R = d/a$ and of the electron positions, the following can be written:

$$E = \frac{e^2}{4a} \frac{1}{4\pi\varepsilon_0 \varepsilon_r} \sum_{i \neq j} \frac{s_{ij}}{\sqrt{(n_{ij} R + l_{ij})^2 + m_{ij}^2}}, \tag{12.13}$$

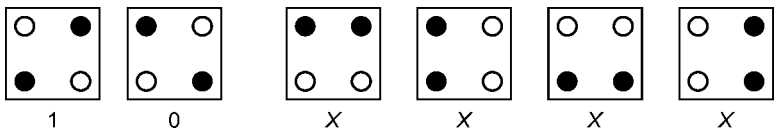


Figure 12.13 Possible configurations of a four-dot cell with two electrons. The configurations marked with *X* do not correspond to a well-defined logic state.

where $n_{ij} \in \{0, \ldots, N_{\text{cell}} - 1\}$ is the separation, in terms of number of cells, between the cell with dot i and the cell with dot j, $s_{ij} \in \{-1, 1\}$ is the sign of $q_i q_j$, $l_{ij} \in \{-1, 0, 1\}$ and $m_{ij} \in \{0, 1\}$, is the position of dots i and j within the relative cells. The quantity l_{ij} is equal to 0 if both the i and the j dots are on the left side or on the right side of the cell, to -1 if dot i is on the right side and dot j is on the left side, and to 1 if dot i is on the left side and dot j is on the right. Analogously, m_{ij} is equal to 0 if both dots i and j are on the top or on the bottom of a cell, to 1 if one dot is on the top and the other is on the bottom.

The most direct approach consists of computing the energy associated with each possible configuration by means of the direct evaluation of Eq. (12.13) and choosing the configuration that corresponds to the minimum energy. With this procedure the complete energy spectrum for the circuit is also obtained; that is, the energy values corresponding to all possible configurations. However, such a method soon becomes prohibitively expensive from a computational point of view, as the number of configurations to be considered is 6^N, where N is the number of active cells (i.e. of cells whose polarization is not enforced from the outside, as in the case of the driver cells). As the number of cells is increased, a simplified model can be used in which only the two basic states of a cell are considered, thus reducing the total number of configurations down to 2^N. This does not introduce significant errors, as long as the X states are unlikely (which is in general true), except for the case of intercell separation equal to or smaller than the cell size, when X states with both electrons vertically aligned may occur.

An example of application of the semi-classical simulation technique with six states per cell is shown in Figure 12.14, where the maximum operating temperature of a binary wire is reported as a function of the number of cells, for a 60%, 90% and 99% probability of obtaining the correct logical output, assuming an interdot distance of 40 nm, an intercell separation of 100 nm and GaAs material parameters. It should

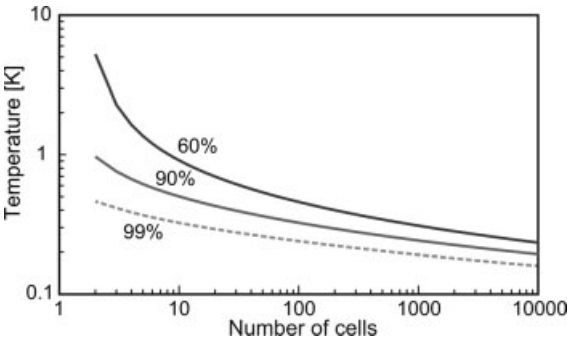


Figure 12.14 Maximum operating temperature, as a function of the number of cells, for a binary wire made up of GaAs cells with an interdot separation of 40 nm and an intercell separation of 100 nm. The maximum operating temperature has been computed for a 99%, 90% and 66% probability of obtaining the correct logical output.

be noted that the probability of obtaining the correct logical output is in general larger than the probability of being in the ground state, as there are also a number of excited states in which the polarization of the output cell has the correct value.

It is apparent that, even with the simplification down to just two states per cell, large circuits cannot be simulated with the semi-classical approach just described, because of the exponential increase in the time required to perform a complete exploration of the configuration space. This has led to the development of techniques based on an incomplete, targeted exploration of the configuration space, such as that described in the following subsection.

12.3.4 Simulated Annealing

The concept of simulated annealing derives from that of thermal annealing, whereby a material is brought into a more crystalline and regular phase by heating it and allowing it to cool slowly. Analogously, in simulated annealing the aim is to reach the ground state of the system, starting from a generic state at a relatively high temperature, and then to perform a Monte Carlo simulation in which at each step an elementary transition within a cell (chosen at random) is accepted with a probability P depending on the energy E_{old} of the system before the transition, and on the energy E_{new} after the transition:

$$P = \begin{cases} 1 & \text{if} \quad E_{\text{new}} \leq E_{\text{old}} \\ \exp[-(E_{\text{new}} - E_{\text{old}})/kT] & \text{if} \quad E_{\text{new}} > E_{\text{old}} \end{cases} \tag{12.14}$$

It is apparent that, in this way, the evolution of the system is steered along a path of decreasing energy, whilst at the same time trapping in a local minimum is prevented in most cases by the non-zero probability of climbing to a higher energy configuration. This procedure is iterated many times, gradually decreasing the temperature, until convergence to a stable configuration is achieved [17].

The application of simulated annealing to QCA circuits was originally proposed for their operation [25], and has since been applied to their modeling [26]. This has allowed the analysis of circuits with a number of cells of the order of 100 with limited computational resources and with just a few hours of CPU time. With large circuits, the simulated evolution of the circuit may occasionally become stuck in a local energy minimum, which would then be erroneously assumed as the ground state. The probability of this happening can be minimized by performing the equivalent of "thermal cycling". Once a stable state has been reached, the temperature is raised again, driving the circuit into an excited state, and a new annealing run is performed, reaching a new stable state. If the whole procedure is repeated several times, there is a better chance of reaching the ground state. It is possible to show that the probability P of the computational procedure stopping in the ground state is given by $P = 1 - (1 - \mathrm{P}_0)^m$, where P_0 is the probability of reaching the ground state without cycling, and m is the number of cycles. With this technique it is possible to reliably simulate QCA circuits with a few hundreds of cells.

12.3.5 Existing Simulators

A number of simulators have been developed to study both the static and dynamic behaviors of QCA circuits. One of the first available was AQUINAS (A Quantum Interconnected Network Array Simulator, from the Notre Dame group), where cells are modeled within a Hartree–Fock approximation and the time-dependent Schrödinger equation is solved with the Crank–Nickolson algorithm. Relatively large systems can be handled, as a result of an approximation consisting in the representation of the state of a single cell with a simplified two-dimensional basis [27]. NASA researchers have added to AQUINAS capabilities for the statistical analysis of data, thus creating TOMAS (Topology Optimization Methodology using Applied Statistics) AQUINAS [28].

A static simulator for the determination of the ground state of a QCA circuit on the basis of a classical electrostatic model has been developed by the group in Pisa, and is currently available on the Phantoms Computational Hub (http://vonbiber.iet.unipi.it). The simulator has been named QCAsim, and operates according to the approach described in Section 12.3.3. In general, it can compute the ground-state configuration of a generic array of cells via a complete exploration of the configuration space, assuming for each cell six possible configurations for the two electrons. It is possible to specify whether neutralization charges should be included and in which positions (on the same plane as the electrons, on a different plane, as image charges, etc.).

The group in Pisa has also developed a dynamic simulator, MCDot (also available on the Phantoms HUB). This was conceived specifically for the QCA implementation based on metal tunnel junctions, and is therefore based on the Orthodox Theory of the Coulomb Blockade [29] with the addition of cotunneling effects treated to first order in perturbation theory [30]. The operation of such a code will be described in more detail in Section 12.4.3 while discussing limitations for the operating speed. Although the code was originally developed for circuits with metallic tunnel junctions, its range of applicability can be easily extended to different technologies, extracting appropriate circuit parameters and defining an equivalent circuit. For example, it has been successfully applied to the simulation of silicon-based QCA cells [17]. To this purpose, linearized circuit parameters can be determined from three-dimensional simulations around a bias point and then used in MCDot. The most challenging part of the parameter extraction procedure is represented by the capacitances and resistances of the tunneling junctions obtained by defining a lithographic constriction in silicon wires [31]: the detailed geometry and the actual distribution of dopants cannot be known exactly, and resort to experimental data is often necessary.

Recently, another simulator has been developed at the University of Calgary, QCADesigner (http://www.qcadesigner.ca). This uses a two-state model for the representation of each cell, derived from the theory developed by the Notre Dame group. QCADesigner is meant to be an actual CAD (computer-aided design) tool, applicable to the design of generic QCA circuits and with the capability for testing their operation with a targeted or exhaustive choice of input vectors.

12.4
Challenges and Characteristics of QCA Technology

12.4.1
Operating Temperature

As mentioned previously, the maximum achievable operating temperature is one of the main challenges in QCA technology. Indeed, the energy separation ΔE between the first excited state and the ground state of the system must be much larger than the thermal energy, if disruption of the operation as a result of thermal fluctuations is to be prevented. Unfortunately, the magnitude of the dipole interaction between cells is very small, of the order of millielectronvolts for cells with a size of a few tens of nanometers, and is further reduced by the screening action of nearby conducting electrodes and surfaces. This is why, with currently available technologies, the operation of a QCA circuit is not conceivable at temperatures beyond 10–20 K, and has so far been demonstrated only in the 100 mK range.

An increase in operating temperature requires an increase in the strength of the dipole interaction between cells, which can be achieved by reducing the size of the cell, by decreasing the dielectric permittivity of the material in which cells are embedded, or by resorting to a new type of interaction. As far as the dielectric permittivity is concerned, for semiconductor implementations silicon is more promising than gallium arsenide, because silicon dots can be defined by etching and be embedded in silicon oxide, which has a permittivity of 3.9 (much smaller than that of gallium arsenide, which is about 12). In Figure 12.15 the maximum operating temperature is plotted as a function of cell size for the silicon–silicon oxide and the gallium arsenide–aluminum gallium arsenide material systems. While for the GaAs/AlGaAs system the variation of the permittivity between the two materials is small, and can be neglected in approximate calculations, for silicon it is assumed that most of the electric field lines go through silicon-oxide (which encompasses the dots on all sides) and therefore its permittivity is used in the calculations. It is apparent that the

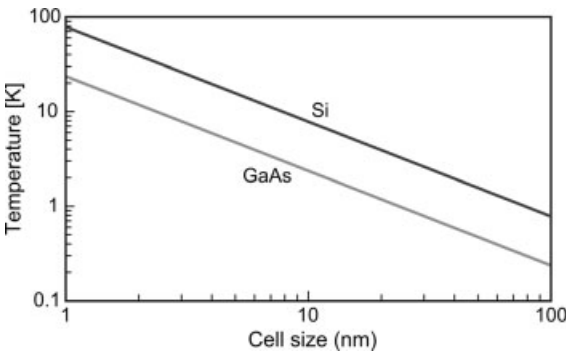


Figure 12.15 Maximum operating temperature as a function of the interdot separation within the cell, for gallium arsenide and silicon material systems.

stronger electrostatic interaction in silicon dots makes them suitable for relatively higher operating temperatures.

From Figure 12.15 it is however clear that an extremely small feature size would be needed to achieve operation at temperatures that are easily attainable.

An interaction that could allow QCA operation at room temperature is the one between nanomagnets characterized by a bistable behavior [17] (this will be discussed further in the next section). The magnetic interaction can be made strong enough to allow proper behavior of the circuit up to room temperature, but the achievable data processing speed is probably very low, of the order of a few kilohertz. On the other hand, a magnetic QCA circuit could exhibit an extremely reduced power dissipation, which could make it of interest for specific applications where speed is not a major issue, while keeping power consumption low is essential.

12.4.2 Fabrication Tolerances

The issue of fabrication tolerances has been introduced previously, and is probably the major limitation of the QCA concept, particularly for its implementation with semiconductor technologies. Detailed simulations [21] have shown that an approach based on the creation of an array of cells with dots defined by means of openings in a depletion gate cannot possibly lead to a working circuit. This is due to the fact that even extremely small errors in the geometry of such holes are sufficient to perturb the value of the confinement energy for the corresponding quantum dot to make it permanently empty or permanently occupied, no matter what the occupancy of the nearby dots is. Although shrinking the size of the cell the electrostatic interaction energy is increased, the above-mentioned problem becomes even more serious, due to the larger increase of the quantum confinement energy. An evaluation was made of the precision that could be achieved with state-of-the-art lithographic techniques and compared with the requirements for proper operation of a QCA circuit [32]. An array of square holes was obtained on a metal gate by means of electron beam lithography (Figure 12.16), after which the contour of the holes was extracted from a scanning

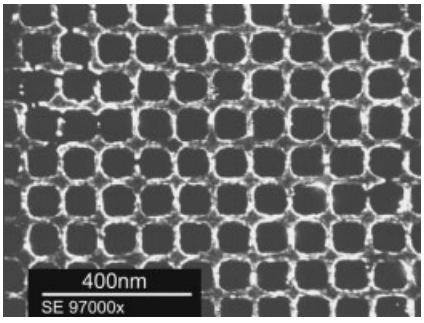


Figure 12.16 Scanning electron microscope image of the "hole array" that has been defined with state-of-the-art electron beam lithography for the purpose of evaluating the achievable precision.

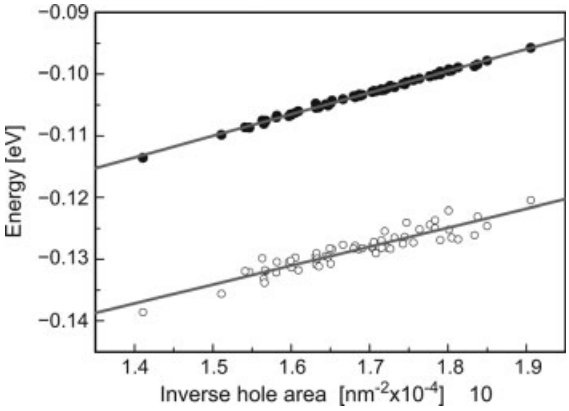


Figure 12.17 Scattering plot of the ground-state energy of single isolated quantum dots (closed circles) or of dots included in a cell (open circles) as a function of the inverse hole area.

electron microscope image and a solution of the Schrödinger equation was performed for the confinement potential obtained from each group of four holes (corresponding to a cell). The results showed a significant variance for the values of the confinement energy, as shown in Figure 12.17, where the ground-state energy for single dots and for four-dot cells is reported as a function of the inverse area of the holes defining them. From the almost linear dependence of the energy on the inverse area, it can be deduced that the local irregularities on the contour do not play an essential role, while the overall area is quite critical. It is clear that there is a dispersion of about 4 meV around the average value, while, from configuration–interaction calculations, it is shown that the allowed dispersion would be only 3 μeV, more than three orders of magnitude smaller.

Sensitivity to fabrication tolerances is ultimately the consequence of the same issue preventing operation at higher temperatures – that is, the smallness of the electrostatic interaction between cells. Imperfections are expected to play a role also with molecular-scale QCA circuits because, although molecules are in principle identical, once they are attached to a substrate any defects and stray charges from the substrate will disrupt the symmetry of the cells.

12.4.3 Limitations for the Operating Speed

The maximum operating speed of QCA circuits is ultimately limited by the dynamics of the evolution toward the ground state (if the ground-state relaxation paradigm is used), or by the tunneling rate between quantum dots. First, consider a non-clocked circuit, such as that represented for the case of a binary wire in Figure 12.18. The polarization state of the first cell is switched by inverting the bias voltages, and the cells of the wire will follow; however, according to a time evolution that may be rather

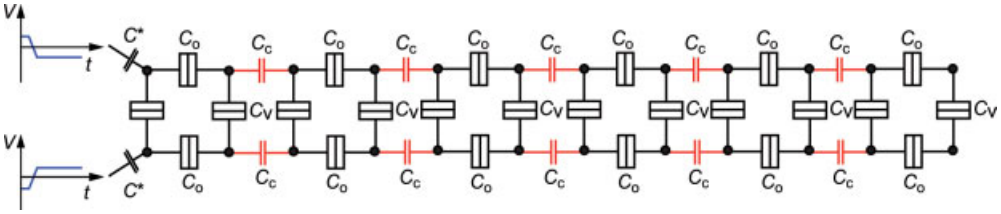


Figure 12.18 Equivalent circuit of a non-clocked binary wire; the voltages applied at the left end enforce the polarization state of the driver cell.

complex and involved. In particular, the presence of states that are very close in energy to the ground state, although corresponding to different configurations, will increase the time required for settling.

It is possible to obtain estimates of the time required for completion of the computation in a QCA array by performing simulations with a Monte Carlo approach. A Monte Carlo simulator specifically devised for QCA circuits was presented in Ref. [33]. This is based on the Orthodox Coulomb Blockade theory: the transition rate between two configurations differing by the position of one electron (which has tunneled through one of the junctions) and by a free energy variation ΔE can be expressed as

$$\Gamma = \frac{1}{e^2 R_T} \frac{\Delta E}{1 - \exp \frac{-\Delta E}{kT}}, \tag{12.15}$$

where R_T is the tunneling resistance of the junction.

Such a quantity is computed for all possible transitions, after which one of the transitions is chosen with a probability proportional to the corresponding rate. This procedure is repeated for each elementary time step into which the simulation period is divided, and the time-dependent currents in the branches of the circuit can be calculated from the contribution of the electron transitions.

This simulator can be used for the analysis of both clocked and unclocked circuits [17], as well as of a wide variety of single-electron circuits. By applying it to a six-cell binary wire with capacitances of the order of a few attofarads (values that are within the reach of lithographic technologies in the near future [33]), relaxation times of the order of 0.1 ms have been obtained; these are quite large, considering the extremely advanced technology needed for the fabrication of such devices. The reason for the slow operation is in the stochastic relaxation process, which brings the system to the ground state through a rather irregular path in the configuration space.

The Monte Carlo simulator MCdot can also be applied to the simulation of clocked QCA circuits. It has, for example, been used for the investigation of a clocked binary wire, as represented in Figure 12.19. The capacitance values are of the order of a few attofarads [33] and, at a temperature of 2.5 K, clock periods down to 0.1 μs can be achieved, as can be deduced from Figure 12.20, where the probability P_{co} of obtaining the correct logical state is plotted as a function of the clock interval for the second

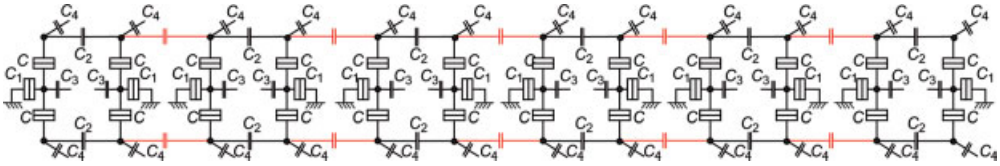


Figure 12.19 Equivalent circuit of a clocked binary wire; the C_4 capacitors are connected to voltage sources providing the bias required for proper operation, while the C_3 capacitors are connected to the clock signals.

(circles) and the last (squares) cells in the chain. As the tunneling resistances are assumed to be 200 kΩ, the resulting *RC* constant is of the order of 10^{-12} s. There is therefore a difference of about five orders of magnitude between the *RC* time constant of the circuit and the minimum clock period. This is due to a series of reasons [17]: the average time an electron takes to tunnel through a 200-kΩ junction with a 1.5-mV voltage is around 20 ps; furthermore the time during which the cell is active is only about one-tenth of the actual ramp duration; the active time, to be reasonably sure of regular operation, must be at least ten times the tunneling time; a clock period is made up of four ramps; and the intercell coupling is about five times smaller than the intracell coupling, which involves a further slow-down. Taken together, all of these effects lead to the above-mentioned reduction of the speed by five orders of magnitude with respect to the *RC* time constant, and make QCA technology not very suitable for high-speed applications.

On the other hand, the relatively slow operation of QCA circuits further limits their power dissipation, and in particular makes the power dissipated in capacitor charging–discharging negligible, as will be discussed in the next subsection.

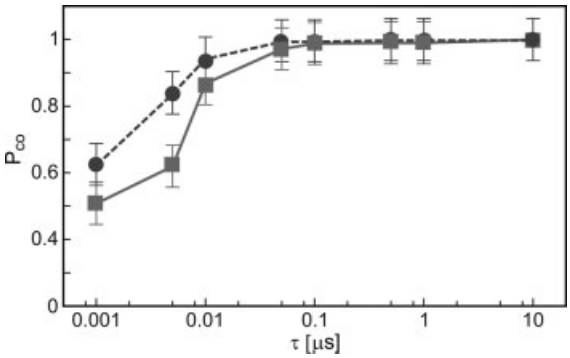


Figure 12.20 Probability (P_{co}) of correct logical output for the second (circles) and the third (squares) cell in a clocked binary wire as a function of the clock period. The *RC* constants of the circuit are of the order of 10 ps.

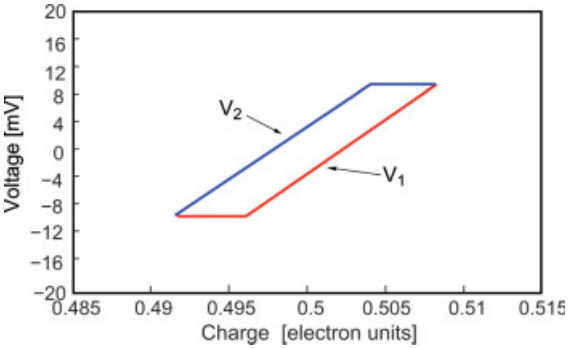


Figure 12.21 Representation of the voltages applied to a driver cell as a function of the charge flowing in the leads: the area inside the parallelogram corresponds to the work performed by the voltage sources on the QCA circuit.

12.4.4
Power Dissipation

One of the most attractive features of QCA circuits is represented by the limited power dissipation, which results mainly from the fact that there is no net transfer of charge across the whole circuit: electrons move only within the corresponding cell. The energy dissipated can be computed by integrating over the $V_i - Q$ plane [34] (Figure 12.21), where Q is the charge transferred from the source, for each external voltage source V_i and taking the algebraic sum of the results. The voltage V_2 is varied linearly until the unbalance is reversed: up to this point the charge variation corresponds to charging of the equivalent capacitance seen by V_2; then, some time after the new bias condition has been established, the electrons in the cell will tunnel, thus leading to a charge variation at constant voltage, which is represented by the horizontal segment. It is this tunneling event that makes the switch operation irreversible: without it, the area comprised between the two curves would be zero, as the voltage would simply be reversed across an equivalent capacitor, without changing its magnitude.

The energy dissipation depends on the voltage unbalance that is applied to the input cells, and that is reversed when the input data change: the larger the unbalance, the faster the switch, but the larger the dissipation, too. In the case of a single driven cell, for a typical unbalance of a few millivolts the power dissipation for a single switching event is about 10^{-22} J [35]. When considering a binary wire, the energy dissipated when the polarization of the driver cell is reversed, followed by all the other cells, and then increases very slowly as the number of cells is increased: the value for a six-cell wire is just 1% larger than that for a three-cell wire. This is due to the fact that the external voltage sources that provide energy are directly connected only to the driver cell, and the electrostatic action of the electrodes of the driver cell decays rapidly when moving along the chain. This leads to a very marginal contribution to the dissipated energy from the cells further away but, at the same

time, is the fundamental reason for the above-mentioned slow and irregular switching of a long chain.

So far, the energy loss associated with the capacitor charging process has not been included. If the unbalance reversal were abrupt, such an energy loss would be (as well known) equal to the electrostatic energy stored in the capacitor, and therefore it would represent the main contribution to dissipation. However, due to the other limitations in switching speed, there is no reason to perform such a switching with very steep ramps. It transpires from calculations, applying the expressions typically used in adiabatic logic [36], that the energy loss in capacitor charging performed with a speed compatible with the response of the circuit is negligible with respect to that due to electron tunneling in all practical cases [35].

For the case of clocked circuits the simulation is more complex and must be performed over a complete clock cycle; the conclusion is however similar, as far as a single cell is concerned: about 6×10^{-22} J dissipated in a clock cycle for the above-mentioned typical circuit parameters. In the clocked case, however, the dissipated energy is supplied by the clock electrodes directly to each cell (or clocking zone, in the case of groups of cells sharing the same clock phase), and therefore there is a linear increase in the energy dissipation as the number of cells is increased, contrary to what happens with the unclocked circuits. Indeed, with the clocked architecture there is an improvement in terms of speed and regularity of operation, but the power consumption in increased. Also in this case, it is possible to show that the contribution from the energy loss associated with capacitor charging is negligible, because the clock ramps can be much slower than the relevant time constants in the circuit.

Overall, the dissipation for a switching event of a single QCA cell is four orders of magnitude smaller than that projected for an end-of-the-roadmap MOS transistor by the ITRS Roadmap. However, the transistor operates at 300 K, while the simulations have been performed, for the clocked case, at 2.5 K. Cooling down to such a temperature requires energy, which can be estimated on the basis of the efficiency of a Carnot cycle refrigerator [37]. Inclusion of the energy lost for cooling reduces the ratio of the energy dissipated in a transistor to that dissipated in a QCA cell by two orders of magnitude. The advantage of the QCA cell still remains two orders of magnitude, but a fair comparison would require a relatively large effort, as a larger number of QCA cells is in general needed than transistors in order to obtain the same logic function. Furthermore, the energy savings that can be obtained in CMOS adiabatic logic should also be taken into consideration.

12.5
Physical Implementations of the QCA Architecture

12.5.1
Implementation with Metallic Junctions

The implementation of QCA circuits with metal islands connected by tunnel junctions was introduced in the previous sections. Tunnel junctions with extremely

small area (and therefore very small capacitance) can be fabricated between slightly overlapping electrodes, on top of a silicon oxide substrate, using the shadow mask evaporation technique. The QCA array in this case corresponds to a single-electron circuit, with tunnel junctions, capacitors, and voltage sources. From a technological point of view, a circuit with metallic junctions is relatively simple to implement, but has the major drawback, with currently available fabrication capabilities, of yielding capacitances no smaller than a few hundred attofarads [17], thus making operation possible only at temperatures of 100 mK or lower.

With such a technique, several QCA circuits have been demonstrated by the Notre Dame group: the basic driver cell-driven cell interaction [2], the operation of a clocked cell [12], a clocked QCA latch [38], and power gain in a QCA chain [34].

The problems connected with the undesired influence of each electrode on the proper balance of the cells via stray capacitances have been solved with a clever experimental scheme, based on an initial evaluation of the capacitance matrix among all electrodes. Once this is known, when the voltage of one electrode is varied, the voltages applied to all the other electrodes can be corrected in such a way as to compensate for the effects deriving from undesired capacitive couplings.

Although this technology has been very successful in the experimental demonstration of QCA operation, it appears very difficult to scale it down in order to increase the operating temperature so that it can be applied to large-scale circuits.

12.5.2 Semiconductor-Based Implementation

There are two main semiconductor implementations of QCA technology that have been attempted: one based on the Si/SiO_2 material system, and the other on the GaAs/AlGaAs material system. As previously stated, silicon has the advantage of the reduced permittivity of SiO_2, which allows the operating temperature to be raised but the fabrication of nanostructures in GaAs (defined by means of depletion gates) is more developed and tested.

For the approach based on GaAs/AsGaAs, a high-mobility, two-dimensional electron gas (2DEG) is formed at the heterointerface, and the quantum dots forming a cell are obtained by means of electrostatic depletion performed with properly shaped metal gates (Figure 12.22). In the experiments conducted to date, there is a hint of QCA effect, but it has not been possible to obtain full cell operation due to the too-small value of the capacitance coupling the upper with the lower dots across the barrier created by the horizontal electrode.

As the 2DEG is a few tens of nanometers below the surface, it is not possible to effectively define (at the 2DEG level) features that are significantly smaller; this also implies that dots cannot be made very close to each other, which represents a limitation on the maximum achievable interdot capacitance.

The advantage of GaAs technology is that tunnel barriers between dots can be finely tuned (contrary to what happens with the silicon–silicon oxide material system; see Section 12.5.3) by adjusting the bias voltage applied to the split gates defining them. In the cell represented in Figure 12.22, tunneling can occur between the top dots and

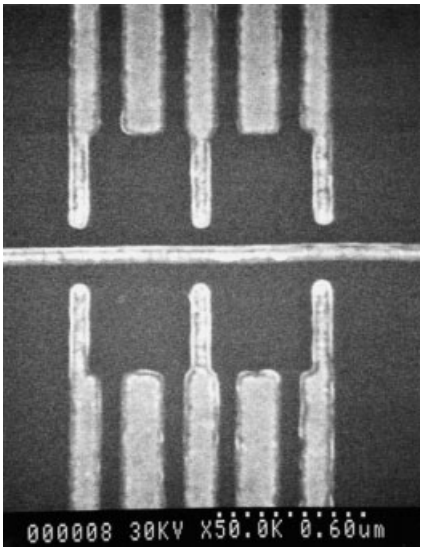


Figure 12.22 Scanning electron microscope image of the gate layout used for the investigation of the feasibility of a QCA cell in the GaAs/AlGaAs material system.

between the bottom dots, but not between one of the top dots and one of the bottom dots. This is not a problem, however, as the two configurations, with the electrons aligned along either diagonal, can still be achieved, and thus cell operation is unaltered.

A series of experiments has been performed on the prototype GaAs cell, operating, for example, the bottom part, while using one of the split gates in the top part as a non-invasive charge detector [39, 40], to monitor the motion of electrons between the two bottom dots. The outer quantum point contacts (QPC) in the bottom part are pinched off, to guarantee that the total number of electrons remains constant. Therefore, as a result of a variation of the voltage applied to the plunger gate of the dot at the bottom left (the shorter gate in the middle of the dot region), it is possible to observe motion of an electron from one dot to the other: as the plunger gate becomes more negative, an electron is moved from left to right. It has been observed [39] that the motion of an electron between the two bottom dots causes a shift of the Coulomb blockade peaks relative to one of the upper dots by about 20% of the separation between two consecutive peaks: this coupling is estimated to be sufficient to determine a reverse motion of an electron between the upper dots (i.e. the basic QCA effect).

The gate layout used for this experiment can also be applied to the implementation of a binary wire, but not for general logic circuits, because lateral branching is not possible due to the presence of the leads reaching each gate. It should also be pointed out that, even for the implementation of a simple binary wire, a careful balancing procedure would be needed, because even the finite length of the wire may be sufficient to create a fatal unbalance for all the cells, except for the one in the middle [23].

The other semiconductor implementation that has been attempted is based on silicon dots embedded in silicon oxide. As mentioned above, this material system has

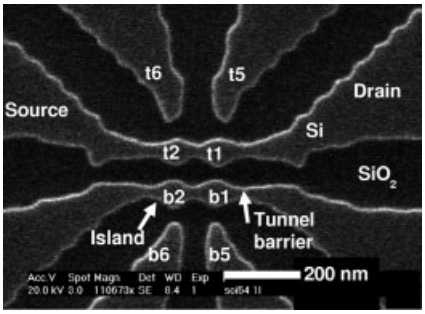


Figure 12.23 Scanning electron microscope image of a prototype QCA cell fabricated with the silicon/silicon oxide material system.

the advantage of the lower permittivity of silicon oxide with respect to gallium arsenide. However, although smaller feature sizes are achievable, control of the tunnel barriers is quite difficult as they are obtained by lithographically defining a narrower silicon region between two adjacent dots [31]. A prototype silicon QCA cell was fabricated at the University of Tübingen starting from a SOI (Silicon-On-Insulator) substrate, defining the structure by means of electron-beam lithography and reactive ion etching. The lithographically defined features are then further shrunk by means of successive oxidations [41]. It can be seen in Figure 12.23 that the tunneling junctions (between the two upper and the two lower dots) have been obtained by creating a narrower region, with a cross-section small enough that it does not allow propagation of electrons at the Fermi energy.

Such tunnel junctions are not easily controllable and, depending on the value of the Fermi energy and on the distribution of charged impurities, they may contain multiple barriers. However, it has been shown [17, 42] that, by properly tuning the back-gate voltage and the bias voltages applied to the gates, it is possible to achieve a condition in which both junctions contain a single barrier. Clear control of dot occupancy by means of the external gates has been demonstrated, by monitoring the conductance of the upper and lower double-dot systems. A clear demonstration of the QCA effect has not yet been possible due to the limited capacitive coupling between the upper and the lower double dots. However, simulations have shown [17] that a modified layout, with reduced spacing between the upper and the lower parts, should allow the observation of cell operation at a temperature of 0.3 K (probably up to 1 K), which is definitely higher than that required for metal dots and for GaAs. Unfortunately, also in this case, the basic layout used for the experiments should be significantly modified to make it suitable for complex circuits.

12.5.3
Molecular QCA

Another possible approach to the implementation of QCA circuits, as pioneered by Lent [4], is based on single molecules: this would satisfy the ultimate miniaturization requirement (a single molecule for a single computational function) and possibly

reduce the precision constraints, exploiting the fact that molecules are identical by construction. Furthermore, approaches to fabrication based on self-assembly could be envisioned, which would significantly decrease fabrication costs.

The molecular QCA concept relies on molecules containing four (or possibly two) sites where excess electrons can be located and which are separated by potential barriers. It has been demonstrated that potential barriers do exist at the molecular level and that they do lead to bistability effects [7]: a simple example is represented by a methylene group ($–CH_2–$) placed between two phenyl rings.

For the implementation of a complete cell, several candidate molecular structures have been proposed, such as metallorganic mixed-valence compounds containing four ruthenium atoms that represent the four dots. However, investigations are continuing to determine whether sufficient coupling is achievable between cells, because the screening action of the electronic clouds of the ligand atoms may determine too large a suppression of the electrostatic interaction. Furthermore, the problem of attaching the molecules to a substrate, in order to create properly structured arrays, is still only partially solved. In particular, the presence of imperfections or unavoidable stray charges at the surface of the substrate may create asymmetries preventing correct QCA operation, notwithstanding the identity of all molecules.

A simple molecule that has been proposed by the Notre Dame group as a model system for half of a cell is the so-called Aviram molecule [43], in which the two dots are represented by allyl groups at the ends of an alkane chain. Quantum chemistry calculations have shown that some bistability effects can be obtained, as well as sufficient electrostatic interaction between neighboring molecules, although, due to the reactivity of the allyl groups and to the difficulty to attach this molecule to a substrate, it does not seem a likely candidate for experiments.

Whilst overall the molecular approach seems the most appropriate solution for the implementation of the QCA concept in the long term, many problems – some of which are fundamental in nature – remain unsolved, such as finding a reliable way to assemble molecular arrays, managing the effect of stray charges, and determining whether the interplay of molecular sizes and screening effects will allow reasonably high operating temperatures.

12.5.4 Nanomagnetic QCA

To date, implementations of the QCA concept have been considered that rely on an electrostatic interaction between dots within a cell and between neighboring cells. It is also possible, as mentioned in the introduction, to exploit other forms of interaction, which may be less susceptible to the effects of temperature and of imperfections. One such alternative solution is represented by nanomagnetic QCA circuits. The concept is rather simple: an array of elongated single-domain dots obtained from properly chosen magnetic material will relax into an antiferromagnetic ordering, and it will be possible to drive the evolution of the system with an external clock consisting in an oscillating magnetic field that also supplies the energy needed for power gain

along the circuit. The first experimental investigation into the possibility of propagating the magnetic polarization along a chain of nanomagnets was performed by Cowburn and Welland [44], who managed to show the operation of a chain of magnetic nanodots.

The specific nanomagnetic approach to QCA circuits has been investigated mainly by Csaba and Porod, who have determined that an energy difference between the ground state and the first excited state of $150\,kT$ at room temperature can be achieved with elongated nanomagnets that are manufacturable with existing technologies. However, there is also a relatively high barrier ($100\,kT$) between the two states, and therefore at room temperature the system would be stable in both configurations. Thus, a pure ground-state relaxation scheme is not applicable, and resort must be made to the above-mentioned oscillating clock field. Such a field is used to turn the magnetic moments of all nanomagnets into a neutral state, from which they can relax into the ground state (as long as they remain in the instantaneous ground state).

A chain with an even number of cells (including the driver cell) will act as an inverter, as antiferromagnetic ordering is present. Thus, implementation of the NOT gate is straightforward; the majority voting gate can be implemented [17] in a way similar to that for electrostatically coupled QCA systems, with three binary wires converging on a single cell, from which another binary wire representing the output departs. Therefore, a generic combinatorial network can be realized in a way quite similar to what has been seen for other QCA technologies.

Simulators for nanomagnetic QCA circuits have been developed by Csaba and Porod [45], in which the complete micromagnetic equations are solved numerically. It has also been noticed that, for dot sizes below 100 nm, a significant simplification can be used – the single-domain approximation – in which the magnetization condition of the dots can be represented by means of single vectors instead of vector fields. Such an approximation is valid because magnetic dots below a size of 100 nm operate as single domains. The equations governing the evolution of single domains can be written as a system of ordinary differential equations and may then be recast into the form of a standard SPICE model. This allows efficient and easy simulation of relatively complex architectures of nanomagnetic QCAs, and has made it possible to show that logic signal restoration and power gain can be achieved, at the expense of the external oscillating magnetic field.

12.5.5 Split-Current QCA

An alternative approach to QCA implementation has been proposed by Walus *et al.* [46], who suggested a QCA cell in which tunneling of electrons between the dots does not take place; rather, the interaction is between tunneling currents that flow vertically through a double resonant tunneling diode structure. The cross-section of the cell proposed by Walus *et al.* is shown in Figure 12.24: the lower quantum well region of the double-resonant tunneling structure is partitioned into four dots in the horizontal plane.

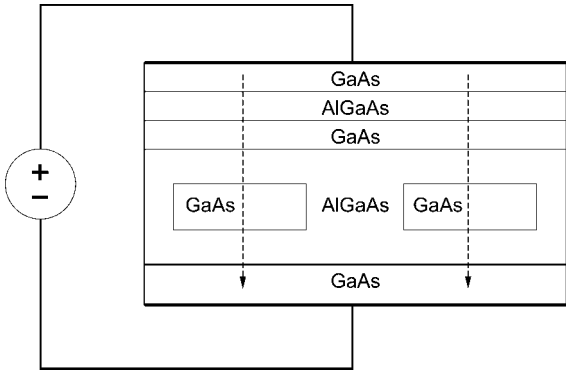


Figure 12.24 Cross-section of the split-current QCA cell, based on four parallel double-resonant tunneling structures.

The current flows mainly through these four dots, and the actual value of the current through each dot is strongly dependent on the position of the alignment of the energy levels in the upper and lower GaAs wells. Starting from a situation where the upper and lower levels are aligned, the flow of current will create a charge density that, in turn, will perturb the position of the resonant levels in the nearby dots: the larger the current, the greater the induced level shift. Therefore, for an isolated cell, the rest condition will be with current flowing mainly through either pair of antipodal dots (i.e. the dots that are furthest from each other), so that the resonant level shift is minimized. If another, driver, cell is placed next to it, with a well-defined polarization, the same polarization state will be obtained, as a result of Coulomb interaction between dots. Thus, operation similar to that of previously discussed electrostatically coupled cells will be achieved. The authors of Ref. [46] suggest that clocking is also possible, by controlling the voltage applied in the vertical direction across the resonant tunneling structures.

Although interesting in principle, this approach forfeits one of the main advantages of the QCA architecture, namely the potentially extremely low power consumption, since non-zero currents flowing in the vertical direction through the resonant tunneling diodes are always present, except for the "null" phase of the clock.

12.6 Outlook

The QCA concept has been the subject of significant research activity throughout the past decade, leading to results of general interest in the field of nanoelectronics. The practical implementation of QCA circuits is, however, still elusive, because of a few major problems connected with the weakness of the proposed cell-to-cell interaction mechanisms and with the extreme sensitivity to fabrication tolerances. Novel concepts are being explored, in particular aimed at the ultimate miniaturization, with cells consisting of single molecules, or aimed at an increase of inter-cell interaction, with cells made up of single-domain nanomagnets.

It is possible that these will lead to applications in niche markets in which extremely low power consumption is essential and high data processing speed is not a requirement. On the basis of the limited achievable switching speed and of the functional density not expected (if realistically evaluated) to be much higher than that achievable with CMOS technology, it is unlikely that QCA circuits will find application in large-scale integration. The QCA concept, however, can be at the basis of applications that go beyond its original purpose, for example in the field of metrology, where some of its weaknesses, such as the extreme sensitivity to external charges, may become important assets.

Overall, in the – so far unsuccessful – quest for a technology capable of succeeding CMOS, QCA have represented a very interesting diversion. Although such an approach may be too bold in relation to existing and near-future technological capabilities, it has contributed a wealth of novel understanding about the ultimate limitations of computation at the nanoscale.

References

1 C. S. Lent, P. D. Tougaw, W. Porod, *Appl. Phys. Lett*. 1993, **62**, 714.

2 I. Amlani, A. O. Orlov, G. L. Snider, C. S. Lent, G. H. Bernstein, *Appl. Phys. Lett*. 1998, **72**, 2179.

3 C. S. Lent, P. D. Tougaw, *Proc. IEEE* 1997, **85**, 541.

4 C. S. Lent, *Science* 2000, **288**, 1597.

5 G. Csaba, A. Imre, G. H. Bernstein, W. Porod, V. Metlushko, *IEEE Trans. Nanotechnol*. 2002, **1**, 209.

6 P. Bakshi, D. A. Broido, K. Kempa, *J. Appl. Phys*. 1991, **70**, 5150.

7 M. Girlanda, M. Macucci, *J. Phys. Chem. A* 2003, **107**, 706.

8 C. S. Lent, P. D. Tougaw, W. Porod, in: Proceedings of the Workshop on Physics and Computation - Physcomp, Dallas, Texas, November 17–20, IEEE Computer Press, pp. 1–13, 1994.

9 C. Ungarelli, S. Francaviglia, M. Macucci, G. Iannaccone, *J. Appl. Phys*. 2000, **87**, 7320.

10 A. N. Korotkov, *Appl. Phys. Lett*. 1995, **67**, 2412.

11 D. J. Griffths, *Introduction to Quantum Mechanics*, Prentice-Hall, Englewood Cliffs, NJ, 1994.

12 A. O. Orlov, I. Amlani, R. K. Kummamuru, R. Ramasubramaniam, G. Toth, C. S. Lent, G. H. Bernstein, G. L. Snider, *Appl. Phys. Lett*. 2000, **77**, 295.

13 G. Toth, C. S. Lent, *J. Appl. Phys*. 1999, **85**, 2977.

14 M. Girlanda, M. Governale, M. Macucci, G. Iannaccone, *Appl. Phys. Lett*. 1999, **75**, 3198.

15 P. D. Tougaw, C. S. Lent, *J. Appl. Phys*. 1994, **75**, 1818.

16 C. A. Stafford, S. Das Sarma, *Phys. Rev. Lett*. 1994, **72**, 3590.

17 M. Macucci (Ed.), *Quantum Cellular Automata: Theory, Experimentation and Prospects*, Imperial College Press, London, 2006.

18 R. McWeeny, *Methods of Molecular Quantum Mechanics*, Academic Press, London, 1989.

19 G. W. Bryant, *Phys. Rev. Lett*. 1987, **59**, 1140.

20 M. Brasken, M. Lindberg, D. Sundholm, J. Olsen, *Phys. Rev. B* 2000, **61**, 7652.

21 M. Governale, M. Macucci, G. Iannaccone, C. Ungarelli, J. Martorell, *J. Appl. Phys*. 1999, **85**, 2962.

22 S. Gardelis, C. G. Smith, J. Cooper, D. A. Ritchie, E. H. Linfield, Y. Jin, *Phys. Rev. B* 2003, **67**, 033302.

23 M. Girlanda, M. Macucci, *J. Appl. Phys*. 2002, **92**, 536.

24 J. D. Jackson, *Classical Electrodynamics*, Wiley, New York, 1962.

25 M. Akazawa, Y. Amemiya, N. Shibata, *J. Appl. Phys.* 1997, **82**, 5176.

26 M. Macucci, G. Iannaccone, S. Francaviglia, B. Pellegrini, *Int. J. Circul. Theoret. Appl.* 2001, **29**, 37.

27 P. D. Tougaw, C. S. Lent, *J. Appl. Phys.* 1996, **80**, 4722.

28 C. D. Armstrong, W. M. Humphreys, A. Fijany, The design of fault-tolerant quantum dot cellular automata based logic, in: *Proceedings, 2nd International Workshop on Quantum Dots for Quantum Computing and Classical Size Effect Circuits*, University of Notre Dame, August 7–9, 2003. Also available at: http://www.cambr.uidaho.edu/symposiums/symp11.asp.

29 V. Averin, K. K. Likharev, in: B. L. Altshuler, P. A. Lee, R. A. Webb (Eds.), *Mesoscopic Phenomena in Solids*, Elsevier, Amsterdam, 1991.

30 L. R. C. Fonseca, A. N. Korotov, K. K. Likharev, A. A. Odinstov, *J. Appl. Phys.* 1995, **78**, 3238.

31 R. Augke, W. Eberhardt, C. Single, F. E. Prins, D. A. Wharam, D. P. Kern, *Appl. Phys. Lett.* 2000, **76** 2065.

32 M. Macucci, G. Iannacone, C. Vieu, H. Launois, Y. Jin, *Superlatt. Microstruct.* 2000, **27**, 359.

33 L. Bonci, M. Gattobigio, G. Iannaccone, M. Macucci, *J. Appl. Phys.* 2002, **92**, 3169.

34 R. K. Kummamuru, J. Timler, G. Toth, C. S. Lent, R. Ramasubramaniam, A. O. Orlov, G. H. Bernstein, G. L. Snider, *Appl. Phys. Lett.* 2002, **81**, 1332.

35 L. Bonci, M. Macucci, in: Proceedings of the European Conference on Circuit Theory and Design, Cork, Ireland, vol. II, p. 239. Also available at: http://ieeexplore.ieee.org/xpl/RecentCon.jsp?punumber=10211.

36 R. C. Merkle, *Nanotechnology* 1993, **4**, 21.

37 International Technology Roadmap for Semiconductors (ITRS), 2005 Edition Semiconductor Industry Association.

38 A. O. Orlov, R. K. Kummamuru, R. Ramasubramaniam, G. Toth, C. S. Lent, G. H. Bernstein, G. L. Snider, *Appl. Phys. Lett.* 2001, **78**, 1625.

39 M. Field, C. G. Smith, M. Pepper, D. A. Ritchie, J. E. F. Frost, G. A. C. Jones, D. G. Hasko, *Phys. Rev. Lett.* 1993, **70**, 1311.

40 G. Iannaccone, C. Ungarelli, M. Macucci, E. Amirante, M. Governale, *Thin Solid Films* 1998, **336**, 145.

41 C. Single, R. Augke, F. E. Prins, D. P. Kern, *Semicond. Sci. Technol.* 1999, **14**, 1165.

42 C. Single, F. E. Prins, D. P. Kern, *Appl. Phys. Lett.* 2001, **78**, 1421.

43 A. Aviram, *J. Am. Chem. Soc.* 1988, **110**, 5687.

44 R. P. Cowburn, M. E. Welland, *Science* 2000, **287**, 1466.

45 G. Csaba, W. Porod, *J. Comput. Electronics* 2002, **1**, 87.

46 K. Walus, A. Budiman, G. A. Jullien, *IEEE Trans. Nanotech.* 2004, **3**, 249.

13
Quantum Computation: Principles and Solid-State Concepts

Martin Weides and Edward Goldobin

> *. . . how can we simulate the quantum mechanics? . . . Can you do it with a new kind of computer - a quantum computer? It is not a Turing machine, but a machine of a different kind.* R. P. Feynman, 1982 [1]

13.1
Introduction to Quantum Computing

For half a century the conventional electronic information processing based on Boolean logic and using a von Neumann-type architecture has been very successful in solving many numerical problems, and its computational power is still increasing. The term *Neumann-type architecture* describes a device, which implements a so-called *Turing machine* by a specifying sequential architectures of information processing. In short, a *Turing machine* contains a program (software), a finite state control, a tape (memory) and a read-write tape-head [2]. It can be proven that conventional computers are equivalent to a Turing machine.

However, there are – and will continue to be – some restrictions for conventional computation, as will be seen below. Most of today's electronics is based on devices with digital logics, apart from the very specialized analog computers, which can solve for example differential equations up to a certain size. Both the feature size and energy consumption per logic step of conventional computers have been much reduced in the recent decades, and will continue to do so for some more years [3]. The *total* energy dissipation per unit area is still increasing due to increasing packaging density. A power density of $\sim 100\,\mathrm{W\,cm^{-2}}$ can be regarded as a reasonable limit, given by the thermal conductivity of materials and the geometry for setting up temperature gradients. Note that a common kitchen heating plate at full power has as a tenth of this power density. This limit will be reached in a few decades.

If one day the energy dissipation per a single logic step is equal to $k_B T \ln 2$, where $k_B = 1.38 \times 10^{-23}\,\mathrm{J\,K^{-1}}$ is the Boltzmann constant and T the theoretical limiting

Nanotechnology. Volume 4: Information Technology II. Edited by Rainer Waser

ISBN: 978-3-527-31737-0

value of temperature, the device has to use a so-called reversible logic (see Section 13.2). The implementation of reversible computing demands a precise control of the physical dynamics of the computation machine to prevent a partial dissipation of the input information (i.e. energy) into the form of heat. One type of reversible computer is the *quantum computer* which, by definition, relies on the time-reversible quantum mechanics.

A quantum computer is a device for information processing that is based on distinctively quantum mechanical phenomena, such as quantization of physical quantities, superposition and entanglement of states, to perform operations on data. The amount of data in a quantum computer is measured in quantum bits or *qubits*, whereas a conventional digital computer uses binary digits, in short: *bits*. The quantum computation relies on the quantum information carriers, which are single quantities, whereas the conventional computer uses a huge number of information carriers. In addition, the quantum devices may be much more powerful than the conventional devices, as a different class of (quantum) algorithm exploiting quantum parallelism can be implemented. Some specific problems cannot be solved efficiently on classical computers because the computation time would be astronomically large. However, they could be solved in reasonable time by quantum computers. This emerging technology attracts much attention and effort, both from the scientific community and industry, although the possibility of building such a device with a large number ($\gg$10) of qubits is still not answered.

For a more detailed introduction to classical and quantum computation, the interested reader is referred to a great selection of excellent textbooks such as, for example, Feynman Lectures on Computation [2] and others [4, 5].

13.1.1 The Power of Quantum Computers

In recent years there has been a growing interest in quantum computation, as some problems, which are practically intractable with classical algorithms based on digital logic, can be solved much faster by massive parallelism provided by the superposition principle of quantum mechanics.

In theoretical computer science, all problems can be divided into several classes of complexity, which represents the number of steps of the most efficient algorithm needed to solve the problem. The class **P** consists of all problems that can be solved on a Turing machine in a time which is a *polynomial* function of the input size. The class **BPP** (Bound-error, Probabilistic, Polynomial time) is solvable by a *probabilistic* Turing machine in polynomial time, with a given small (but non-zero) error probability for all instances. It contains all problems that could be solved by a conventional computer within a certain probability. It comprises all problems of **P**. The class **NP** consists of all problems whose solutions can be *verified* in polynomial time, or equivalently, whose solution can be *found* in polynomial time on a non-deterministic machine. Interestingly, no proof for $\mathbf{P} \neq \mathbf{NP}$ – that is, whether **NP** is the same set as **P** – has yet been found. Thus, the possibility that $\mathbf{P} = \mathbf{NP}$ remains, although this is not believed by many computer scientists.

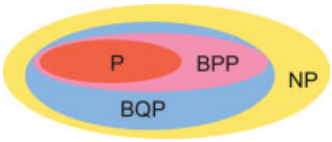


Figure 13.1 Diagram of complexity classes. A quantum computer can solve **BQP** (Bounded error, Quantum, Polynomial time) problems, whereas digital logic can just solve problems from the **P** (polynomial) class efficiently.

The class of problems that can be efficiently solved by *quantum computers*, called **BQP** (bounded error, quantum, polynomial time), is a strict superset of **BPP** (see Figure 13.1). For details on complexity classes, see Nielsen and Chuang [4].

The most important problems from the **BQP** class that cannot be solved efficiently by conventional computation (i.e. they do not belong to **P** or **BPP** class) but by quantum computation, are summarized below.

13.1.1.1 Sorting and Searching of Databases (Grover's Algorithm)

Quantum computers should be able to search unsorted databases of N elements in $\sim\sqrt{N}$ queries, as shown by Grover [6], rather than the linear classical search algorithm which takes $\sim N$ steps of a conventional machine (see Figure 13.2). This speedup is considerable when N is getting large.

13.1.1.2 Factorizing of Large Numbers (Shor's Algorithm)

A quantum algorithm for the factorization of large numbers was proposed by Shor [7], who showed that quantum computers could factor large numbers into their prime factors in a polynomial number of steps, compared to an exponential

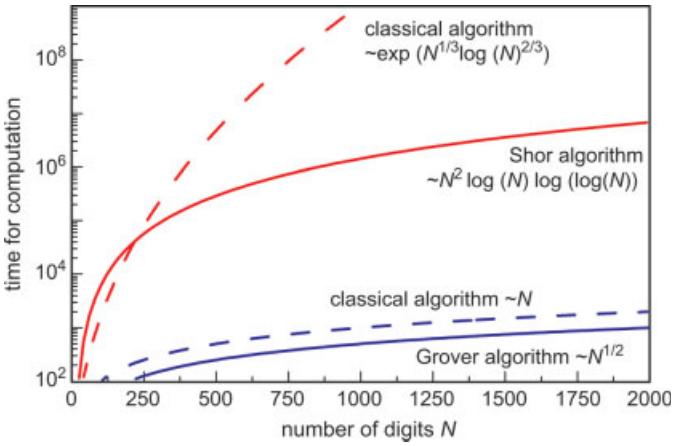


Figure 13.2 (In red): Factorization of a N digit number. (In blue): searching an unsorted database of N elements. Time requested for classical (dashed line) versus quantum (continuous line) algorithm. Note the logarithmic scale of the ordinate.

number of steps on classical computers (see Figure 13.2). The difficulty of prime factorization is a cornerstone of modern public key cryptographic protocols. The successful implementation of Shor's algorithm may lead to a revolution in cryptography.

13.1.1.3 Cryptography and Quantum Communication

Contrary to the classical bit, an arbitrary quantum state cannot be copied (no-cloning theorem; see Section 13.3.4) and may be used for secure communication by means of quantum key distribution or quantum teleportation. By sharing an entangled pair of qubits (so-called EPS-pair after a famous paper from Einstein, Podolski and Rosen [8]), signals can be transmitted that cannot be eavesdropped upon without leaving a trace; that is, performing a measurement and thereby destroying the entangled state.

All of these three important challenges for quantum computations have been implemented on NMR qubits or photons; that is, their feasibility has been proven using a small set of qubits.

13.2 Types of Computation

Here, information theory and logic will be briefly reviewed.

13.2.1 Mathematical Definition of Information

From statistical thermodynamics it is known that the entropy S of a system is defined as

$$S = k_B \ln \Omega$$

where Ω is the number of possible configurations or microscopic states. If an ideal gas is considered in a given volume and compressed under *isothermal* ($T = \text{const.}$) conditions, the mean translational kinetic energy of the gas is not changed. However, its entropy is reduced, because fewer possible positions are available for the gas atoms in the new volume. Conservation of the total energy leads to a dissipation of thermal energy

$$\Delta W = T \Delta S$$

to the outside thermal reservoir.

In information theory the situation is similar, the aim being to measure information [9]. The number Ω of possible configurations of a binary code with m elements is

$$\Omega = 2^m.$$

A stored bit can be in one of two states. If these states have the same probability (i.e. $\Omega = 2$), the minimum entropy associated with this bit is

$$S = k_B \ln\Omega = k_B \ln 2.$$

In case that the number of bits m is reduced during the computation process, for example by logic gates with less output than input bits, the energy dissipation per lost bit is given by

$$\Delta W = T\Delta S = k_B T \ln 2.$$

This is the so-called *Landauer principle* [10], which states that erasure is not thermodynamically reversible. Any loss of information inherently leads to a (minimum) energy dissipation of this amount per reduced bit in the case of an isothermal operation, as the entropy of the systems changes. In contrast, to prevent the thermal energy destroying the stored information, the minimum power for storage of information is $k_B T \ln 2$. The *energy consumption* in computation is closely linked to the *reversibility* of the computation.

13.2.2 Irreversible Computation

In all switching elements with more input than output bits, a loss of information occurs upon the information processing. For example, the Boolean function AND is defined as

$$(0,0) \to 0; \quad (0,1) \to 0; \quad (1,0) \to 0; \quad (1,1) \to 1\,.$$

The 0 output of this conventional two-valued gate can be caused by three possible input signal configurations. This type of computing is called *irreversible*.

13.2.3 Reversible Computation

Reversible computation needs to be based on switching elements which do not lose information. An example is the NOT gate, which is a single bit in- and output

$$(0) \to (1); \quad (1) \to (0)$$

or the controlled NOT (CNOT) gate, which makes the XOR operation reversible $(x, y) \to (x, x\ XOR\ y)$:

$$(0,0) \to (0,0); \quad (0,1) \to (0,1); \quad (1,0) \to (1,1); \quad (1,1) \to (1,0)\,.$$

The input can always be deduced from the output.

Now, this should be considered from the physical point of view. The classical two-state system is prepared and stored in the two stable states 0 or 1, respectively. It can be characterized by a particle placed in a double-well potential (see Figure 13.3a). To start the controlled reversible switching of a bit, some definite energy must be fed into the system to overcome the energy barrier separating the two minima of the double well potential. The energy is then available to perform a switching in an adjacent minimum of the potential, and remains available for subsequent switching processes

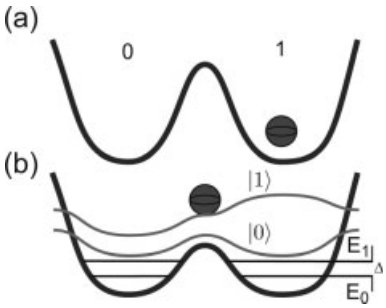


Figure 13.3 Representation of information in (a) a classical computer by a bit with two states (0, 1) and (b) a quantum computer by a quantum-mechanical two-level systems. $|0\rangle$, $|1\rangle$ denote the two quantum states.

or to switch the system back into its original state. Thus, both forward- and backward-switching processes have the same probabilities and the net speed of the computational process is zero. The flow of information must be determined by the gates and an adequate feedback prevention must be built into the logic gates.

13.2.4
Information Carriers

In classical information-processing systems the information transfer is based on the flow of particles, that is, electronic charges. A huge number of information carriers is involved for fault-tolerant operations. The information processing itself can be done by either reversible or irreversible logic gates.

A pure reversible information transfer could be implemented by means of discrete stationary states of a microscopic system which interacts by fields to move the information and to set the logic states. For example, the Quantum Cellular Automata (QCA) (see Chapter 12) are based on elementary cells with two stable states (0 or 1), which can be toggled by fields emerging from neighboring cells. There is no flow of charges or particles, as single atomistic entities (electrons, electron spin, small molecules) solely change their position in a potential well. In principle, an ideal QCA circuit would operate in the thermodynamic limit of information processing and at the same time is still calculating with conventional Boolean logic.

13.3
Quantum Mechanics and Qubits

In quantum computation the information is represented by the quantum properties of particles, so-called *qubits*, and its operations are devised and built by quantum mechanisms. The information content of a single qubit is obtained by a measurement process. An ideal quantum mechanical measurement of a single qubit can only measure one degree of freedom, and returns either 0 or 1. The information encoded

in a quantum mechanical system is described by a vector in the Hilbert space. The components of the Hilbert space vector denote probability amplitudes related to the outcome of certain measurements. Physical pure states are unit-norm vectors in the Hilbert space. Any observable of a system – that is, any quantity which can be measured in a physical experiment – should be associated with a self-adjoint linear operator. The measurable values are the eigenvalues of this operator and their probability is related to the projection of the physical state on the respective subspace. To keep the norm of the physical state fixed the operator should be unitary; that is, all eigenvalues of the operator matrix are complex numbers having absolute value 1.

Shortly, the aim of this subchapter is to summarize the important features of quantum mechanics. A good introduction can be found in various textbooks, for example in Ref. [11].

- *Quantization of states:* The observable quantities do not vary continuously but come in discrete *quanta*. In real systems they can be represented by electric charge, spin, magnetic flux, phase of electromagnetic wave, chemical structure, and so on.
- *Superposition of states:* The linear combination of two or more eigenstates results in a quantum superposition. If the quantity is measured, the state will randomly collapse onto one of the values in the superposition with a probability proportional to the square of the amplitude of that eigenstate in the linear combination. This superposition of states makes quantum computation qualitatively more powerful than classical one, because if we have $\alpha|0\rangle + \beta|1\rangle$ we simultaneously make computations with $|0\rangle$ and $|1\rangle$, roughly speaking.
- *Entanglement of states:* Two spatially separated and non-interacting quantum systems (qubits) that have been interacting may have some locally inaccessible information in common. An entangled state cannot be written as a direct product of two states from two subsystem Hilbert spaces. The entanglement of states makes quantum cryptography possible.

13.3.1 Bit versus Qubit

In a classical computer the information is encoded in a sequence of bits, having two distinguishable and stable states, for example, as a particle placed in a double-well potential, which are conventionally read as either a 0 or a 1 (Figure 13.3a). Apart from some similarities with a classical bit the qubit is overall very different. As in case of the bit, two distinguishable states – that is, different eigenstates of an operator – are needed. For example, a spin $\frac{1}{2}$ particle has two possible states ($\uparrow$ or $\downarrow$), or a photon can be polarized either vertically or horizontally. The state of a qubit may be expressed by basis states (or vectors) of the Hilbert space. The Dirac, or the so-called bra-ket, notation is used to represent them. This means that the two logical basis states, so-called eigenstates, are conventionally written as $|0\rangle$ and $|1\rangle$ (pronounced: "ket 0" and "ket 1").

Unlike the bit (being either 0 or 1) the qubit is not necessarily in the $|0\rangle$ or $|1\rangle$ state, but it can be rather in a superposition of both states. A particle representing a qubit is

described by a quantum-mechanical wave function and can tunnel under the barrier which separates two wells (i.e. the two states). As a consequence, a quantum system in a double-well potential has the two lowest energy states $|0\rangle = \frac{1}{\sqrt{2}}(|\uparrow\rangle + |\downarrow\rangle)$ and $|1\rangle = \frac{1}{\sqrt{2}}(|\uparrow\rangle - |\downarrow\rangle)$. Here $\uparrow$ and $\downarrow$ are two classical states, like the states 0 and 1 in the double-well potential (Figure 13.3a). In the ground state $|0\rangle$ the wave function is symmetric, whereas for the first excited state $|1\rangle$ it is antisymmetric, Figure 13.3b). A system prepared in a superposition state exhibits coherent oscillations between the two wells, and the measuring probability for finding the particle in each well oscillates with frequency $\omega = \Delta\,\hbar^{-1}$, where Δ is the splitting of the lowest energy level (see Figure 13.3b).

13.3.2 Qubit States

A pure qubit state is a linear superposition of both eigenstates. This means that the qubit can be represented as a linear combination of $|0\rangle$ and $|1\rangle$:

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle, \tag{13.1}$$

where α and β are probability amplitudes and can, in general, be complex.

When measuring this qubit in the standard eigen-basis, the probability of outcome $|0\rangle$ is $|\alpha|^2$ and of $|1\rangle$ is $|\beta|^2$. Because the absolute squares of the amplitudes are equal to probabilities, α and β must be constrained by the equation $|\alpha|^2 + |\beta|^2 = 1$. The Eq. (13.1) could be rewritten as

$$|\psi\rangle = |\psi(\varphi,\theta)\rangle = e^{-i\varphi/2}\cos\frac{\theta}{2}|0\rangle + e^{i\varphi/2}\sin\frac{\theta}{2}|1\rangle.$$

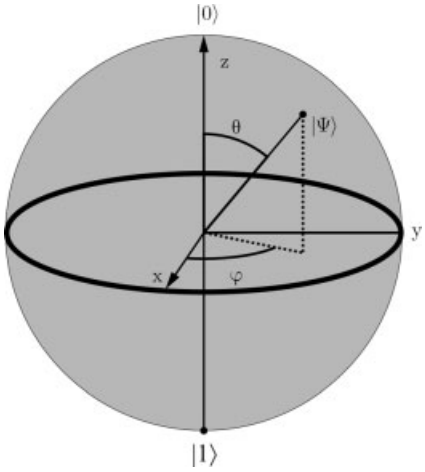


Figure 13.4 The qubit $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$ is represented as point (θ,φ) on a unit sphere, the so-called Bloch sphere. θ and φ are defined by $\alpha = e^{-i\varphi/2}\cos(\theta/2)$ and $\beta = e^{i\varphi}\sin(\theta/2)$. $|\psi\rangle$ is represented by the unit vector $[\cos(\varphi)\sin(\theta), \sin(\varphi)\sin(\theta), \cos(\theta)]$, called the Bloch vector.

The common phase factor resulting from the complex nature of α and β was neglected. The state of a single qubit can be represented geometrically by a point on the surface of the Bloch sphere (see Figure 13.4). A single qubit has two degrees of freedom φ, θ. A classical bit can only be represented by two discrete values 0 or 1. Note that the two complex numbers α, β in Eq. 13.1 in fact correspond to four numbers: Re(α), Im(α), Re(β), Im(β). However, these numbers are not independent; they are linked by the unity norm and the physical irrelevant common phase factor can be neglected. Any two-level quantum physical system can be used as a qubit; for example, the electron charge, polarization of photons, the spin of electrons or atoms and the charge, flux or phase of Josephson junctions could be used for implementation of qubits.

13.3.3 Entanglement

The entanglement of qubits is a subtle non-local correlation that has no classical analog. It allows a set of qubits to express a higher correlation than is possible in classical systems. As an example, consider the two entangled qubits A and B

$$|\Phi\rangle = |\phi\rangle_A|\psi\rangle_B = |\phi_A\psi_B\rangle = |\phi\psi\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle).$$

The first system is in state $|\phi\rangle_A$ and the second in state $|\psi\rangle_B$. Both systems have the two basis vectors $|0\rangle$ and $|1\rangle$. When measuring state $|\Phi\rangle$ the outcomes $|00\rangle$ and $|11\rangle$ have equal probabilities. It is impossible to attribute to either system A or system B a definite pure state as their states are superposed with one another. It is seen that if a 0 state is measured in A, then there will be an obligatory 0 state when measuring B. So A and B are not independent, they are *entangled*.

Entanglement allows multiple states to be acted on simultaneously, unlike classical bits that can only have one value at a time.

A number of entangled qubits taken together is a *qubit register*, with basis states of the form $|x_1x_2\ldots x_n\rangle$. An n-qubit register has a 2^n dimensional space, being much larger than a classical n-bit register.

Entanglement of states in quantum computing has been referred to as *quantum parallelism*, as the state can be in a quantum superposition of many different classical computational paths which can all proceed concurrently.

13.3.4 Physical State

The quantum Hamiltonian operator $\hat{H}$ generates the time evolution of quantum states and applied to the state vector yields the observable corresponding to the total energy of the system. The eigenvectors of $\hat{H}$, denoted by $|x\rangle$, provide an orthonormal basis for the Hilbert space of the system. The equation

$$\hat{H}|x\rangle = E_x|x\rangle.$$

yields the spectrum of allowed energy levels of the system, given by the set of eigenvalues E_x. Since $\hat{H}$ is a Hermitian operator, the energy is always a real number.

The time evolution $|\psi(t)\rangle$ of quantum states is given by the time-dependent Schrödinger equation:

$$\hat{H}\,|\psi(t)\rangle = \mathrm{i}\hbar\frac{\partial}{\partial t}|\psi(t)\rangle.$$

If $\hat{H}$ is independent of time this equation can be integrated to obtain the state at any time:

$$|\psi(t)\rangle = \exp\left(-\frac{\mathrm{i}\hat{H}\,t}{\hbar}\right)|\psi(0)\rangle,$$

where $|\psi(0)\rangle$ is the state at some initial time ($t = 0$).

13.3.4.1 Measurement

A measurement of a quantum state inevitably alters the system, as it projects the state onto the basis states of the measuring operator. Only if the state is already the eigenstate of the measuring operator then the state does not change. Thus, the superposition of states collapse into one or the other eigenstate of the operator, defined by the probabilities amplitudes. The precise amplitudes (α and β of a single qubit) can be found by multiple recreation of the superposition and subsequent measurements.

13.3.4.2 No-Cloning Theorem

Unlike the classical bit, of which multiple copies can be made, it is not possible to make a copy (clone) of an unknown quantum state [12, 13]. This so-called *no-cloning theorem* has profound implications for error correction of quantum computing as well as quantum cryptography. For example, such as it prevents the use of classical error correction techniques on quantum states, no backup copy of a state to correct subsequent errors could be made. However, error correction in quantum computation is still possible (see Section 13.5 for details). The no-cloning theorem protects the uncertainty principle in quantum mechanics, as the availability of several copies of an unknown system, on which each dynamical variable could be measured separately with arbitrary precision would bypass the Heisenberg uncertainty principle $\Delta x \Delta p$, $\Delta E \Delta t \geq \hbar/2$.

13.4 Operation Scheme

Quantum bits must be coupled and controlled by gates in order to process the information. At the same time, they must be completely decoupled from external influences such as thermal noise. It is only during well-defined periods that the control, write and readout operations take place to prevent an untimely readout.

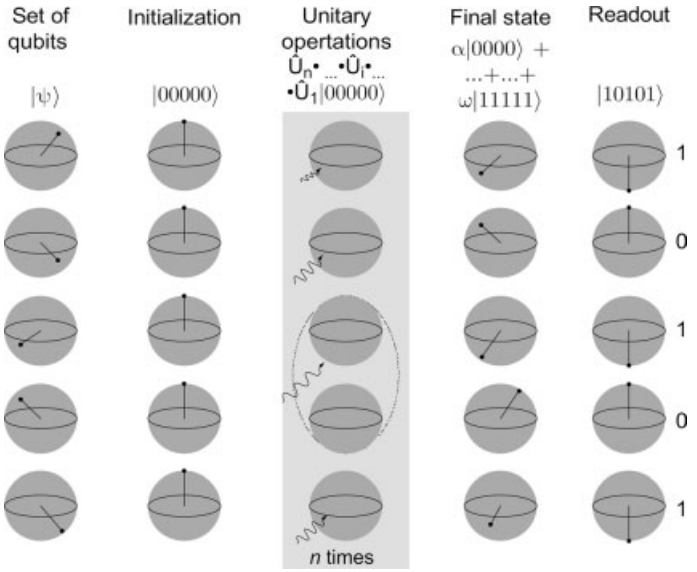


Figure 13.5 Operation scheme of quantum computation. After the system is prepared as a quantum register, controlled unitary operations $\hat{U}$ on single or entangled qubits are performed by gate operations in a controlled manner $\hat{U} = \exp(-i\hat{H}t/\hbar)$ ($\hat{H}$: Hamiltonian, t: time). The readout is done by projection onto basis states, yielding probability distributed Boolean values.

13.4.1
Quantum Algorithms: Initialization, Execution and Termination

The operation scheme of quantum computation is depicted in Figure 13.5. To start an quantum algorithm the qubit register must be initialized in some specified well-defined state, for example by a dissipative process to the ground state $|00\ldots0\rangle$.

Then, the computation is done by an appropriate sequence of applied unitary operations $\hat{U} = \exp(-i\hat{H}t/\hbar)$ with Hamiltonian $\hat{H}$ to the qubits. The actual mechanism – that is, electromagnetic waves, voltages or magnetic fields of well-defined energy/amplitude and time t – depends strongly on the type of qubit. In each step of the algorithm, the qubit vector is modified by multiplying it by a unitary matrix. The components of the matrix are determined by the physics of the device. The unitary character of the matrix ensures that the matrix is invertible, making the computation reversible.

For a given algorithm the operations will always be done in exactly the same order. Since there is no way to read the qubit state before the final destructive readout measurement, there are no conditional statements such as 'IF... THEN...'. However, there are conditional gate operations such as the controlled NOT (CNOT) gate.

After termination of the algorithm the qubit vector must be read out by measurement. Quantum mechanics ensure that the measurement will destroy the qubit

vector by projection onto the eigenstate of the corresponding observable, and only a probability distributed n-bit vector is obtained.

Even when neglecting the decoherence sources during the unitary transformations, the experimental readout schemes can never be perfectly efficient. Thus, it should be possible to repeat the measurement to enhance the probability of the obtained results by a majority polled output.

13.4.2 Quantum Gates

Once a quantum register is initialized the qubits must be manipulated in order to process the information by quantum gates, just as in case of the classical logic gates for conventional digital information processing. Quantum logic gates are represented by unitary matrices, as they are reversible, unlike many classical logic gates. However, some universal classical logic gates, such as the *Toffoli* gate, also provide reversibility, and can be directly mapped onto quantum logic gates. Mostly quantum gates operate on spaces of one or two qubits, thus written as matrices the quantum gates can be described by 2×2 or 4×4 matrices with orthonormal rows.

Examples of a single qubit operation are the *Hadamard* gate, which puts the initialized qubit in superposition state, represented by the Hadamard matrix

$$\hat{H}_{\text{Hadamard}} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix},$$

and the CNOT gate for two qubit operations, defined as

$$\hat{H}_{\text{CNOT}} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{X} \end{bmatrix},$$

with $\mathbf{1}$ the identity matrix and $\mathbf{X}$ the first Pauli matrix.

Both, in classical and quantum computation, all possible operations can be reduced to a finite set of *universal* gates, which can be used to construct the specific algorithm of the information processing. To achieve universality for classical reversible gates, three-bit operations are needed, whereas in the quantum regime only one- and two-qubit gate operations are sufficient. This underlines the versatile character of quantum logic.

13.5 Quantum Decoherence and Error Correction

Decoherence is the mechanism by which the information encoded in the superposed and entangled qubits register degrades over time. For example, dephasing caused by fluctuations of the energy level of two quantum mechanical states gives an additional phase proportional to the energy change of the superposition states. With increasing number of qubits the computation power, as well as the probability for decoherence,

will increase, and the need for decoherence control becomes predominant. This could be done by the implementing of quantum error correcting gates. In general, the sources of error can be: (i) non-ideal gate operations; (ii) interaction with environment causing relaxation or decay of phase coherence; and (iii) deviations of the quantum system from an idealized model system.

In classical computers every bit of information can be re-adjusted after every logical step by using non-linear devices to re-set the information bit to the 0 or 1 state. Contrary in a quantum system, no copy can be made of a qubit state without projecting it onto one of its eigenstates and thus destroying the superposition state.

Quantum information processing attracted much attention after Shor's surprising discovery that quantum information stored in a superposition state can be protected against decoherence. The single qubit is encoded in multiple qubits, followed by a measurement yielding the type of error, if any, which happened on the quantum state. With this information the original state is recovered by applying a proper unitary transformation to the system. This stimulated much research on quantum error correction, and led to the demonstration that arbitrarily complicated computations could be performed, provided that the error per operation is reduced below a certain error threshold.

By repeated runs of the quantum algorithm and measurement of the output, the correct value can be determined to a high probability. In brief, quantum computations are probabilistic.

13.6 Qubit Requirements

Di Vincenzo [14] listed criteria that any implementation for quantum computers should fulfill to be considered as *useful*:

- A scalable physical system with well-characterized qubits. A quantum computer consisting of a few qubits is not sufficient for useful computation.
- The ability to initialize the state of all qubits to a simple basis state.
- Relative long decoherence times τ_{dec}, much longer than the gate-operating time. To observe quantum-coherent oscillations the requirement $\tau_{dec}\Delta \gg \hbar$ must be fulfilled, and the fidelity loss per single quantum gate operation should be below some criteria.
- A universal set of quantum gates is needed to control the quantum system.
- A qubit-specific measurement capability to perform a readout and to transfer the information to conventional computers.

13.7 Candidates for Qubits

Various quantum mechanical two-level systems have been examined as potential candidates for qubits. In this overview, the aim is to concentrate on the first system

where the feasibility of quantum computation was demonstrated, and highlight some potential *solid-state* systems. The solid-state technology is easily scalable and can be integrated into conventional electric circuits. At the time of writing [15] there is no notable clear favorite for quantum computing, because on one hand the known systems must be improved, while on the other hand new promising candidates for quantum computing hardware are continuously appearing.

13.7.1
Nuclear Magnetic Resonance (NMR)-Based Qubits

Atomic nuclei are relatively isolated from the environment and thus well protected against decoherence. Their spins can be manipulated by properly chosen radio frequency irradiation. Elementary quantum logic operations have been carried out on the spin $\frac{1}{2}$ nuclei as qubits by using nuclear magnetic resonance (NMR) spectroscopy [16]. This system is restricted by the low polarization and the limited numbers of resolvable qubits. The nuclei (see Figure 13.6) were in molecules in a solution, and a magnetic field defined the two energy-separated states of the nuclei. These macroscopic samples with many nuclear qubits provide massive redundancy, as each molecule serves as a single quantum computer. The qubits interact through the electronic wave function of the molecule. Quantum calculations with up to seven qubits have been demonstrated. For example, the Shor algorithm was implemented to factorize the number 15 [17].

13.7.2
Advantages of Solid-State-Based Qubits

Large number of qubits could be assembled using existing solid-state technology. The main drawback is that the accuracy of the devices is not as high as established NMR-based qubits described above. Solid-state quantum computers encounter a new set of problems, as the spatial inhomogeneity during processing causes differences between nominally identical devices. Digital logic tolerates imperfection by restoring a signal to its intended value, whereas quantum computing forsakes this

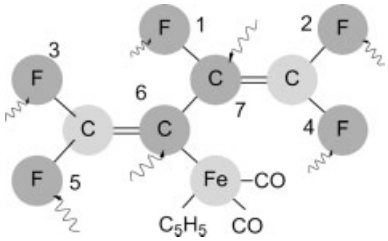


Figure 13.6 One molecule acts as seven-qubit system. The nuclei of five fluorine and two carbon atoms in the molecule form seven nuclear spin qubits (light gray), which are programmed by radiofrequency pulses and can be detected by nuclear magnetic resonance (NMR) spectroscopy [16].

methodology. Until now the imprecision of conventional solid-state devices prevents the storage of more than a few bits in a single device. A solid-state quantum computing device should be cooled down to a few mK to prevent thermal noise, caused for example, by phonons and destroying the coherent superposition of states.

The computations are performed by microwave pulses, that decrease the barrier separating the two states and brings the system into quantum realm. Subsequent pulses manipulate the Hamiltonian, *viz*. the probability weighting of the state vector to perform useful operations.

The spread between qubit parameters could be compensated by individually chosen biases, managed by a conventional computer. However, leaving the reduced coherence time due to additional noise aside, this would complicate the layout of a quantum computer, as at least one additional wire per qubit is needed.

Fortunately, it transpires that quantum error correction can even take care of error caused by defective implementation of a quantum gate, in case that the error is not larger than 10^{-4} per operation [18].

13.7.3
Kane Quantum Computer

In 1998, Kane [19] suggested imbedding a large number of nuclear spin qubits, made up by isotopically pure ^{31}P (nuclear spin 1/2) donor atoms, at a moderate depth in a ^{28}Si (nuclear spin 0) crystal to build a quantum computer (see Figure 13.7). When placed in a magnetic field, two distinguishable states for the spin 1/2 nuclei of the P atoms appear. By a controllable overlap of the wave functions of electrons bound to this atom in the deliberately doped semiconductor, some interaction may take place between adjacent nuclear qubits. Voltages applied to electrodes (gate A) placed above the phosphorus atoms can modify the strength of the electron–nuclear coupling. Individual qubits are addressed by the A-gates to shift their energies in and out of resonance with an external radiofrequency. The strength of the qubit–qubit coupling by overlapping wavefunctions is controlled by electrodes placed midway between the P atoms (gate B). The operation principle is similar to the NMR-based qubits in a macroscopic molecule solution, except that the nuclei are addressed by potentials rather than by the radiofrequency pulses.

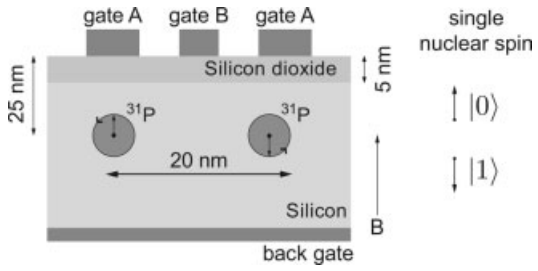


Figure 13.7 The Kane quantum computer: ^{31}P in a ^{28}Si matrix are controlled by gate electrodes (A, B) placed above them [19].

13.7.4
Quantum Dot

Beside the nucleus, the spin and charge of electrons may also be used to construct a double degenerate system. The electron spin-based qubits have the advantage that the Hilbert space consists of only two spin states, which strongly suppresses the leakage of quantum information into other states. In addition, the spin is less coupled to the environment than the charge, which results in longer decoherence times.

Quantum dots of 5 to 50 nm are thin, semiconducting multilayers on the substrate surface, where confined electrons with discrete energy spectra appear. By using Group III–V compound semiconductor materials such as GaAs and InAs of different lattice sizes, two-dimensional electron gas systems with high electron mobility can be constructed. To use these as qubits, their quantized electron number or spin is utilized [20]. The switching of the quantum state can be achieved either by optical or electrical means. In case of the spin-based quantum dot the electrons are localized in electrostatically defined dots (see Figure 13.8). The coupling between the electron spins is made via the exchange interaction, and the electron tunneling between the dots is controlled by a gate-voltage (gate B in Figure 13.8).

13.7.5
Superconducting Qubits

Superconductivity is a macroscopic quantum mechanical state, in which electrons with opposite spin and momenta form the so-called Cooper pairs and an energy gap in the quasi-particle spectrum appears. The interaction between the electrons is mediated by phonons. The superconducting qubits are based on *Josephson junctions* – that is, two superconductors separated by a tunnel barrier. The supercurrent that

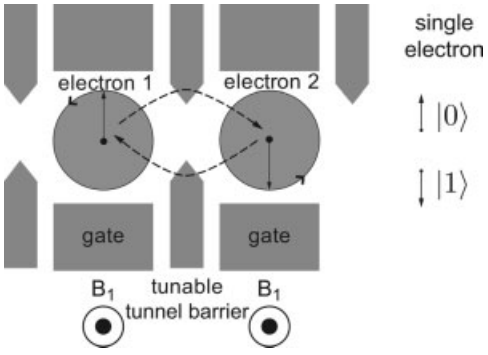


Figure 13.8 Scheme of spin-based quantum dots as a qubit system. The coupling of spins of localized electrons is formed by exchange interaction due to gate-voltage-controlled tunneling. Two electrons are localized in the regions defined by the gates. Single spin operations are performed by local magnetic fields [20].

crosses the weak link is the Josephson supercurrent. For details on superconductivity and the Josephson effect, see Refs. [21, 22].

Until now, several possible systems differing by 2e have been described for constructing a superconducting qubit. In the *charge* qubit a coherent state with a well-defined charge of individual Cooper pairs is used, while the *flux* qubit employs two degenerate magnetic flux states and the *phase* qubit is based on the phase difference of superconducting wavefunctions in two electrodes for quantum computation [23, 24].

13.7.5.1 Charge Qubits

The basis states of a charge qubit are two charge states. The qubit is formed by a tiny ~100-nm superconducting island – a Cooper pair box. Thus the charging electrostatic energy of a Cooper pair dominates in comparison with all other energies. An external gate voltage controls the charge, and the operation can be performed by controlling the applied gate voltages V_g and magnetic fields. A superconducting reservoir, coupled by a Josephson junction to the Cooper pair box, supplies the neutral reference charge level (see Figure 13.9). The state of the qubit is determined by the number of Cooper pairs which have tunneled across the junction. The readout is performed by a single-electron transistor attached to the island (not shown). By applying a voltage to this transistor, the dissipative current through the probe destroys the coherence of charge states and its charge can be measured [25].

13.7.5.2 Flux Qubits

A flux qubit is formed by a superconducting loop (1 μm size) interrupted by several Josephson junctions with well-chosen parameters (c.f. Figure 13.10) [26]. To obtain a double-well potential, as in Figure 13.10, either an external flux $\Phi_0/2$ or a π junction is needed. By including a π junction [27–29] in the loop the persistent current,

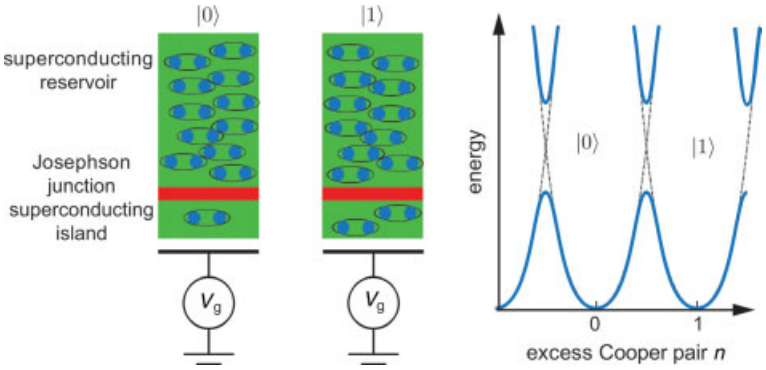


Figure 13.9 The charge qubit is formed by a Cooper pair box (CPB), separated from the superconducting reservoir (top) by a Josephson tunnel junction [25]. The basis states $|0\rangle$ and $|1\rangle$ differ by the number of excess Cooper pairs *n* on the CPB.

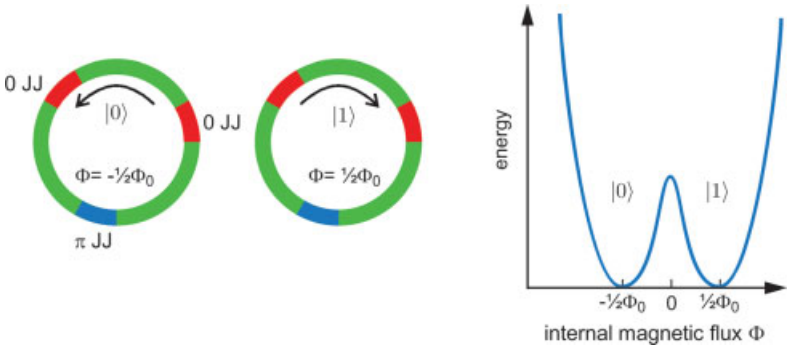


Figure 13.10 The basis states $|0\rangle$ and $|1\rangle$ of the flux qubit are determined by the direction of a persistent current circulating in three-junctions qubit [26]. The basis states $|0\rangle$ and $|1\rangle$ differ by the direction of the current in the superconducting loop. The π-Josephson junction (JJ) self-biases the qubit to the working point and, thus, substitutes an external magnetic flux.

generating the magnetic flux, may spontaneously appear and flow continuously, even in absence of an applied magnetic field [30]. The basis states of the qubit are defined by the direction of the circulating current (clockwise or counter-clockwise). The currents screen the intrinsic phase shift of π of the loop, such that the total flux through the loop is equal to $\pm\Phi_0/2$, i.e. half a magnetic flux quanta. The two energy levels corresponding to the two directions of circulating supercurrent are degenerate. If the system is in the quantum mechanical regime (low temperature to suppress thermal contributions) and the coupling between the two states (clockwise/counter-clockwise current flow) is strong enough (*viz*. the barriers are low), the system can be in the superposition of clockwise and counter clockwise states. This *quiet* qubit [31], is expected to be robust to the decoherence by the environment because it is self-biased by a π Josephson junction. Note that this flux-qubit device with a π junction is an optimization of the earlier scheme, where the phase shift of π was generated by an individual outer magnetic field of $\pm\Phi_0/2$ for each qubit [26].

The readout could by made by an additional superconducting loop with one or two Josephson junctions (i.e. SQUID loop) that is inductively coupled to the qubit.

To process the input and output of flux qubits an interface hardware based on the rapid single flux quantum (RSFQ) circuits could be used. These well-developed superconducting digital logics work by manipulating and transmitting a single flux quanta. In fact, this logic overcomes many problems of conventional CMOS-logics as it has a very low power consumption, an operating frequency of several hundred GHz, and is compatible with the flux qubit technology [32].

13.7.5.3 **Fractional Flux Qubits**

One variation of a flux π qubit is a fractional flux qubit. At the boundary between a 0 and a π coupled Josephson junction (i.e. a 0–π JJ) a *spontaneous* vortex of supercurrent

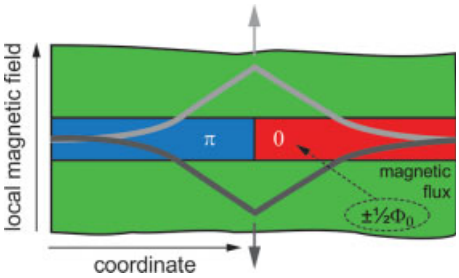


Figure 13.11 Sketch of a 0–π JJ with circulating supercurrent around 0–π phase boundary. The magnetic flux is equal to half of a flux quantum Φ_0 (semifluxon).

may appear under certain circumstances. Depending on the length L of the junction, the supercurrent carries a half-integer flux quantum $\Phi_0/2$ (called *semifluxon*) or fractions of it. Figure 13.11 depicts the cross-section of a symmetric 0–π *long* JJ. Classically, the semifluxon has a degenerate ground state of either positive or negative polarity, that corresponds to clockwise and counter-clockwise circulation of supercurrent around the 0–π boundary. The magnetic flux localized at the 0–π boundary is $\Phi_0/2$ and represents two degenerate classical states [33].

0–π Josephson junctions with a spontaneous flux in the ground state are realized with various technologies. The presence of spontaneous flux has been demonstrated experimentally in *d*-wave superconductor-based ramp zigzag junctions [34], in long Josephson 0–π junctions fabricated using the conventional Nb/Al-Al_2O_3/Nb technology with a pair of current injectors [35], in the so-called tricrystal grain-boundary long junctions [36–38] or in SIFS Josephson junctions [39] with a *step-like* ferromagnetic barrier. In the latter systems the Josephson phase is set to 0 or π by choosing a proper F-layer thicknesses d_F. The advantages of this system are that it can be prepared in a multilayer geometry (allowing topological flexibility) and it can be easily combined with the well-developed Nb/Al-Al_2O_3/Nb technology.

A *single* semifluxon ground state is double degenerate with basis flux states $|\uparrow\rangle, |\downarrow\rangle$. It transpires that the energy barrier scales proportionally to the junction length L, and the probability of tunneling between $|\uparrow\rangle$ and $|\downarrow\rangle$ decreases exponentially for increasing L [40]. Hence, a single semifluxon will always be in the classical regime with thermal activated tunneling for long junctions. As a modification, a junction of finite, rather small length L may be considered. In this case, the barrier height is finite and approaches zero when the junction length $L \to 0$. At this limit, the situation is not really a semifluxon, as the flux Φ present in the junction is much smaller than $\Phi_0/2$.

A 0–π–0 Josephson junction (see Figure 13.12) has *two* antiparallel coupled semifluxons for a distance a larger than the critical distance a_c [40]. The ground state of this system is either $|\uparrow\downarrow\rangle$ or $|\downarrow\uparrow\rangle$. For symmetry reasons both states are degenerate. The tunnel barrier can be made rather small, which results in a rather strong coupling with appreciable energy level splitting due to the wave functions overlap. Estimations show that this system can be mapped to a particle in a double-well potential, and thus can be used as a qubit like other Josephson junctions-based

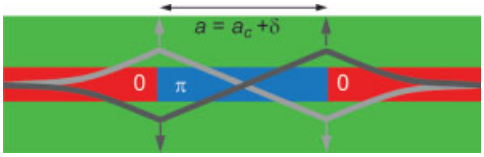


Figure 13.12 The two basis states are $|\uparrow\downarrow\rangle$ and $|\downarrow\uparrow\rangle$ of two coupled fractional vortices in a long linear 0–π–0 Josephson junction.

qubits. Thus, the 0–π–0 junctions are supposed to show the motion of a point-like particle in a double-well potential, and may be used as the basis cell of a fractional flux qubit.

13.8 Perspectives

Today, quantum computer algorithms allow to solve some specific **NP** problems, including factoring, sorting, calculating discrete logarithms, and simulating quantum physics. However, conventional computers will still be needed for addressing the quantum computer, for databases and for applications where either high computation power is not needed or sufficient coherence could not be provided, as in mobile devices. Quantum computers could be used as coprocessors for specific tasks such as en/decryption. The use of quantum computers may have several consequence, as today's data security algorithms could stay secure only if the keylength exceeds the storage capacity of quantum computers. The simulations of quantum mechanics might contribute to a variety of scientific and practical applications based on physics, chemistry, biology, medicine and related fields.

References

1 R. P. Feynman, *Int. J. Theoret. Physics* 1982, **21**, 467.

2 R. P. Feynman, *Feynman Lectures on Computation*, Addison-Wesley, Reading, MA, 1996.

3 International Technology Roadmap for Semiconductors, http://www.itrs.net/.

4 M. A. Nielsen, I. L. Chuang, *Quantum Computation and Quantum Information*, University Press, Cambridge, 2000.

5 J. Stolze, D. Suter, *Quantum Computation*, Wiley-VCH, 2004.

6 L. K. Grover, Proceedings 28th Annual ACM Symposium on the Theory of Computing, Philadelphia, Pennsylvania, ACM Publishers, New York, 1996.

7 P. W. Shor, *Proceedings, 35th Annual Symposium on the Foundations of Computer Science*, IEEE Computer Society Press, Los Alamitos, CA, 1994.

8 A. Einstein, B. Podolsky, N. Rosen, *Phys. Rev.* 1935, **47**, 777.

9 R. W. Hamming, *Coding and Information Theory*, Prentice-Hall, 1986.

10 R. Landauer, *IBM J. Res. Dev.* 1961, **5**, 183.

11 C. Cohen-Tannoudji, B. Diu, F. Laloë, *Quantum Mechanics*, Wiley-VCH, 1977.

12 W. K. Wootters, W. H. Zurek, *Nature* 1982, **299**, 802.

13 D. Dieks, *Phys. Lett. A* 1982, **92**, 271.

14 D. P. DiVincenzo, *Fortschr. Phys. Prog. Physics* 2000, **48**, 771.

15 GDEST , *EU/US Workshop on Quantum Information and Coherence*, December 8–9, Munich, Germany, 2005.

16 L. M. K. Vandersypen, I. L. Chuang, *Rev. Mod. Phys.* 2004, **76**, 1037.

17 L. M. K. Vandersypen, M. Steffen, G. Breyta, C. S. Yannoni, M. H. Sherwood, I. L. Chuang, *Nature* 2001, **414**, 883.

18 R. W. Keyes, *Appl. Phys. A* 2003, **76**, 737.

19 B. E. Kane, *Nature* 1998, **393**, 133.

20 D. Loss, D. P. DiVincenzo, *Phys. Rev. A* 1998, **57**, 120.

21 W. Buckel, R. Kleiner, *Superconductivity. Fundamentals and Applications*, Wiley-VCH, 2004.

22 A. Barone, G. Paterno, *Physics and Applications of the Josephson Effect*, John Wiley & Sons, 1982.

23 A. Ustinov, in: R. Waser (Ed.), *Nanoelectronics and Information Technology - Advanced Electronic Materials and Novel Devices*, Wiley-VCH, 2005.

24 Y. Makhlin, G. Schön, A. Shnirman, *Rev. Mod. Phys.* 2001, **73**, 357.

25 Y. Nakamura, Y. A. Pashkin, J. S. Tsai, *Nature* 1999, **398**, 786.

26 J. E. Mooji, T. P. Orlando, L. Levitov, L. Tian, C. H. van der Wal, S. Lloyd, *Science* 1999, **285**, 1036.

27 L. Bulaevskii, V. Kuzii, A. Sobyanin, *J. Exp. Theoret. Physics Lett.* 1977, **25**, 290.

28 V. V. Ryazanov, V. A. Oboznov, A. Y. Rusanov, A. V. Veretennikov, A. A. Golubov, J. Aarts, *Phys. Rev. Lett.* 2001, **86**, 2427.

29 M. Weides, M. Kemmler, E. Goldobin, D. Koelle, R. Kleiner, H. Kohlstedt, A. Buzdin, *Appl. Phys. Lett.* 2006, **89**, 122511.

30 T. Yamashita, S. Takahashi, S. Maekawa, *Appl. Phys. Lett.* 2006, **88**, 132501.

31 L. B. Ioffe, V. B. Geshkenbein, M. V. Feigel'man, A. L. Fauchère, G. Blatter, *Nature* 1999, **398**, 679.

32 M. Siegel, in: R. Waser (Ed.), *Nanoelectronics and Information Technology - Advanced Electronic Materials and Novel Devices*, Wiley-VCH, 2005.

33 E. Goldobin, D. Koelle, R. Kleiner, *Phys. Rev. B* 2002, **66**, 100508.

34 H. Hilgenkamp, A. Ariando, H. J. H. Smilde, D. H. A. Blank, G. Rijnders, H. Rogalla, J. R. Kirtley, C. C. Tsuei, *Nature* 2003, **422**, 50.

35 E. Goldobin, A. Sterck, T. Gaber, D. Koelle, R. Kleiner, *Phys. Rev. Lett.* 2004, **92**, 57005.

36 J. R. Kirtley, C. C. Tsuei, K. A. Moler, *Science* 1999, **285**, 1373.

37 J. R. Kirtley, C. C. Tsuei, M. Rupp, J. Z. Sun, L. S. Yu-Jahnes, A. Gupta, M. B. Ketchen, K. A. Moler, M. Bhushan, *Phys. Rev. Lett.* 1996, **76**, 1336.

38 A. Sugimoto, T. Yamaguchi, I. Iguchi, *Physica C* 2002, **367**, 28.

39 M. Weides, M. Kemmler, H. Kohlstedt, R. Waser, D. Koelle, R. Kleiner, E. Goldobin, *Phys. Rev. Lett.* 2006, **97**, 247001.

40 E. Goldobin, K. Vogel, O. Crasser, R. Walser, W. P. Schleich, D. Koelle, R. Kleiner, *Phys. Rev. B* 2005, **72**, 054527.

Part One:
Nanomedicine: The Next Waves of Medical Innovations

Nanotechnology, Volume 5: Nanomedicine. Edited by Viola Vogel

ISBN: 978-3-527-31736-3

1
Introduction

Viola Vogel

1.1
Great Hopes and Expectations are Colliding with Wild Hype and Some Fantasies

What is nanomedicine? Will nanomedicine indeed help to cure major diseases and live up to the great hopes and expectations? What innovations are on the horizon and how can sound predictions be distinguished from wild hype and plain fantasy? What are realistic timescales in which the public might benefit from their ongoing investments?

When first exploring whether nanotechnology might reshape the future ways of diagnosing and treating diseases, the National Institutes of Health stated in the report of their very first nanoscience and nanotechnology workshop in 2000 (http://www.becon.nih.gov/nanotechsympreport.pdf. Bioengineering Consortium):

> *Every once in a while, a new field of science and technology emerges that enables the development of a new generation of scientific and technological approaches. Nanotechnology holds such promise.*

Our macroscopic bodies and tissues are highly structured at smaller and smaller length scales, with each length scale having its own secrets as to how life-supporting tasks are mastered. While we can still touch and feel our organs, they are all composed of cells which are a little less than one million times smaller and only visible under the light microscope (microscopic). Zooming further into the cell, about one thousand times, we find the nanoscale molecular machineries that drive and control the cellular biochemistry, and thereby distinguish living systems from dead matter. Faced with a rest of new technologies that has enabled researchers to visualize and manipulate atoms and molecules, as well as to engineer new materials and devices at this tiny length scale [1], major think tanks have begun since the late 1990's to evaluate the future potential of nanotechnology [2], and later at the interface to medicine [3–11]. These efforts were paralleled by a rapid worldwide increase in funding and research activities since 2000. The offset of a gold rush into the 'nano', by which the world of the very small is currently discovered, will surely also lead to splendid new entrepreneurial opportunities. Progress impacting on human health came much faster than expected.

Nanotechnology, Volume 5: Nanomedicine. Edited by Viola Vogel

ISBN: 978-3-527-31736-3

1.2
The First Medical Applications are Coming to the Patients' Bedside

The public most commonly associates nanomedicine with engineered nanoparticles in the context of drug delivery devices or advanced medical imagine applications. Novel is that the molecules which are coassembled into nanoparticles can nowadays carry many different chemical functionalities. It has thereby become feasible to integrate multiple tasks into drug delivery device, from targeting specific tissues to releasing drugs, from enhancing contrast to probing their environment. How is this all done? First of all, the nanoparticles are loaded with drugs. The particles might then carry molecules on their surfaces that bind with great specificity to complementary molecules that are unique to cancers or to other diseased tissues, as reviewed later in this volume [12, 13]. For example, by using antibody– antigen recognition such engineered nanoparticles can be accumulated in the targeted tissues rather than being distributed over the entire body. Local accumulation greatly enhances the efficiency of drugs and reduces any unintended adverse side effects that might otherwise harm other organs. The selectivity by which disease can be treated by using engineered nanoparticles is thus in stark contrast to how conventional drugs operate; as conventional drugs lack the capacity to target specific tissues, they are distributed much more uniformly over the entire body, and must therefore be administered at much higher doses. Beyond tissue targeting, the nanoparticles might further be engineered to absorb therapeutic radiation, which might heat them up when they have reached the diseased tissues to damage the local tissue, either by heat or by the release of drugs [11, 12, 14]. Alternatively, the nanoparticles might hold on to drugs by bonds that can be locally cleaved by enzymes or be broken by light or radiation, thereby releasing the drug under the control of a physician (as reviewed elsewhere in this volume [12, 13]). The goal here is to design new strategies to inflict damage only to the aberrant cells, while leaving the surrounding tissues unharmed. This multifunctional integration of many different diagnostic and therapeutic tasks in single particles thereby enables applications that go far beyond those of conventional drugs.

Engineered nanoparticles can also change the future of medical imaging, as they enable us to combine structural imaging with spatially resolved diagnostics and interventions. Only eight years after wondering whether nanotechnology will revolutionize medicine, the US National Cancer Institute (NCI) stated that [15]:

> *Nanodevices are used in detecting cancer at its earliest stages, pinpointing its location within the body, delivering anticancer drugs specifically to malignant cells, and determining if these drugs are killing malignant cells.*

Increasingly sophisticated medical imaging technologies continue to revolutionize medicine. X-ray imaging, which was later complemented by ultrasound, positron emission tomography (PET) and magnetic resonance imaging (MRI), opened the possibility to visualize *nonintrusively* first bones, and then the inner organs, of our bodies. The objectives now are to obtain images of the structure of organs, at much

higher resolution, together with spatially resolved biochemical information which is reflective of how well cells and organs function. This includes probing noninvasively whether certain organs produce the hormones and enzymes at normal rates, and whether other metabolic activities might deviate from the norm. Major advances are about to come from the usage of nanoparticles that are engineered to serve both, as drug delivery systems and to enhance the contrast in ultrasound, PET and MRI images (for a review, see Ref. [16] and elsewhere in this volume [12, 13]). Enhancing the contrast and spatial resolution of images will enable the detection of cancers and other structural abnormalities of organs at much earlier stages, which in turn will enhance the chances of an effective therapy. Such multitasking approaches might also one day substitute for a variety of surgical interventions. Today, many books and articles have been published discussing the various medical applications of such engineered nanoparticles, while the pharmaceutical industry continues to invest heavily in their development (for reviews, see Refs [6, 15–22]).

1.3 Major Advances in Medicine Have Always been Driven by New Technologies

During the past few decades, the deciphering of the molecular origins of many diseases has had a most profound impact on improving human health. One historical step was the deciphering of the first protein structure in 1958 [23]. This opened a new era in medicinal chemistry, as drugs could since then be designed in a rational manner – that is, drugs that fit tightly into essential binding pockets thereby regulating protein and DNA functions. The invention of how to harness DNA polymerase in order to amplify genetic material in the test tube – which we now know as the polymerase chain reaction (PCR) [24] – then opened the field of molecular biology during the 1980s. PCR also enabled targeted genetic alterations of cells to identify the functional roles of many proteins, and this in turn led to the discovery that cell signaling pathways of many interacting proteins existed, and could be altered by diseases. The explosion of knowledge into how cell behavior is controlled by biochemical factors opened the door to target drugs to very specific players in cell signaling pathways. This also led to a host of new biotechnology start-up companies, the first of which became profitable only around 2000.

The next major breakthrough came with the solving of the human genome in 2001 [25–29]. Access to a complete genetic inventory of more than 30 000 proteins in our body, combined with high-resolution structures for many of them, enables a far more systematic search for correlations between genetic abnormalities and diseases. Finally, various diseases could for the first time be traced to inherited point mutations of proteins. In achieving this, much insight was gained into the regulatory roles of these proteins in cell signaling and disease development [30]. This includes recognizing genetic predispositions to various cancers [31], as well as to inherited syndromes where larger sets of seemingly uncorrelated symptoms could finally be explained by the malfunctioning of particular proteins or cell signaling pathways [32–38], including ion channel diseases [39–42].

1.4 Nanotechnologies Foster an Explosion of New Quantitative Information How Biological Nanosystems Work

Far less noticed by the general public are the next approaching waves of medical innovations, made possible by an explosion of new quantitative information how biological systems work.

The ultimate goal is to achieve an understanding of all the structural and dynamic processes by which the molecular players work with each other in living cells and coregulate cellular functions. Driven by the many technologies that have emerged from the physical, chemical, biological and engineering sciences, to visualize (see elsewhere in this volume [43, 44]) and manipulate the nanoworld, numerous discoveries are currently being made (as highlighted in later chapters [43–51]). These findings result from the new capabilities to create, analyze and manipulate nanostructures, as well as to probe their nanochemistry, nanomechanics and other properties within living and manmade systems. New technologies will continuously be developed that can interrogate biological samples with unprecedented temporal and spatial resolution [52]. Novel computational technologies have, furthermore, been developed to simulate cellular machineries in action with atomic precision [53]. New engineering design principles and technologies will be derived from deciphering how natural systems are engineered and how they master all the complex tasks that enable life. Take the natural machineries apart, and then learn how to reassemble their components (as exemplified here in Chapter 8 for molecular motors [48]).

How effectively will these novel insights into the biological nanoworld are recognized in their clinical significance, and translated into addressing grand medical challenges? This defines the time that it takes for the emergence of a next generation of diagnostic and therapeutic tools. As these insights change the way we think about the inner workings of cells and cell-made materials, totally new ways of treating diseases will emerge. As described in detail elsewhere in this volume, new developments are already under way of how to probe and control cellular activities [45, 47, 49–51, 54, 55]. This implicates the emergence of new methodologies of how to correct tissue and organ malfunctions. Clearly, we need to know exactly how each disease is associated with defects in the cellular machinery before medication can be rationally designed to effectively cure them.

Since every one of the new (nano)analytical techniques has the potency of revealing something never seen before, a plethora of opportunities can be envisioned. Their realization, however, hinges on the scientists' ability to recognize the physiological and medical relevance of their discoveries. This can best be accomplished in the framework of interdisciplinary efforts aimed at learning from each other what the new technologies can provide, and how this knowledge can be effectively translated to address major clinical challenges. Exploring exactly how these novel insights into the nanoworld will impact medicine has been the goal of many recent workshops [3–11, 56–58]). This stimulated the creation of the NIH

Roadmap Initiative in Nanomedicine [57, 58], and is the major focus of this volume.

1.5 Insights Gained from Quantifying how the Cellular Machinery Works will lead to Totally New Ways of Diagnosing and Treating Disease

Which are some of the central medical fields that will be impacted? Despite these stunning scientific advances, and the successful suppression or even eradication of a variety of infectious diseases during the past 100 years, the goal has not yet been reached of having medication at hand to truly cure many of the diseases that currently kill the largest fraction of humans per year, including cancers, cardiovascular diseases and AIDS. Much of the current medication against these diseases fights symptoms or inhibits their progression, often inflicting considerable side effects. Unfortunately, however, much of the medication can slow down but it cannot *reverse* disease progression in any major way – all of which contributes to healthcare becoming unaffordable, even in the richest nations of the world. For instance, intense cancer research over the past decades has revealed that the malignancy of cancer cells progresses with the gradual accumulation of genetic alterations [12, 50, 59–65]. Yet, little remains known as to how cancer stem cells form, in the first place, and about the basic mechanisms that trigger the initiation of their differentiation into more malignant cancer cells after having remained dormant, sometimes for decades [66–69]. While much has been learned in the past about the molecular players and their interactions, the above-mentioned shortcomings in translating certain advances in molecular and cell biology into more effective medication reflect substantial gaps in our knowledge of how all these components within the cells work in the framework of an integrated system. How can so many molecular players be tightly coordinated in a crowded cellular environment to generate predictable cell and tissue responses? Whilst lipid membranes create barriers that enclose the inner volumes of cells and control which molecules enter and leave (among other tasks), it is the proteins that are the 'workhorses' that enable most cell functions. In fact, some proteins function as motors that ultimately allow cells to move, as discussed in different contexts in the following chapters [46, 48, 50]. Other proteins transcribe and translate genetic information, and efforts to visualize these in cells have been summarized in Chapter 6 [45]. Yet other sets of proteins are responsible for the cell signaling through which all metabolic functions are enabled, orchestrated and regulated. But what are the underlying rules by which they play and interact together to regulate diverse cell functions? How do cells sense their environments, integrate that information, and translate it to ultimately adjust their metabolic functions if needed? Can this knowledge help to develop interventions which could possibly reverse pathogenic cells such that they performed their normal tasks again? Deciphering how all of these processes are regulated by the physical and biochemical microenvironment of cells is key to addressing various biomedical challenges with new perceptions, and is described as one of the major foci in this volume. But, how can this be accomplished?

1.6
Engineering Cell Functions with Nanoscale Precision

The engineering of nanoenvironments, nanoprobes and nanomanipulators, together with novel modalities to visualize phenomena at this tiny scale, have already led to the discovery of many unexpected mechanisms of how cellular nanoparts function, and how they cooperate synergistically when integrated into larger complex systems [43, 46–51]. Today, nanotechnology tools are particularly well suited to explore and quantify the physical aspects of biology, thereby complementing the tool chests of biochemists, molecular biologists and geneticists. Such nanotechnology approaches could already reveal that not only biochemical factors but also mechanical aspects as well as the micro- and nanoscale features of a cell's microenvironment, play pivotal roles in regulating their fate. The insights and implications thereof are described in chapters 9 to 14 [47, 49–51].

These discoveries are particularly relevant since most of our biomedical knowledge of how cells function has been derived in the past from the study of cells cultured on flat polymer surfaces (Petri dishes or on multi-welled plates). Cells in a more tissue-like environment, however, often show a vastly different behavior [70–75] (and chapters 9 to 14). With a still poorly understood cell signaling response system, cells in tissues thus 'see' and 'feel' an environment that is poorly mimicked by the common cell culturing conditions or scaffolds used in tissue engineering. Nanotechnologies will thus be pivotal to deciphering how cells sense and integrate a broad set of cues that regulate cell fate, from the moment that a sperm fertilizes an egg, to sustained, normal tissue functions. Moreover, these dependencies must be known in order to develop far more efficient drugs and treatment methods. However, ultimately it is the combination of many different technologies – some of which may originate from the physical sciences and others from biology – that must be combined to understand and quantify biology. Unfortunately, today an insufficient number of research workers are trained to perform these tasks [76].

1.7
Advancing Regenerative Medicine Therapies

Virtually any disease that results from malfunctioning, damaged or failing tissues may be potentially cured through regenerative medicine therapies, as was recently stated in the first NIH report on Regenerative Medicine [4]. But, how will nanotechnology make a difference? The repair – or ultimately replacement – of diseased organs, from larger bone segments to the spinal cord, or from the kidneys to the heart, still poses major challenges as discussed in chapters 9 to 14 [47, 49, 50, 54, 55, 77]. The current need for organ transplants surpasses the availability of suitable donor organs by at least an order of magnitude, and the patients who finally receive an organ transplant must receive immune suppressant drugs for the rest of their life. Thus, one goal will be to apply the mounting insights into how cells work, and how their functions are controlled by matrix interactions, to design alternate therapies that stimulate regenerative healing

processes of previously irreparable organs. In a most promising approach, some molecules have been designed that can self-assemble in the body into provisional matrices [55]. If these are injected shortly after injury, they help to repair spinal cord injuries and heart tissues damaged by an infarction. And if such strategies do not work, then another possibility might be to seed the patient's cells or stem cells into engineered biohybrid matrices to grow simple tissues *ex vivo* – that is, in the laboratory – and later implant them to support or regenerate failing organ functions [51, 54]. This could provide new ways of treating diabetes, liver and kidney failures, cardiovascular and many other diseases, or of replacing or repairing organs damaged in accidents or removed during surgery. Learning how to control the differentiation of stem cells in engineered matrices is therefore central to advancing our technical abilities in tissue engineering and regenerative medicine, and the challenges ahead as discussed in chapters 9 to 14 [50, 51, 54]. Nanofibers thereby mimic much better the fibrous nature of extracellular matrices [54], and the nanoscale patterning of ligands can control cell activation [49], including the activation of cells that play central roles in the immune response system (for a review, see [49]). In summary, the insights derived with the help of nanotechnology will enable the engineering of tissue-mimetic scaffolds that better control and regulate tissue function and repair. Improving human health will thus critically hinge upon translating nanotechnology-derived insights about cellular and tissue functions into novel diagnostic and therapeutic technologies.

1.8
Many More Relevant Medical Fields Will be Innovated by Nanotechnologies

Whilst the major focus of this volume is to outline the biomedical implications derived from revealing the underpinning mechanisms of how human cells function, it should also be mentioned that fascinating developments that are prone to alter medicine are being made in equally relevant other biomedical sectors. Future ways to treat infection will change when the underpinning mechanisms of how microbial systems function are deciphered, and how they interact with our cells and tissues. Many beautiful discoveries have already been made that will help us for example to interfere more effectively with the sophisticated machinery that bacteria have evolved to target, adhere and infect cells and tissues. Nanotechnology tools have revealed much about the function of the nano-engines that bacteria and other microbes have evolved for their movement [78–80], how bacteria adhere to surfaces [81–83], and how microbes infect other organisms [81–85]. Equally important when combating infection is an ability to exploit micro- and nanofabricated tools in order to understand the language by which microorganisms communicate with each other [86, 87] and how their inner machineries function [88–90]. A satisfying understanding of how a machine works can only be reached when we are capable of reassembling it from its components. It is thus crucial to learn how these machineries can be reassembled *ex vivo*, potentially even in nonbiological environments, as this should open the door to many technical and medical applications [84, 91–93]. Today, we have only just started along the route to combining the natural and synthetic worlds, with the community

seeking how bacteria might be used as 'delivery men' for nanocargoes [94], or in manmade devices to move fluids and objects [95, 96].

Finally, microfabricated devices with integrated nanosensors, nanomonitors and nanoreporters – all of which are intrinsic to a technology sector enabled by (micro/nano)biotechnology – will surely also lead to changes in medical practice. In the case of chemotherapies and many other drugs, it is well known that they may function well in some patients, but fail in others. It is feasable that this 'one-size-fits-all' approach might soon be replaced by a more patient-specific system. *Personalized medicine* refers to the use of genetic and other screening methods to determine the unique predisposition of a patient to a disease, and the likelihood of them responding to particular drugs and treatments [30, 97–100]. Cheap diagnostic systems that can automatically conduct measurements on small gas or fluid volumes, such as human breath or blood, will furthermore enable patients to be tested rapidly, without the need to send samples to costly medical laboratories. Needless to say, portable integrated technologies that will allow the testing and treatment of patients on the spot (point-of-care) will save many lives, and are urgently needed to improve human health in the Third World [101–104].

Faced with major challenges in human healthcare, an understanding what each of the many nanotechnologies can do – and how they each can best contribute to address the major challenges ahead – is crucial to drive innovation forwards. An improved awareness of how new technologies will help to unravel underpinning mechanisms of disease is crucial to setting realistic expectations and timescales, as well as to prepare for the innovations to come.

Since ultimately, thriving towards providing access to efficient and affordable healthcare, by improving upon technology, is not just an intellectual luxury, but our responsibility.

Acknowledgments

Many of the thoughts about the future of nanomedicine were seeded during exciting discussions with Drs Eleni Kousvelari (NIH) and Jeff Schloss (NIH), as well as with Dr Mike Roco (NSF), when preparing for the interagency workshop on Nanobiotechnology [9]. The many inspirations and contributions from colleagues and friends worldwide are gratefully acknowledged, several of whom have contributed chapters here, as well as from my students and collaborators and the many authors whose articles and conference talks have left long-lasting impressions.

References

1 Nalwa, H.S. (2004) *Encyclopedia of Nanoscience and Nanotechnology*, American Scientific Publishers.

2 Roco, M.C. (2003) National Nanotechnology Initiative: From Vision to Implementation. http://www.nano.gov/nni11600/sld001.htm.

3 National Institutes of Health (2000): Nanoscience and Nanotechnology: Shaping Biomedical Research. Organized

by the Bioengineering Consortium. Available at http://www.becon.nih.gov/nanotechsympreport.pdf.

4 National Institutes of Health (2003) 2020: A New Vision – A Future for Regenerative Medicine, U.S. Department of Health and Human Services. Available at http://www.hhs.gov/reference/newfuture.shtml

5 National Institutes of Health (2003) Nanotechnology in Heart, Lung, Blood, and Sleep Medicine. http://www.nhlbi.nih.gov/meetings/nano_sum.htm.

6 Duncan, R. (2005) Nanomedicine. Forward Looking Report. http://www.esf.org/publication/214/Nanomedicine.pdf, European Science Foundation (ESF)-European Medical Research Council (EMRC). Report Number 2-912049-52-0.

7 NIBIB/DOE (2005) Workshop on Biomedical Applications of Nanotechnology. Organized by the National Institute of Biomedical Imaging and BioEngineering, Department of Energy, USA. Available at http://www.capconcorp.com/nibibdoenanotech/

8 Vogel, V. and Baird, B. (2005) 'BioNanotechnology'. National Nanotechnology Initiative Grand Challenge Workshop Report. Available at http://nano.gov/nni_nanobiotechnology_rpt.pdf.

9 European-Technology-Platform (2006) Nanomedicine: Nanotechnology for Health. Available at ftp://ftp.cordis.europa.eu/pub/nanotechnology/docs/nanomedicine_bat_en.pdf

10 Nantional Institute of Health (2008) Roadmap Initiative in Nanomedicine. Available at http://nihroadmap.nih.gov/nanomedicine/index.

11 Schmid, G., Brune, H., Ernst, H., Grünwald, W., Grunwald, A., Hofmann, H., Janich, P., Krug, H., Mayor, M., Rathgeber, W., Simon, U., Vogel, V. and Wyrwa, D. (2006) *Nanotechnology. Assessment and Perspectives*, Springer Verlag.

12 Godin, B., Serda, R.E., Sakamoto, J., Decuzzi, P. and Ferrari, M. (2009) Nanoparticles for cancer detection and therapy, in *Nanomedicine* (ed. V. Vogel), Ch. 3, Wiley-VCH, Weinheim.

13 Wallace, K.D., Hughes, M.S., Marsh, J.N., Caruthers, S.D., Lanza, G.M. and Wickline, S.A. (2009) From i*n vivo* ultrasound and MRI imaging to therapy: Contrast agents based on target-specific nanoparticles, in *Nanomedicine* (ed. V. Vogel), Ch. 2, Wiley-VCH, Weinheim.

14 Johannsen, M., Gneveckow, U., Thiesen, B., Taymoorian, K., Cho, C.H., Waldofner, N., Scholz, R., Jordan, A., Loening, S.A. and Wust, P. (2007) Thermotherapy of prostate cancer using magnetic nanoparticles: feasibility, imaging, and three-dimensional temperature distribution. *European Urology*, **52** (6), 1653–1661.

15 National Cancer Institute (2008). Alliance for Nanotechnology in Cancer. http://nano.cancer.gov/resource_center/tech_backgrounder.asp.

16 Phelps, M.E. (2004) *PET: Molecular Imaging and Its Biological Applications*, Springer.

17 National Cancer Institute (2004) Cancer Nanotechnology. Going Small for Big Advances, NIH.

18 Jain, K.K. (2007) Applications of nanobiotechnology in clinical diagnostics. *Clinical Chemistry*, **53** (11), 2002–2009.

19 Nie, S., Xing, Y., Kim, G.J. and Simons, J.W. (2007) Nanotechnology applications in cancer. *Annual Review of Biomedical Engineering*, **9**, 257–288.

20 Kumar, C.S.S.R. (ed.) (2008) *Nanomaterials for Cancer Diagnosis*, Nanotechnologies for the Life Sciences, Wiley-VCH, Weinheim.

21 Kumar, C.S.S.R. (ed.) (2008) *Nanomaterials for Cancer Therapy*. Nanotechnologies for the Life Sciences, Wiley-VCH, Weinheim.

22 Kumar, C.S.S.R. (ed.) (2008) *Nanomaterials for Medical Diagnosis and*

Therapy Nanotechnologies for the Life Sciences, Wiley-VCH, Weinheim.

23 Perutz, M.F. (1962) X-ray analysis of haemoglobin. Nobel Lecture.

24 Mullis, K.B. (1993) The Polymerase Chain Reaction. Nobel Lecture.

25 Cheung, V.G., Nowak, N., Jang, W., Kirsch, I.R., Zhao, S., Chen, X.N., Furey, T.S., Kim, U.J., Kuo, W.L., Olivier, M., Conroy, J., Kasprzyk, A., Massa, H., Yonescu, R., Sait, S., Thoreen, C., Snijders, A., Lemyre, E., Bailey, J.A., Bruzel, A., Burrill, W.D., Clegg, S.M., Collins, S., Dhami, P., Friedman, C., Han, C.S., Herrick, S., Lee, J., Ligon, A.H., Lowry, S., Morley, M., Narasimhan, S., Osoegawa, K., Peng, Z., Plajzer-Frick, I., Quade, B.J., Scott, D., Sirotkin, K., Thorpe, A.A., Gray, J.W., Hudson, J., Pinkel, D., Ried, T., Rowen, L., Shen-Ong, G.L., Strausberg, R.L., Birney, E., Callen, D.F., Cheng, J.F., Cox, D.R., Doggett, N.A., Carter, N.P., Eichler, E.E., Haussler, D., Korenberg, J.R., Morton, C.C., Albertson, D., Schuler, G., de Jong, P.J. and Trask, B.J. (2001) Integration of cytogenetic landmarks into the draft sequence of the human genome. *Nature* **409** (6822), 953–958.

26 Helmuth, L. (2001) Genome research: map of the human genome 3.0. *Science*, **293** (5530), 583–585.

27 McPherson, J.D., Marra, M., Hillier, L., Waterston, R.H., Chinwalla, A., Wallis, J., Sekhon, M., Wylie, K., Mardis, E.R., Wilson, R.K., Fulton, R., Kucaba, T.A., Wagner-McPherson, C., Barbazuk, W.B., Gregory, S.G., Humphray, S.J., French, L., Evans, R.S., Bethel, G., Whittaker, A., Holden, J.L., McCann, O.T., Dunham, A., Soderlund, C., Scott, C.E., Bentley, D.R., Schuler, G., Chen, H.C., Jang, W., Green, E.D., Idol, J.R., Maduro, V.V., Montgomery, K.T., Lee, E., Miller, A., Emerling, S., Kucherlapati, Gibbs, R., Scherer, S., Gorrell, J.H., Sodergren, E., Clerc-Blankenburg, K., Tabor, P., Naylor, S., Garcia, D., de Jong, P.J., Catanese, J.J., Nowak, N., Osoegawa, K., Qin, S., Rowen, L., Madan, A., Dors, M., Hood, L., Trask, B., Friedman, C., Massa, H., Cheung, V.G., Kirsch, I.R., Reid, T., Yonescu, R., Weissenbach, J., Bruls, T., Heilig, R., Branscomb, E., Olsen, A., Doggett, N., Cheng, J.F., Hawkins, T., Myers, R.M., Shang, J., Ramirez, L., Schmutz, J., Velasquez, O., Dixon, K., Stone, N.E., Cox, D.R., Haussler, D., Kent, W.J., Furey, T., Rogic, S., Kennedy, S., Jones, S., Rosenthal, A., Wen, G., Schilhabel, M., Gloeckner, G., Nyakatura, G., Siebert, R., Schlegelberger, B., Korenberg, J., Chen, X.N., Fujiyama, A., Hattori, M., Toyoda, A., Yada, T., Park, H.S., Sakaki, Y., Shimizu, N., Asakawa, S., Kawasaki, K., Sasaki, T., Shintani, A., Shimizu, A., Shibuya, K., Kudoh, J., Minoshima S., Ramser J., Seranski P., Hoff C., Poustka A., Reinhardt R. and Lehrach H. (2001) A physical map of the human genome. *Nature* **409** (6822), 934–941.

28 Olivier, M., Aggarwal, A., Allen, J., Almendras, A.A., Bajorek, E.S., Beasley, E.M., Brady, S.D., Bushard, J.M., Bustos, V.I., Chu, A., Chung, T.R., De Witte, A., Denys, M.E., Dominguez, R., Fang, N.Y., Foster, B.D., Freudenberg, R.W., Hadley, D., Hamilton, L.R., Jeffrey, T.J., Kelly, L., Lazzeroni, L., Levy, M.R., Lewis, S.C., Liu, X., Lopez, F.J., Louie, B., Marquis, J.P., Martinez, R.A., Matsuura, M.K., Misherghi, N.S., Norton, J.A., Olshen, A., Perkins, S.M., Perou, A.J., Piercy, C., Piercy, M., Qin, F., Reif, T., Sheppard, K., Shokoohi, V., Smick, G.A., Sun, W.L., Stewart, E.A., Fernando, J., Tejeda, J., Tran, N.M., Trejo, T., Vo, N.T., Yan, S.C., Zierten, D.L., Zhao, S., Sachidanandam, R., Trask, B.J., Myers, R.M. and Cox, D.R. (2001) A high-resolution radiation hybrid map of the human genome draft sequence. *Science*, **291** (5507), 1298–1302.

29 Reich, D.E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P.C., Richter, D.J.,

Lavery, T., Kouyoumjian, R., Farhadian, S.F., Ward, R. and Lander, E.S. (2001) Linkage disequilibrium in the human genome. *Nature*, **411** (6834), 199–204.

30 Ideker, T. and Sharan, R. (2008) Protein networks in disease. *Genome Research*, **18** (4), 644–652.

31 King, M.C., Marks, J.H. and Mandell, J.B. (2003) Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2. *Science*, **302** (5645), 643–646.

32 Lin, M.T. and Beal, M.F. (2006) Mitochondrial dysfunction and oxidative stress in neurodegenerative diseases. *Nature*, **443** (7113), 787–795.

33 Kato, T. (2007) Molecular genetics of bipolar disorder and depression. *Psychiatry and Clinical Neurosciences*, **61** (1), 3–19.

34 Madsen, E. and Gitlin, J.D. (2007) Copper and iron disorders of the brain. *Annual Review of Neuroscience*, **30**, 317–337.

35 Oti, M. and Brunner, H.G. (2007) The modular nature of genetic diseases. *Clinical Genetics*, **71** (1), 1–11.

36 Abrahams, B.S. and Geschwind, D.H. (2008) Advances in autism genetics: on the threshold of a new neurobiology. *Nature Reviews. Genetics*, **9** (5), 341–355.

37 Ferrell, R.E. and Finegold, D.N. (2008) Research perspectives in inherited lymphatic disease: an update. *Annals of the New York Academy of Sciences*, **1131**, 134–139.

38 Judge, D.P. and Dietz, H.C. (2008) Therapy of Marfan syndrome. *Annual Review of Medicine*, **59**, 43–59.

39 Weinreich, F. and Jentsch, T.J. (2000) Neurological diseases caused by ion-channel mutations. *Current Opinion in Neurobiology*, **10** (3), 409–415.

40 Wilde, A.A. and van den Berg, M.P. (2005) Ten years of genes in inherited arrhythmia syndromes: an example of what we have learned from patients, electrocardiograms, and computers. *Journal of Electrocardiology*, **38** (4 Suppl), 145–149.

41 Fiske, J.L., Fomin, V.P., Brown, M.L., Duncan, R.L. and Sikes, R.A. (2006) Voltage-sensitive ion channels and cancer. *Cancer Metastasis Reviews*, **25** (3), 493–500.

42 Terrenoire, C., Simhaee, D. and Kass, R.S. (2007) Role of sodium channels in propagation in heart muscle: how subtle genetic alterations result in major arrhythmic disorders. *Journal of Cardiovascular Electrophysiology*, **18** (8), 900–905.

43 Baker, M.L., Marsh, M.P. and Chiu, W. (2009) Electron cryo-microscopy of molecular nanomachines and cells, in *Nanomedicine* (ed. V. Vogel), Ch. 4, Wiley-VCH, Weinheim.

44 Kukura, P., Renn, A. and Sandoghdar, V. (2009) Pushing optical microscopy to the limit: from single-molecule fluorescence microscopy to label-free detection and tracking of biological nano-objects, in *Nanomedicine* (ed. V. Vogel), Ch. 5, Wiley-VCH, Weinheim.

45 Bao, G., Santangelo, P., Nitin, N. and Rhee, W.-J. (2009) Nanostructured probes for *in vivo* gene detection, in *Nanomedicine* (ed. V. Vogel), Ch. 6, Wiley-VCH, Weinheim.

46 Applegate, K.T., Yang, G. and Danuser, G. (2009) High-content analysis of cytoskeleton functions by fluorescent speckle microscopy, in *Nanomedicine* (ed. V. Vogel), Ch. 7, Wiley-VCH, Weinheim.

47 Fratzl, P., Gupta, H.S., Roschger, P. and Klaushofer, K. (2009) Bone nanostructure and its relevance for mechanical performance, disease and treatment, in *Nanomedicine* (ed. V. Vogel), Ch. 12, Wiley-VCH, Weinheim.

48 Goel, A. and Vogel, V. (2009) Transport, assembly and proof-reading: harnessing the engineering principles of biological nanomotors, in *Nanomedicine* (ed. V. Vogel), Ch. 8, Wiley-VCH, Weinheim.

49 Dunlop, I.E., Dustin, M.L. and Spatz, J.P. (2009) The Micro- and Nanoscale

Architecture of the Immunological Synapse, in *Nanomedicine* (ed. V. Vogel), Ch. 11, Wiley-VCH, Weinheim.

50 Vogel, V. and Sheetz, M.P. (2009) Mechanical Forces Matter in Health and Disease: From Cancer to Tissue Engineering, in *Nanomedicine* (ed. V. Vogel), Ch. 9, Wiley-VCH, Weinheim.

51 Rehfeldt, F., Engler, A.J. and Discher, D.E. (2009) Stem Cells and Nanomedicine: Nanomechanics of the Microenvironment, in *Nanomedicine* (ed. V. Vogel), Ch. 10, Wiley-VCH, Weinheim.

52 Shorokhov, D. and Zewail, A.H. (2008) 4D electron imaging: principles and perspectives. *Physical Chemistry Chemical Physics*, **10** (20), 2879–2893.

53 Sotomayor, M. and Schulten, K. (2007) Single-molecule experiments in vitro and *in silico*. *Science*, **316** (5828), 1144–1148.

54 Khademhosseini, A., Rajalingam, B., Jinno, S. and Langer, R. (2009) Nanoengineered Systems for Tissue Engineering and Regeneration, in *Nanomedicine* (ed. V. Vogel), Ch. 13, Wiley-VCH, Weinheim.

55 Capito, R.M., Mata, A. and Stupp, S.I. (2009) Self-Assembling Peptide-Based Nanostructures for Regenerative Medicine, in *Nanomedicine* (ed. V. Vogel), Ch. 14, Wiley-VCH, Weinheim.

56 Alper, M.D. and Stupp, S.I. (2003) Biomolecular Materials. Basic Energy Sciences Advisory Committee to the Office of Science (DOE) Report http://www.sc.doe.gov/bes/besac/BiomolecularMaterialsReport.pdf.

57 NIH (2008) Roadmap Initiative in Nanomedicine. http://nihroadmap.nih.gov/nanomedicine/index.

58 NIH (2008) US National Cancer Institute (NCI), Alliance for Nanotechnology in Cancer.

59 Sekido, Y., Fong, K.M. and Minna, J.D. (2003) Molecular genetics of lung cancer. *Annual Review of Medicine*, **54**, 73–87.

60 Ishikawa, T., Zhang, S.S., Qin, X., Takahashi, Y., Oda, H., Nakatsuru, Y. and Ide, F. (2004) DNA repair and cancer: lessons from mutant mouse models. *Cancer Science*, **95** (2), 112–117.

61 Sogn, J.A., Anton-Culver, H. and Singer, D.S. (2005) Meeting report: NCI think tanks in cancer biology. *Cancer Research*, **65** (20), 9117–9120.

62 Makrantonaki, E. and Zouboulis, C.C. (2007) Molecular mechanisms of skin aging: state of the art. *Annals of the New York Academy of Sciences*, **1119**, 40–50.

63 Wren, B.G. (2007) The origin of breast cancer. *Menopause (New York, NY)*, **14** (6), 1060–1068.

64 Frey, A.B. and Monu, N. (2008) Signaling defects in anti-tumor T cells. *Immunological Reviews*, **222**, 192–205.

65 Savage, S.A. and Alter, B.P. (2008) The role of telomere biology in bone marrow failure and other disorders. *Mechanisms of Ageing and Development*, **129** (1–2), 35–47.

66 Indraccolo, S., Favaro, E. and Amadori, A. (2006) Dormant tumors awaken by a short-term angiogenic burst: the spike hypothesis. *Cell Cycle (Georgetown, Tex)*, **5** (16), 1751–1755.

67 Townson, J.L. and Chambers, A.F. (2006) Dormancy of solitary metastatic cells. *Cell Cycle (Georgetown, Tex)*, **5** (16), 1744–1750.

68 Vessella, R.L., Pantel, K. and Mohla, S. (2007) Tumor cell dormancy: an NCI workshop report. *Cancer Biology & Therapy*, **6** (9), 1496–1504.

69 Riethdorf, S. and Pantel, K. (2008) Disseminated tumor cells in bone marrow and circulating tumor cells in blood of breast cancer patients: current state of detection and characterization. *Pathobiology*, **75** (2), 140–148.

70 Goodman, S.L., Sims, P.A., Albrecht, R.M. (1996) Three-dimensional extracellular matrix textured biomaterials. *Biomaterials*, **17** (21), 2087–2095.

71 Friedl, P. and Brocker, E.B. (2000) The biology of cell locomotion within three-

dimensional extracellular matrix. *Cellular and Molecular Life Sciences*, **57** (1), 41–64.

72 Cukierman, E., Pankov, R., Stevens, D.R. and Yamada, K.M. (2001) Taking cell-matrix adhesions to the third dimension. *Science*, **294** (5547), 1708–1712.

73 Grinnell, F. (2003) Fibroblast biology in three-dimensional collagen matrices. *Trends in Cell Biology*, **13** (5), 264–269.

74 Li, S., Moon, J.J., Miao, H., Jin, G., Chen, B.P., Yuan, S., Hu, Y., Usami, S. and Chien, S. (2003) Signal transduction in matrix contraction and the migration of vascular smooth muscle cells in three-dimensional matrix. *Journal of Vascular Research*, **40** (4), 378–388.

75 Larsen, M., Artym, V.V., Green, J.A. and Yamada, K.M. (2006) The matrix reorganized: extracellular matrix remodeling and integrin signalling. *Current Opinion in Cell Biology*, **18** (5), 463–471.

76 Stryer, L. (2003) Bio2010: Transforming Undergraduate Education for Future Research Biologists. National Research Council Report website.

77 Ratner, B.D. and Bryant, S.J. (2004) Biomaterials: where we have been and where we are going. *Annual Review of Biomedical Engineering*, **6**, 41–75.

78 Berg, H.C. (2003) The rotary motor of bacterial flagella. *Annual Review of Biochemistry*, **72**, 19–54.

79 Weibel, D.B., Diluzio, W.R. and Whitesides, G.M. (2007) Microfabrication meets microbiology. *Nature Reviews Microbiology*, **5** (3), 209–218.

80 Jarrell, K.F. and McBride, M.J. (2008) The surprisingly diverse ways that prokaryotes move. *Nature Reviews Microbiology*, **6** (6), 466–476.

81 Sokurenko, E., Vogel, V. and Thomas, W.E. (2008) Catch bond mechanism of force-enhanced adhesion: counter-intuitive, elusive but . . . widespread? *Cell Host & Microbe, October*.

82 Thomas, W.E., Vogel, V. and Sokurenko, E. (2008) Biophysics of catch bonds. *Annual Review of Biophysics*, **37**, 399–416.

83 Biais, N., Ladoux, B., Higashi, D., So, M. and Sheetz, M. (2008) Cooperative retraction of bundled type IV pili enables nanonewton force generation. *PLoS Biology*, **6** (4), e87.

84 Fletcher, D.A. and Theriot, J.A. (2004) An introduction to cell motility for the physical scientist. *Physical Biology*, **1** (1–2), T1–T10.

85 Brandenburg, B. and Zhuang, X. (2007) Virus trafficking – learning from single-virus tracking. *Nature Reviews Microbiology*, **5** (3), 197–208.

86 Balagadde, F.K., Song, H., Ozaki, J., Collins, C.H., Barnet, M., Arnold, F.H., Quake, S.R. and You, L. (2008) A synthetic *Escherichia coli* predator-prey ecosystem. *Molecular Systems Biology*, **4**, 187.

87 Welch, M., Mikkelsen, H., Swatton, J.E., Smith, D., Thomas, G.L., Glansdorp, F.G. and Spring, D.R. (2005) Cell-cell communication in Gram-negative bacteria. *Molecular BioSystems*, **1** (3), 196–202.

88 Xie, X.S., Choi, P.J., Li, G.W., Lee, N.K. and Lia, G. (2008) Single-molecule approach to molecular biology in living bacterial cells. *Annual Review of Biophysics*, **37**, 417–444.

89 Johansson, M., Lovmar, M. and Ehrenberg, M. (2008) Rate and accuracy of bacterial protein synthesis revisited. *Current Opinion in Microbiology*, **11** (2), 141–147.

90 Zorrilla, S., Lillo, M.P., Chaix, D., Margeat, E., Royer, C.A. and Declerck, N. (2008) Investigating transcriptional regulation by fluorescence spectroscopy, from traditional methods to state-of-the-art single-molecule approaches. *Annals of the New York Academy of Sciences*, **1130**, 44–51.

91 Forero, M., Thomas, W., Bland, C., Nilsson, L., Sokurenko, E. and Vogel, V. (2004) A catch-bond based nano-adhesive sensitive to shear stress. *Nano Letters*, **4**, 1593–1597.

92 Sleytr, U.B., Huber, C., Ilk, N., Pum, D., Schuster, B. and Egelseer, E.M. (2007) S-layers as a tool kit for nanobiotechnological applications. *FEMS Microbiology Letters*, **267** (2), 131–144.

93 Chevance, F.F. and Hughes, K.T. (2008) Coordinating assembly of a bacterial macromolecular machine. *Nature Reviews Microbiology*, **6** (6), 455–465.

94 Diao, J.J., Hua, D., Lin, J., Teng, H.H. and Chen, D. (2005) Nanoparticle delivery by controlled bacteria. *Journal of Nanoscience and Nanotechnology*, **5** (10), 1749–1751.

95 Darnton, N., Turner, L., Breuer, K. and Berg, H.C. (2004) Moving fluid with bacterial carpets. *Biophysical Journal*, **86** (3), 1863–1870.

96 Hiratsuka, Y., Miyata, M., Tada, T. and Uyeda, T.Q. (2006) A microrotary motor powered by bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, **103** (37), 13618–13623.

97 Turner, S.T., Schwartz, G.L. and Boerwinkle, E. (2007) Personalized medicine for high blood pressure. *Hypertension*, **50** (1), 1–5.

98 Katsanis, S.H., Javitt, G. and Hudson, K. (2008) Public health. A case study of personalized medicine. *Science*, **320** (5872), 53–54.

99 Zhong, J.F., Chen, Y., Marcus, J.S., Scherer, A., Quake, S.R., Taylor, C.R. and Weiner, L.P. (2008) A microfluidic processor for gene expression profiling of single human embryonic stem cells. *Lab on a Chip*, **8** (1), 68–74.

100 van't Veer, L.J. and Bernards, R. (2008) Enabling personalized cancer medicine through analysis of gene-expression patterns. *Nature*, **452** (7187), 564–570.

101 Toner, M. and Irimia, D. (2005) Blood-on-a-chip. *Annual Review of Biomedical Engineering*, **7**, 77–103.

102 Yager, P., Edwards, T., Fu, E., Helton, K., Nelson, K., Tam, M.R. and Weigl, B.H. (2006) Microfluidic diagnostic technologies for global public health. *Nature*, **442** (7101), 412–418.

103 Park, J.Y. and Kricka, L.J. (2007) Prospects for nano- and microtechnologies in clinical point-of-care testing. *Lab on a Chip*, **7** (5), 547–549.

104 Phillips, K.A., Liang, S.Y. and Van Bebber, S. (2008) Challenges to the translation of genomic information into clinical practice and health policy: Utilization, preferences and economic value. *Current Opinion in Molecular Therapeutics*, **10** (3), 260–266.

Part Two:
Imaging, Diagnostics and Disease Treatment by Using Engineered Nanoparticles

Nanotechnology, Volume 5: Nanomedicine. Edited by Viola Vogel

ISBN: 978-3-527-31736-3

2
From *In Vivo* Ultrasound and MRI Imaging to Therapy: Contrast Agents Based on Target-Specific Nanoparticles

Kirk D. Wallace, Michael S. Hughes, Jon N. Marsh, Shelton D. Caruthers, Gregory M. Lanza, and Samuel A. Wickline

2.1
Introduction

Advances and recent developments in the scientific areas of genomics and molecular biology have created an unprecedented opportunity to identify clinical pathology in pre-disease states. Building on these advances, the field of molecular imaging has emerged, leveraging the sensitivity and specificity of molecular markers together with advanced noninvasive imaging modalities to enable and expand the role of noninvasive diagnostic imaging. However, the detection of small aggregates of precancerous cells and their biochemical signatures remains an elusive target that is often beyond the resolution and sensitivity of conventional magnetic resonance and acoustic imaging techniques. The identification of these molecular markers requires target-specific probes, a robust signal amplification strategy, and sensitive high-resolution imaging modalities.

Currently, several nanoparticle or microparticle systems are under development for targeted diagnostic imaging and drug delivery [1]. Perfluorocarbon (PFC) nanoparticles represent a unique platform technology, which may be applied to multiple clinically relevant imaging modalities. They exploit many of the key principles employed by other imaging agents. Ligand-directed, lipid-encapsulated PFC nanoparticles (with a nominal diameter of 250 nm) have inherent physicochemical properties which provide acoustic contrast when the agent is bound to a surface layer. The high surface area of the nanoparticle accommodates 100 to 500 targeting molecules (or ligands), which impart high avidity and provides the agent with a robust 'stick and stay' quality (Figure 2.1). The incorporation of large payloads of lipid-anchored gadolinium chelate conjugates further extends the utility of the agent to detect sparse concentrations of cell-surface biochemical markers with magnetic resonance imaging (MRI) [2]. Moreover, for MRI the high fluorine signal from the nanoparticle core allows the noninvasive quantification of ligand-bound particles, which enables clinicians to confirm tissue concentrations of drugs when

Nanotechnology, Volume 5: Nanomedicine. Edited by Viola Vogel

ISBN: 978-3-527-31736-3

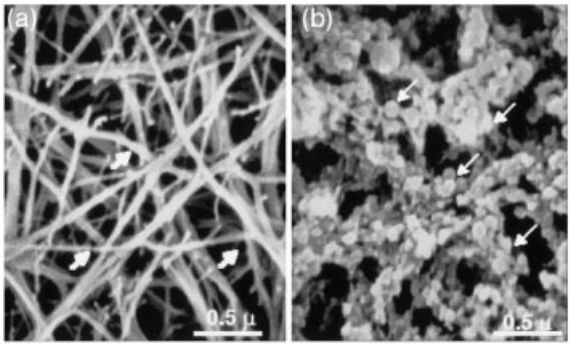


Figure 2.1 Scanning electron microscopy images (original magnification, ×30 000) of (a) a control fibrin clot and (b) fibrin-targeted paramagnetic nanoparticles bound to the clot surface. Arrows in (a) indicate a fibrin fibril; arrows in (b) indicate fibrin-specific nanoparticle-bound fibrin epitopes.

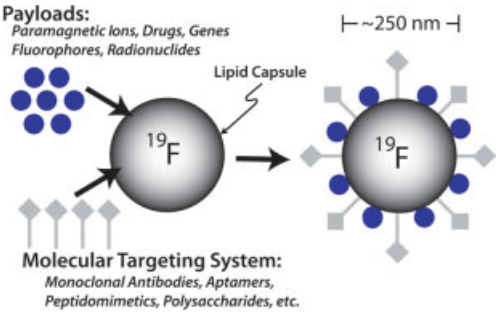


Figure 2.2 Paradigm for targeted liquid perfluorocarbon-based nanoparticle contrast agent. This example has a payload of Gd^{3+} chelates and monoclonal antibodies. The platform is extremely versatile, applicable to almost any imaging modality, and capable of carrying other payloads such as drugs or genes.

the functionality of the nanoparticles is extended to include targeted therapy (Figure 2.2). The detection of sparse concentrations of cell-surface biochemical markers is also possible with ultrasound [3]; however, novel signal processing is required for this application [4–6].

2.2 Active versus Passive Approaches to Contrast Agent Targeting

The passive targeting of a contrast agent is achieved by exploiting the body's inherent defense mechanisms for the clearance of foreign particles. Macrophages of the macrophage phagocytic system are responsible for the removal of most of these contrast agents from the circulation; these are produced, in size-dependent fashion,

from the lung, spleen, liver and bone marrow. Phagocytosis and accumulation within specific sites can be enhanced by biologic tagging (i.e. *opsonization*) with blood proteins such as immunoglobulins, complement proteins or nonimmune serum factors. In general, sequestration in the liver appears to be complement-mediated, while the spleen removes foreign particulate matter via antibody F_c receptors [8]. This natural process of nondirected and nonspecific uptake of particles is generally referred to as '*passive targeting*' (e.g. Feridex in the liver, or iron oxide in the sentinel lymph nodes [9]).

Distinguished from passive contrast agents, targeted (i.e. 'ligand-directed') contrast agents are designed to enhance specific pathological tissue that otherwise might be difficult to distinguish from the surrounding normal tissues. Here, an extensive array of ligands can be utilized, including monoclonal antibodies and fragments, peptides, polysaccharides, aptamers and drugs. These ligands may be attached either covalently (i.e. by direct conjugation) or noncovalently (i.e. by indirect conjugation) to the contrast agent. Engineered surface modifications, such as the incorporation of polyethylene glycol (PEG), are used to delay or avoid the rapid systemic removal of the agents, such that ligand-to-target binding is allowed to occur.

The effectiveness of this concept of contrast agent targeting is demonstrated with the application of paramagnetic MRI contrast agents. Paramagnetic agents influence only those protons in their immediate vicinity, and removal of these contrast agents by the macrophage phagocytic system during passive targeting may decrease their effectiveness via two mechanisms: (i) an accumulation of contrast agent in specific organs that are distal to region of interest; and (ii) endocytosis, which further decreases their exposure to free water protons. By targeting the contrast agent, the paramagnetic ions can be brought in close proximity to the region of interest with sufficient accumulation to overcome the partial dilution effect that plagues some MRI contrast agents. Its efficacy is further enhanced with some targeting platforms by delivering multiple contrast ions per particle [2].

2.3
Principles of Magnetic Resonance Contrast Agents

The fundamental physics underpinning MRI is grounded in the quantum mechanical magnetic properties of the atomic nucleus. All atomic nuclei have a fundamental property known as the *nuclear magnetic momentum* or *spin quantum number*. Individual protons and neutrons are fermions that possess an intrinsic angular momentum, or 'spin', quantized with a value of 1/2 [10, 11].

The overall spin of a nucleus (a composite fermion) is determined by the numbers of neutrons and protons. In nuclei with even numbers of protons and an even numbers of neutrons, these nucleons pair up to result in a net spin of zero. Nuclei with an odd number of protons or neutrons will have a nonzero net spin which, when placed in a strong magnetic field (with magnitude B_0), will have an associated net magnetic moment, $\vec{\mu}$, that will orient either with ('parallel') or against ('anti-parallel') the direction of B_0. For a nucleus with a net spin of 1/2 (e.g. ^{1}H), this results in two

possible spin states with an energy of separation $\Delta E = h\gamma B_0/2\pi$ (where $h = 6.626 \times 10^{-34}$ J s is Planck's constant and γ is the *gyromagnetic ratio*, which for hydrogen is equal to 42.58 MHz T^{-1}).

For a given population of nuclei in a static magnetic field, an equilibrium exists, described by Maxwell–Boltzmann statistics, in which only a slight majority of nuclei are oriented in the 'parallel' position (i.e. a lower energy state); however, this small difference in spin distribution results in a net magnetization that is perceptible. Absorption of the appropriately tuned radiofrequency (RF) radiation by the nuclei can alter the equilibrium distribution of 'anti-parallel' states. On a macroscopic level, this is equivalent to tilting the net magnetization away from the direction of the main magnetic field (B_0). Once the RF energy is removed, decay to the previous lower energy state takes place and occurs in two distinct and independent processes known as longitudinal relaxation (T_1) and transverse relaxation (T_2). The relaxation times, T_1 and T_2, as well as the proton density of the nuclei of interest, determine the signal intensity for various types of tissue in MRI.

Magnetic resonance contrast agents function by accelerating the longitudinal and/or transverse relaxation rates. The most commonly used nontargeted MR contrast agents are paramagnetic ions (e.g. gadolinium chelates), and these predominantly shorten T_1 relaxation to result in a bright signal on T_1-weighted images. The mechanism by which paramagnetic ions affect T_1 relaxation depends upon close nuclear interaction with protons (^{1}H) in water molecules (H_2O). Therefore, T_1 agents only influence protons proximate to themselves and are highly dependent on local water flux [12].

Superparamagnetic and ferromagnetic compounds have a high magnetic susceptibility, and when placed in a magnetic field (B_0) they concentrate the field; this results in a large local net positive magnetization [13]. This large magnetic susceptibility heterogeneity induces spin dephasing in tissue and results in a loss of the T_2-weighted signal. In contrast to T_1 contrast agents, superparamagnetic agents disturb the magnetic field and have a net effect far beyond their immediate vicinity.

2.3.1
Mathematics of Signal Contrast

To elucidate the source of image contrast, let us assume that two adjacent tissue types (A and B) manifest identical longitudinal (T_1) and transverse (T_2) relaxation times prior to nanoparticle binding, but only one tissue (say, type B) expresses the molecular epitope of interest that binds the targeted paramagnetic nanoparticles. The bound paramagnetic nanoparticles affect the relaxation times in the targeted tissue according to the following equations [14]:

$$\frac{1}{T_{1B}} = \frac{1}{T_{1A}} + r_{1P}\langle NP\rangle \tag{2.1}$$

$$\frac{1}{T_{2B}} = \frac{1}{T_{2A}} + r_{2P}\langle NP\rangle \tag{2.2}$$

where T_{1B} and T_{2B} are the observed relaxation times after the nanoparticle binding, T_{1A} and T_{2A} are the original relaxation times, r_{1P} and r_{2P} are the particle-based

relaxivities, and $\langle NP \rangle$ represents the average nanoparticle concentration within the imaging voxel. For the purpose of this example, the assumption is made that targeted binding does not affect particle relaxivity (i.e. r_{1p} and r_{2p} are constant).

The contrast-to-noise ratio (CNR) between the two tissues for a given sequence is calculated as the absolute difference between their signal intensities. If I_A and I_B represent the signal intensities of tissue A and B respectively, and N is the expected level of noise in the resulting image, the CNR ratio is given by:

$$\mathrm{CNR} = \frac{I_A - I_B}{N} \tag{2.3}$$

For a spin echo pulse sequence, the signal intensity of each tissue is related to the chosen scan parameters (echo time, TE, and repetition time, TR) as well as its magnetic properties (T_1 and T_2), which change due to binding of the contrast agent, and is described with the following relationships for tissues A and B [15]:

$$I_A = k_A(1 - 2e^{-(TR - TE/2)/T_{1A}} + e^{-TR/T_{1A}})e^{-TE/T_{2A}} \tag{2.4}$$

$$I_B = k_B(1 - 2e^{-(TR - TE/2)/T_{1B}} + e^{-TR/T_{1B}})e^{-TE/T_{2B}} \tag{2.5}$$

The constants k_A and k_B incorporate factors such as proton density, RF excitation and coil sensitivity. As these tissues are assumed identical, except for binding of the contrast agent, k_A and k_B are identical except for relative coil sensitivity to the positional differences between the two tissues for this simulation. Substituting Equations (2.4) and (2.5) into Equation (2.3) and optimizing the resulting equation for TR provides a relationship between the T_1 values for the two tissues and the repetition time that will create the highest CNR [15].

$$TR_{\mathrm{opt}} = \frac{T_{1A}T_{1B}}{T_{1B} - T_{1A}} \log\left(\frac{k_A T_{1B}}{k_B T_{1A}}\right) \tag{2.6}$$

With use of the field-dependent input parameters specified, model predictions for the minimum concentration of contrast agent required to generate visually apparent contrast between the two tissues may easily be determined [16]. As a point of reference, visually apparent contrast is typically defined as a $\mathrm{CNR} \geq 5:1$ [17].

2.3.2
Perfluorocarbon Nanoparticles for Enhancing Magnetic Resonance Contrast

For use as a T_1-weighted paramagnetic contrast agent, perfluorocarbon nanoparticles can be functionalized by surface incorporation of homing ligands and more than 50 000 gadolinium chelates (Gd^{3+}) per particle [7]. In addition, all of the paramagnetic ions are present in the outer aqueous phase to achieve maximum relaxivity of T_1 [18]. The result is a perfluorocarbon nanoparticle that is capable of overcoming the diluting partial volume effects that plague most magnetic resonance contrast agents [19]. The efficiency of an magnetic resonance contrast agent can be described by its relaxivity ($\mathrm{m}M^{-1}\,\mathrm{s}^{-1}$), which is simply calculated as the change in relaxation rate ($1/T_1$ or $1/T_2$) divided by the concentration of the contrast agent. The relaxivity

of Gd^{3+} in saline (4.5 $mM^{-1}s^{-1}$) [20] is lower when compared to Gd^{3+} bound to the surface of a PFC nanoparticle (33.7 $mM^{-1}s^{-1}$) [18] at a field strength of 1.5 Tesla. Considering that each nanoparticle carries approximately 50 000 to 100 000 Gd^{3+}, the 'particle' relaxivity has been measured at over 2 000 000 $mM^{-1}s^{-1}$ [18]. The high level of relaxivity achieved using this paramagnetic liquid PFC nanoparticle allows for the detection and quantification of nanoparticle concentrations as low as 100 picomolar, with a CNR of 5 : 1 [16].

2.3.3
Perfluorocarbon Nanoparticles for Fluorine (^{19}F) Imaging and Spectroscopy

The intensity of a magnetic resonance signal is directly proportional to the gyromagnetic ratio (γ) and the number of nuclei in the volume of interest [21]. Although there are seven medically relevant nuclei, the ^{1}H proton is the most commonly imaged nuclei in clinical practice because of its high γ and natural abundance. The isotopes, their γ-values, natural abundance and relative sensitivity compared to ^{1}H with a constant field are listed in Table 2.1. With a gyromagnetic ratio second only to ^{1}H and a natural abundance of 100%, ^{19}F is an attractive nucleus for MRI [22]. Its sensitivity is 83% (when compared to ^{1}H) at a constant field strength and with an equivalent number of nuclei. In biological tissue, low ^{19}F concentrations (in the range of micromoles) makes MRI impractical at clinically relevant field strengths without ^{19}F-specific contrast agents [23]. Perfluorocarbon nanoparticles are 98% perfluorocarbon by volume, which for perfluoro-octylbromide (1.98 g ml^{-1}, 498.97 g mol^{-1}) equates to an approximately 100 M concentration of fluorine within a nanoparticle. The paucity of endogenous fluorine in biological tissue allows the use of exogenous PFC nanoparticles as an effective ^{19}F MR contrast agent, without any interference from significant background signal. When combined with local drug delivery, detection of the ^{19}F signal serves as a highly specific marker for the presence of nanoparticles that would permit the quantitative assessment of drug dosing.

^{19}F has seven outer-shell electrons rather than a single electron (as is the case for hydrogen); as a result, the range and sensitivity to the details of the local environment of chemical shifts are much higher for fluorine than hydrogen. Consequently,

Table 2.1 Medically relevant MRI nuclei.

Isotope	γ (MHz T^{-1})	Natural abundance (%)	Relative sensitivity
^{1}H	42.58	99.98	1.00
^{19}F	40.05	100	0.83
^{23}Na	11.26	100	0.093
^{31}P	17.24	100	0.066
^{13}C	10.71	1.11	0.015
^{2}H	6.54	0.015	0.0097
^{15}N	3.08	4.31	0.0010

distinct spectra from different PFC species can be obtained and utilized for simultaneous targeting of multiple biochemical markers.

For use as a clinically applicable contrast agent, the biocompatibility of PFC nanoparticles must be considered. Liquid PFCs were first developed for use as a blood substitute [24], and no toxicity, carcinogenicity, mutagenicity or teratogenic effects have been reported for pure fluorocarbons within the 460 to 520 molecular-weight range. Perfluorocarbons, which inert biologically, are removed via the macrophage phagocytic system and excreted primarily through the lungs and in small amounts through the skin, as a consequence of their high vapor pressure relative to their mass [25]. The tissue half-lives of PFCs range from 4 days for perfluoro-octylbromide up to 65 days for perfluorotripropylamine. The prolonged systemic half-life of PFC nanoparticles, in conjunction with the local concentrating effect produced by ligand-directed binding, permits ^{19}F spectroscopy and imaging studies to be conducted at clinically relevant magnetic field strengths.

2.3.4
Fibrin-Imaging for the Detection of Unstable Plaque and Thrombus

Of the over 720 000 cardiac-related deaths that occur each year in the United States, approximately 63% are classified as sudden cardiac death [26]. Unfortunately, for the majority of patients, this is the first and only symptom of their atherosclerotic heart disease [27]. Atherosclerosis manifests initially as a fatty streak but, without proper treatment, it can progress to a vulnerable plaque that is characterized by a large lipid core, a thin fibrous cap and macrophage infiltrates [28]. These vulnerable plaques are prone to rupture, which can lead to thrombosis, vascular occlusion and subsequent myocardial infarction [29] or stroke. Routine angiography is the most common method of diagnosing atherosclerotic heart disease, with the identification of high-grade lesions (>70% stenosis) being referred for immediate therapeutic intervention. Ironically, most ruptured plaques originate from coronary lesions classified as nonstenotic [28]. Even nuclear and ultrasound-based stress tests are only designed to detect flow-limiting lesions. Because the most common source of thromboembolism comes from atherosclerotic plaques with 50–60% stenosis [30], diagnosis by traditional techniques remains elusive. In addition, there appears to be a 'window of opportunity' that exists between the detection of a vulnerable or ruptured plaque and acute myocardial infarction (measured in a few days to months) [31], when intervention could prove to be beneficial.

The acoustic enhancement of thrombi using fibrin-targeted nanoparticles was first demonstrated *in vitro* as well as *in vivo* in a canine model at frequencies typically used in clinical transcutaneous scanning [32]. The detection of thrombi was later expanded to MRI in a study by Flacke *et al.* [7] Fibrin clots were targeted *in vitro* with paramagnetic nanoparticles and imaged using typical low-resolution T_1-weighted proton imaging protocols with a field strength of 1.5 Tesla. Low-resolution images show the effect of increasing the amount of Gd^{3+} incorporated in the nanoparticles: a higher gadolinium loading results in brighter T_1 signals from the fibrin-bound PFC nanoparticles (Figure 2.3). In the same study, *in vivo* MR images were obtained of

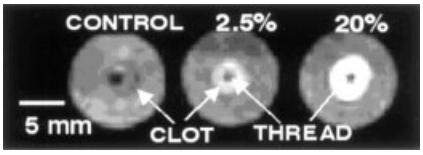


Figure 2.3 Low-resolution images (three-dimensional, T_1-weighted) of control and fibrin-targeted clot with paramagnetic nanoparticles presenting a homogeneous, T_1-weighted enhancement.

fibrin clots in the external jugular vein of dogs. Enhancement with fibrin-targeted PFC nanoparticles produced a high signal intensity in treated clots (1780 ± 327), whereas the control clot exhibited a signal intensity (815 ± 41) similar to that of the adjacent muscle (768 ± 47).

This method was extended to the detection of ruptured plaque in human carotid artery endarterectomy specimens resected from a symptomatic patient (Figure 2.4). Fibrin depositions ('hot spots') were localized to microfissures in the shoulders of the ruptured plaque in the targeted vessel (where fibrin was deposited), but this was not appreciated in the control. Further investigation towards the molecular imaging of small quantities of fibrin in ruptured plaque may someday detect this silent pathology sooner in order to pre-empt stroke or myocardial infarction.

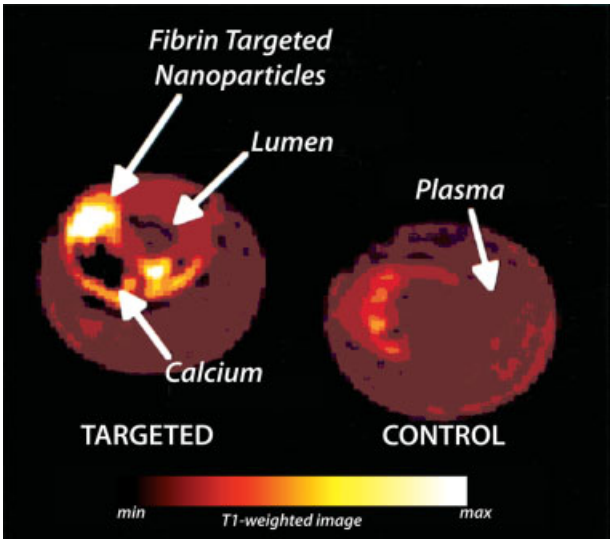


Figure 2.4 Color-enhanced magnetic resonance imaging of fibrin-targeted and control carotid endarterectomy specimens, revealing contrast enhancement (white) of a small fibrin deposit on a symptomatic ruptured plaque. The black area shows a calcium deposit. Three-dimensional, fat-suppressed, T_1-weighted fast gradient echo.

The high fluorine content of fibrin-targeted PFC nanoparticles, as well as the lack of background signal, can also be exploited for ^{19}F MRI and spectroscopy. In a recent study conducted by Morawski *et al.*, several methods were described for quantifying the number of nanoparticles bound to a fibrin clot using the ^{19}F signal [16]. First, fibrin-targeted paramagnetic perfluoro-crown-ether nanoparticles and trichlorofluoromethane ^{19}F spectra were obtained (Figure 2.5a). The relative crown ether signal intensity (with respect to the trichlorofluoromethane peak) from known emulsion volumes provided a calibration curve for nanoparticle quantification (Figure 2.5b). The perfluorocarbon (crown ether) nanoparticles then were mixed in titrated ratios with fibrin-targeted nanoparticles containing safflower oil and bound to plasma clots *in vitro*. As the competing amount of nonsignaling safflower-oil agent was increased, there was a linear decrease in the ^{19}F and Gd^{3+} signal. The number of bound nanoparticles was calculated from the ^{19}F signal and the calibration curve described above, and compared with mass of Gd^{3+} as determined by neutron activation analysis. As expected, there was excellent agreement between measured Gd^{3+} mass and number of bound nanoparticles (calculated from the ^{19}F signal) (Figure 2.5c).

In addition, clots were treated with fibrin-targeted nanoparticles containing either of two distinct PFC cores: crown ether and perfluoro-octyl bromide (PFOB) [33]. These exhibited two distinct ^{19}F spectra at a field strength of 4.7 Tesla, and the signal from the sample was highly related to the ratio of PFOB and crown ether emulsion applied. These findings demonstrated the possibility of simultaneous imaging and quantification of two separate populations of nanoparticles, and hence two distinct biomarkers.

These quantification techniques were applied to the analysis of human carotid endarterectomy samples (see Figure 2.6). An optical image of the carotid reveals extensive plaques, wall thickening and luminal irregularities. Multislice ^{19}F images showed high levels of signal enhancement along the luminal surface due to the binding of targeted paramagnetic nanoparticles to fibrin deposits (not shown in Figure 2.6). The ^{19}F projection images of the artery, taken over approximately 5 min, showed an asymmetric distribution of fibrin-targeted nanoparticles around the vessel wall, corroborating the signal enhancement observed with ^{1}H MRI. Concomitant visualization of ^{1}H and ^{19}F images would permit the visualization of anatomical and pathological information in a single image. In theory, the atherosclerotic plaque burden could be visualized with paramagnetic PFC contrast-enhanced ^{1}H images, while ^{19}F could be used localize identify plaques with high levels of fibrin and thus prone to rupture.

2.3.5
Detection of Angiogenesis and Vascular Injury

As described previously, ligand-directed PFC nanoparticles are well suited to the detection of very sparse biomarkers, such as integrins involved in the process of *angiogenesis* [12, 33, 34]. Although angiogenesis is a critical physiological process in wound healing, inflammation and organ development, it also contributes to the

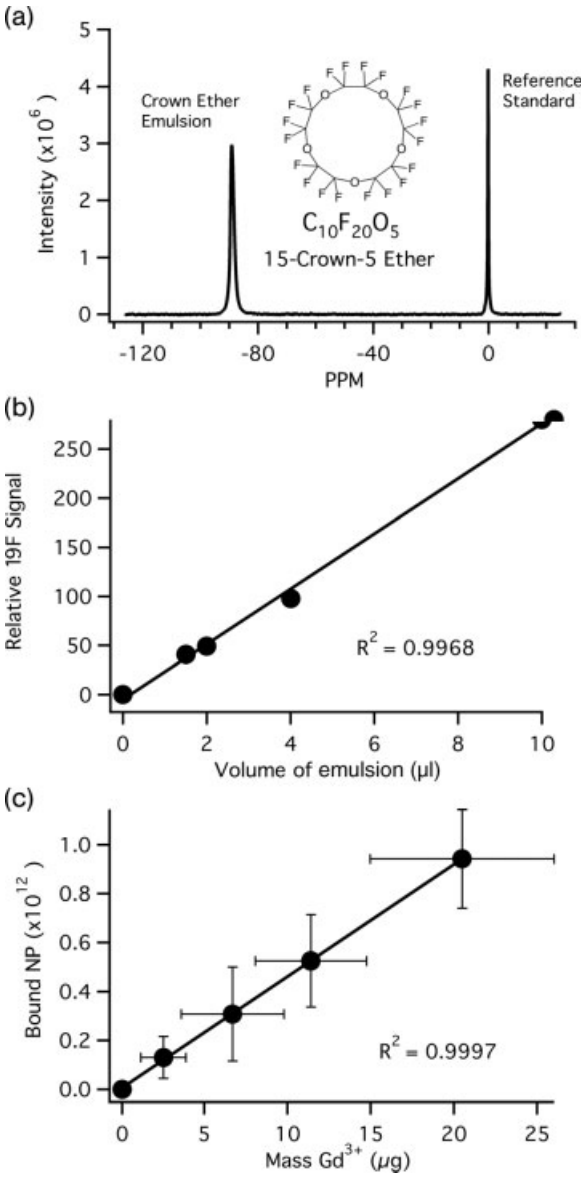


Figure 2.5 (a) Representative spectrum, taken at a field strength of 4.7 T, showing crown ether emulsion (~90 ppm) and trichlorofluoromethane (0 ppm) references; (b) The calibration curve for the crown ether emulsion has a slope of 28.06 with an R^2 of 0.9968; (c) Number of bound nanoparticles (mean ± SE) as calculated from ^{19}F spectroscopy versus the mass of total gadolinium (Gd^{3+}) in the sample as determined by neutron activation analysis, showing excellent agreement as independent measures of fibrin-targeted nanoparticles binding to clots. The linear regression line has an R^2 of 0.9997.

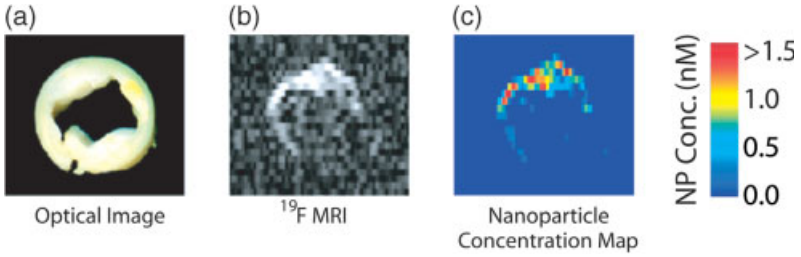


Figure 2.6 (a) Optical image of a 5 mm cross-section of a human carotid endarterectomy sample. The section showed moderate lumenal narrowing, and several atherosclerotic lesions; (b) A ^{19}F projection image acquired at 4.7 T through the entire carotid artery sample, showing a high signal along the lumen due to nanoparticles bound to fibrin; (c) Concentration map of bound nanoparticles in the carotid sample.

pathology of many disease processes such as diabetic retinopathy, rheumatoid arthritis, cancer and atherosclerosis. The process of angiogenesis depends on the adhesive interactions of vascular cells, and the integrin $\alpha_v\beta_3$ has been identified as playing a vital role in angiogenic vascular tissue. The functions of integrin $\alpha_v\beta_3$ includes vascular cell apoptosis (i.e. cell death), smooth muscle cell (SMC) migration and proliferation, and vascular remodeling [35]. The integrin is expressed on the luminal surface of activated endothelial cells, but not on mature quiescent cells. These findings support the fact that the role of $\alpha_v\beta_3$ in pathological conditions characterized by neovascularization may be an important diagnostic and therapeutic target. In fact, the use of a monoclonal antibody against $\alpha_v\beta_3$ has demonstrated an inhibition of angiogenesis, without affecting mature vessels [36]. Although $\alpha_v\beta_3$ integrins are expressed on other mature cells, such as SMCs, tissue macrophages, and on neovascular tissues in gut and developing bone, the PFC-based nanoparticles are too large to escape the normal vasculature and bind in sufficient quantities to create a detectable signal.

Perfluorocarbon nanoparticles have been developed to detect the sparse expression of the $\alpha_v\beta_3$ integrin on the neovasculature, and to deliver anti-angiogenic therapy (Figure 2.7) [37]. This approach has been used to visualize tumor-related angiogenesis in New Zealand White rabbits bearing Vx-2 tumors (<1.0 cm) using a 1.5 Tesla field [38]. MRI signals monitored at 2 h post-injection of $\alpha_v\beta_3$-targeted nanoparticles showed an enhancement of 126%, predominantly in an asymmetrical distribution along the tumor border. These results were consistent with the immunohistochemical staining results. Moreover, *in vivo* competitive blocking with $\alpha_v\beta_3$-targeted nonparamagnetic nanoparticles resulted in a decreased signal enhancement to a level attributable to local extravasation.

In a similar study, athymic nude mice bearing human melanoma tumors (C32, ATCC; 33 mm^3) were injected with $\alpha_v\beta_3$-targeted PFC nanoparticles and imaged at 2 h [39]. MR enhancement was apparent within 30 min and had increased by 173% at 2 h (Figure 2.8). Again, MRI results were correlated with

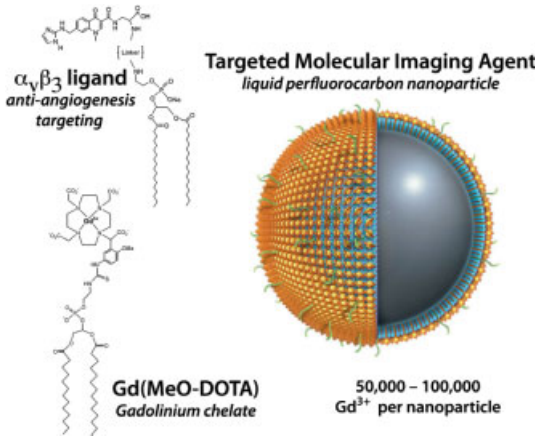


Figure 2.7 Schematic depicting the $\alpha_v\beta_3$ targeting aspect and the Gd^{3+} component that are incorporated into the lipid shell of the liquid perfluorocarbon nanoparticle agent.

histological results. In both studies, *in vivo* competitive blocking with $\alpha_v\beta_3$-targeted nonparamagnetic nanoparticles showed a 50% decrease of signal enhancement. These findings demonstrated the high specificity achievable with $\alpha_v\beta_3$-targeted nanoparticles.

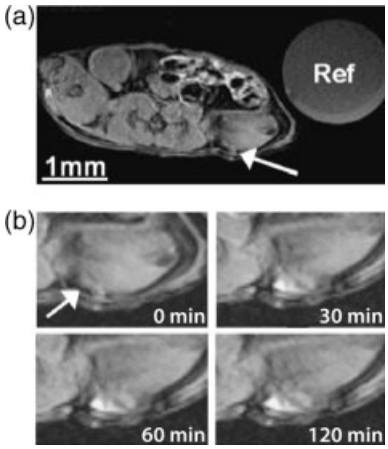


Figure 2.8 (a) T_1-weighted MR image (axial view) of an athymic nude mouse before injection of paramagnetic $\alpha_v\beta_3$-targeted nanoparticles. The arrow indicates a C-32 tumor that is difficult to detect. The reference (Ref) is Gd^{3+} in a 10 ml syringe; (b) Enlarged section of an MR image showing T_1-weighted signal enhancement of angiogenic vasculature of early tumors over 2 h, as detected by $\alpha_v\beta_3$-targeted nanoparticles.

2.4
Perfluorocarbon Nanoparticles as an Ultrasound Contrast Agent

Using the bubbles produced by agitating saline, Gramiak and Shah introduced the concept of an ultrasonic contrast agent in 1968 [40]. Today, commercially available ultrasound contrast agents are based on gas-filled encapsulated microbubbles (average diameter 2–5 μm) that transiently enhance the blood pool signal, which is otherwise weakly echogenic. When insonified by an ultrasound wave, microbubbles improve the gray scale images and Doppler signal via three distinct mechanisms [41–43]. First, at lower acoustic power, microbubbles are highly efficient scatterers due to their large differences in acoustic impedance ($Z = \rho c$, where ρ is the mass density and c is the speed of sound) compared to the surrounding tissue or blood [44]. With increasing acoustic energies, microbubbles begin nonlinear oscillations and emit harmonics of the fundamental (incident) frequency, thus behaving as a source of sound, rather than as a passive reflector [44, 46]. As biological tissue does not display this degree of harmonics, the contrast signal can be exploited to preferentially image microbubbles and improve signal-to-noise ratios (SNRs).

At even higher acoustic power, the destruction of microbubbles occurs allowing the release of free gas bubbles. Although not desirable for most forms of imaging, this results in a strong but transient scattering effect and provides the most sensitive detection of microbubbles. To emphasize these strong echogenic properties, it has been shown that even one microbubble can be detected with medical ultrasound systems [47]. Interestingly, the destruction and cavitation of microbubbles by ultrasound waves have been shown to facilitate drug delivery by 'sonoporating' membranes and allowing drugs and gene therapy to enter the cell [48, 49]. When this process occurs in capillary beds, the permeability increases allowing a subset of particles access to surrounding tissue for further drug deposition [50].

The wide use of microbubbles in everyday clinical applications highlight its effectiveness as a blood pool agent [45]. For example, microbubbles enhance the blood–tissue boundary of the left ventricular cavity, allowing for better diagnostic yield in resting as well as stress echocardiograms [51]. Improved Doppler signals are beneficial in the diagnosis of valvular stenosis and regurgitation [52]. Additionally, microbubbles are removed from the circulation via the macrophage phagocytic system and accumulate in the liver and spleen – that is, passive targeting. This mechanism can be employed for the detection of focal liver lesions and malignancies [48, 53]. When used as targeted contrast agents, microbubbles have been conjugated with ligands for a variety of vascular biomarkers including integrins expressed during angiogenesis, the glycoprotein IIb/IIIa receptor on activated platelets in clots, and L-selectin for the selective enhancement of peripheral lymph nodes, *in vivo* [54–56]. One disadvantage associated with the targeting of microbubbles is the 'tethering' of these particles to a surface. This interaction with a solid structure limits the ability of insonified microbubbles to oscillate, and dampens its echogenicity.

Unlike microbubble formulations that are naturally echogenic, liquid PFC nanoparticles have a weaker inherent acoustic reflectivity, and suspensions of them

have been shown to exhibit backscattering levels 30 dB below that of whole blood [57]. However, when collective deposition occur on the surfaces of tissues or a cell in a layering effect, these particles create a local acoustic impedance mismatch that produces a strong ultrasound signal, without any concomitant increase in the background level [58]. The echogenicity of nanoparticles does not depend upon the generation of harmonics, and therefore is not affected by binding with molecular epitopes. Due to their small size and inherent *in vivo* stability, PFC nanoparticle emulsions have a long circulatory half-life compared to microbubble contrast agents. This is accomplished without modification of their outer lipid surfaces with PEG or the incorporation of polymerized lipids, which may detract from the targeting efficacy. Acquired data have suggested that the PFC nanoparticles remain bound to the tissues for up to 24 h. In additionally, nongaseous PFOB-filled nanoparticles neither easily deform nor cavitate with ultrasound imaging.

The successful detection of cancer *in vivo* depends on a variety of factors when using molecularly targeted contrast agents. The number of epitopes to which the ligand can bind must be sufficient to allow enough of the contrast agent to accumulate for detection, while the ligand specificity must be maintained to ensure that nonspecific binding remains negligible. As stated above, the background signal from unbound, circulating contrast agent is low enough (or even absent) so as to not interfere with the assessment of bound, targeted agent. Previous studies have already demonstrated the use of high-frequency ultrasound in epitope-rich pathologies, such as fibrin in thrombus, where targeted PFC nanoparticles can act as a suitable molecular imaging agent by modifying the acoustic impedance on the surface to which they bind in a configuration that is well-approximated by a reflective layer [59]. However, at lower frequencies and for sparse molecular epitopes, in the typically tortuous vascular bed associated with the advancing front of a growing tumor, the clear delineation between nontargeted normal tissue and angiogenic vessels remains a challenge. The imaging technology itself must be highly sensitive and capable of detecting and/or quantifying the level of contrast agent bound to the pathological tissue. In clinical ultrasonic imaging, the sensitivity of detection depends on a physical difference in the way sound interacts with a surface covered by targeted contrast agent versus one that is not. The data presented below show that, in many cases, the sensitivity of this determination can be improved by applying novel and specific signal-processing techniques based on thermodynamic or information-theoretic analogues.

Site-targeted nanoparticle contrast agents, when bound to the appropriate receptor, must be detected in the presence of bright echoes returned from the surrounding tissue. One approach to the challenge of detecting the acoustic signature of site-specific contrast is through the use of novel signal receivers (i.e. mathematical operations that reduce an entire RF waveform, or a portion of it, to a single number) based on information-theoretic quantities, such as Shannon entropy (H), or its counterpart for continuous signal (H_f). These receivers have been shown to be sensitive to diffuse, low-amplitude features of the signal that often are obscured by noise, or else are lost in large specular echoes and, hence, not usually perceivable by a human observer [60–64].

Although entropy-based techniques have a long history in image processing for image enhancement and the post-processing of reconstructed images, the approach we take is different in that entropy is used directly as the quantity defining the pixel values in the image. Specifically, images are reconstructed by computing the entropy (or a limiting form of it: H_f) of segments of the individual RF A-lines that comprise a typical medical image by applying a 'moving window' or 'box-car' analysis. The computation of an entropy value for each location within an image is therefore possible, and the results can be superimposed over the conventional grayscale image as a parametric map.

2.4.1 Entropy-Based Approach

Radiofrequency data are obtained by sampling a continuous function $y = f(t)$. For an 8-bit digitizer, the sampled waveform is quantized into 256 (2^8) different levels. If we compute the probability, p_k of the k^{th} digitizer value appearing in the digital waveform, then we may compute the Shannon entropy of the resultant probability distribution

$$H - \sum_{k=0}^{255} p_k \log(p_k). \tag{2.7}$$

While this quantity has demonstrated utility for signal characterization [60], it also has the undesirable feature that it depends critically on the attributes of the digitizer used to acquire the data. This dependence may be removed by taking the limit where the sampling rate and dynamic range are taken to infinity [61, 62]. In that case, the probabilities, p_k, are replaced by density function, $wf(y)$, of the signal $f(t)$. While the Shannon entropy H_S becomes infinite in this limit, we may extract a finite portion of it, called H_f, that is also useful for signal characterization.

This well-behaved quantity can be expressed as

$$H_f = \int_{f\min}^{f_{\max}} w_f(y) \log w_f(y) dy. \tag{2.8}$$

This quantity has been shown to be very sensitive to local changes in backscattered ultrasound that arise from the accumulation of targeted nanoparticles in the acoustic field of view [4–6, 57, 65]. In contrast to most methods used to construct medical images, the waveform $f(t)$ does not directly enter the expression used to compute pixel values. Instead, the density function of the waveform is used.

2.4.2 The Density Function $w_f(y)$

The density function $w_f(y)$ corresponds to the density functions that are the primary mathematical objects in statistical signal processing and the description from which other mathematical quantities are subsequently derived (e.g. mean values, variances, covariances) [66–68]. In that setting, the density function constitutes the most

fundamental unit of information that the experimentalist has about a measured variable. It is important to note that the density function $w_f(\gamma)$ may be used to compute not only the entropy H_f but also the signal energy E_f, and hence all conventional energy-based signal analysis may be placed within this same mathematical framework.

Without loss of generality, we may adopt the convention that the domain of $f(t)$ is over the unit interval [0,1], then $w_f(\gamma)$, the density function of $f(t)$, can be defined by the basic integral relationship

$$\int_0^1 \phi(f(t))dt = \int_{f\min}^{f\max} \phi(\gamma)\, w_f(\gamma)d\gamma \tag{2.9}$$

for any continuous function $\phi(\gamma)$. This should be compared with the expression for the expectation value of a function ϕ of a random variable X with density $p_X(x)$, which is given by

$$\int \phi(x)_{p_X}(x)dx,$$

which explains why $w_f(\gamma)$ is referred to as the density function for $f(t)$ [69]. If we chose $\phi(x) = x^2$, then

$$\int_0^1 f(t)^2 dt = \int_{f\min}^{f\max} \gamma^2 w_f(\gamma)d\gamma, \tag{2.10}$$

an expression which represents the signal energy.

Many applications of either probability or information theory to signal processing proceed, usually very early in the discussion, by an *a priori* assumption of a specific underlying density function [70–72]. In contrast, the analysis steps detailed for H_f begin with the measured time-domain waveform data and proceed to calculate the density functions without imposing any additional assumptions.

2.4.3
Ultrasound in a Precancerous Animal Model

The capabilities of entropy-based signal processing for the acoustic detection of nanoparticles targeted to neovasculature has been demonstrated in several animal models [5, 6, 57]. One relevant example was obtained using the transgenic K14-HPV16 mouse [6]. This animal model contains human papilloma virus (HPV)-16 oncoproteins driven by a keratin promoter, so that lesions develop in the skin. Typically, the ears exhibit squamous metaplasia, a precancerous condition, associated with abundant neovasculature that expresses the $\alpha_v\beta_3$ integrin. Eight of these transgenic mice [73, 74] were treated intravenously with 1.0 mg kg^{-1} of either $\alpha_v\beta_3$ integrin-targeted nanoparticles ($n = 4$) or untargeted nanoparticles ($n = 4$), and subsequently imaged subsequently using a research ultrasound system (Vevo 660; Visualsonics, Toronto, Canada). Imaging was accomplished with the mouse ears positioned in the focal zone of a 40 MHz single-element 'wobbler' sector-scan

transducer, with an F-number of 2 (diameter 3 mm, focal length 6 mm). Radio-frequency ultrasonic backscatter waveforms corresponding to a region 80 mm wide × 30 mm deep, were digitized at time points 0, 15, 30 and 60 min after administration of the nanoparticle contrast agent. All of these RF data were processed off-line to reconstruct images using information theoretic (entropy-based) and conventional (energy-based) receivers. Image segmentation was performed automatically using the threshold, which excluded 93% of the area under the composite histogram for all data sets. The mean value of segmented pixels was computed at each time point post injection.

A diagram depicting the placement of transducer, gel standoff and mouse ear is shown in the left side of Figure 2.9, together with a representative B-mode grayscale image (i.e. logarithm of the analytic signal magnitude). The labels indicate the location of the skin (top of image insert), the structural cartilage in the middle of the ear, and a short distance below this, the echo from the skin at the bottom of the ear. To the right of this ultrasound image, is an histological view of a HPV mouse ear that has been magnified 20-fold to permit a better assessment of the thickness and architecture of the sites where the $\alpha_v\beta_3$ integrin-targeted nanoparticle might attach (red by β_3 staining). Both, the skin and tumor are both visible in the image. On either side of the cartilage (center band in image), extending to the dermal–epidermal junction, is the stroma, which is filled with neoangiogenic microvessels. These microvessels are also decorated with targeted $\alpha_v\beta_3$ nanoparticles, as indicated by the fluorescent image (labeled, upper right of Figure 2.9) of a bisected ear from an $\alpha_v\beta_3$-injected K14-HPV16 transgenic mouse. It is in this region that the $\alpha_v\beta_3$-targeted nanoparticles are expected to accumulate, as indicated by the presence of red β_3 stain in the magnified image of an immunohistological specimen also shown in the image.

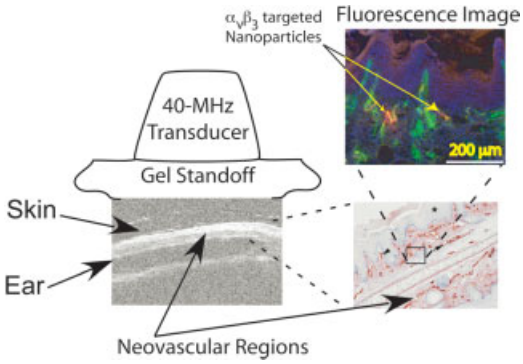


Figure 2.9 Close-up of transducer, standoff and B-mode image of ear, with an enlarged histological view showing the location of the binding sites, and a fluorescent image from the same anatomical region. This shows that the $\alpha_v\beta_3$-targeted nanoparticles accumulate in this portion of the mouse ear.

2.4.3.1 Image Analysis

For this study, in which the same portion of the anatomy was imaged at successive intervals, the objective was to quantify changes in image features as a function of time. The first step in this process was the creation of a composite image from the images obtained at 0 through 60 min. Next, an estimate of the probability density function (PDF) of this composite image was computed by normalizing the pixel value histogram to have unit area. It must be emphasized that this function is not related to the density functions $w_f(\gamma)$ as were defined in Equation (2.8). Rather, it is a calculational device used to objectively segment H_f and $\log[E_f]$ images into 'enhanced' and 'unenhanced' regions. A typical histogram is shown in Figure 2.10. The first, larger maxima, corresponds to the relatively homogeneous gray background visible in most H_f images, while the smaller peak corresponds to tissue interfaces, which appear also as bright features in grayscale B-mode images, such as that shown in the inset of Figure 2.9 [76].

Several different methods of image segmentation based on the PDF were investigated. In all of these a specific value, or threshold, in the histogram was chosen and the images divided into two regions: (i) those having pixel values above the threshold (considered to be unenhanced); and (ii) those having pixel values below (referred to subsequently as enhanced pixels). The PDF of all composite images exhibited a two- peak structure with a large and small peak. Thresholds were set at the second minimum, and at the half-way point between the large and small peaks. The full width at half maximum (FWHM) was also computed, and thresholds set at: 4.5, 3.5, 3.25, 3, 2.75 and 2.5 FWHM below the large peak. Thresholds were also set at points such that 97, 95, 93, 90, 87 and 80% of the pixel values were above the threshold. After selection of a threshold value, regions of interest (ROI) were

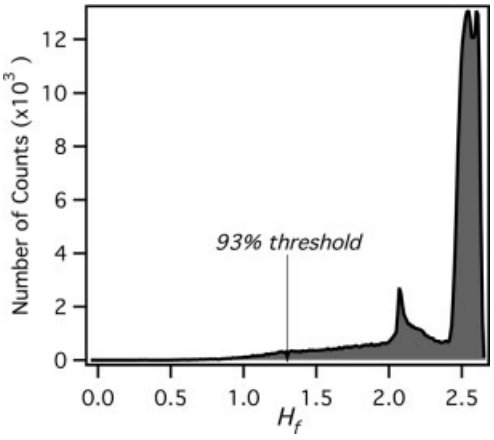


Figure 2.10 A histogram from the composite H_f images acquired at 0, 15, 30, 45 and 60 min post injection. The histogram has two peaks; these have been characteristically observed in several studies using different equipment and animal models. The 93% threshold level used is indicated by the arrow.

selected using NIH ImageJ (http://rsb.info.nih.gov/ij/), and the mean value of the pixels lying below the threshold were computed for each of the images acquired at 0, 15, 30, 45 and 60 min post injection. The mean value at zero minutes was subtracted from the values obtained for all subsequent times, to obtain a sequence of changes in receiver output as a function of time post injection. This was done for all four animals injected with targeted nanoparticles, and also for the four control animals. The sequences of relative changes were then averaged over the targeted and control groups to obtain a sequence of time points for change in receiver output for both groups of animals. The threshold of 93% was finally chosen as it produced the smallest *p*-value (0.00 043) for a *t*-test comparing the mean values of the ROI at 15 min as compared to 60 min. The corresponding *p*-value for the control group was 0.27.

The average change, with time after injection, of the mean value of the enhanced regions of H_f images obtained from all eight of the animals reported in the study are compared in Figure 2.11a [6]. As these data show, the mean value, or enhancements, obtained in the targeted group increased steadily with time. After 30 min the mean

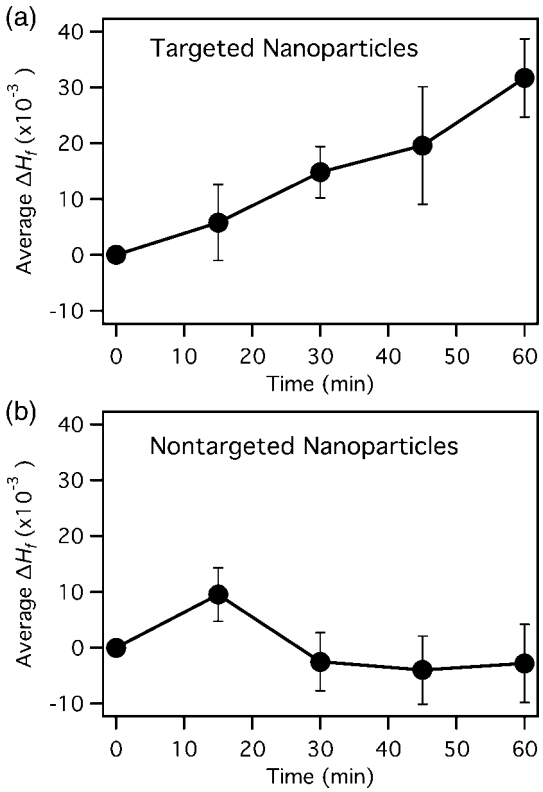


Figure 2.11 Plots of average enhancement obtained by analysis of H_f images from (a) four HPV mice injected with $\alpha_v\beta_3$-targeted nanoparticles, and (b) four HPV mice injected with nontargeted nanoparticles.

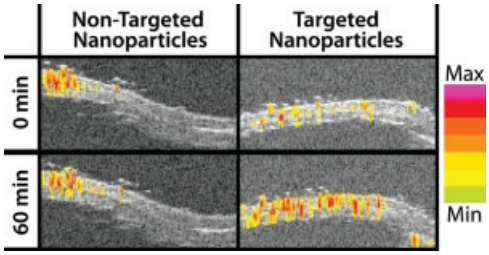


Figure 2.12 Cropped images of transverse cross-sections of HPV mouse ear showing of H_f-enhanced conventional images comparing the effects of targeting (right column) versus nontargeted (left column) both at 0 min (top row) and 60 min (bottom row) after injection.

value of enhancement was measurably different from baseline values ($p < 0.005$). Moreover, the values at 15 and 60 min were also statistically different ($p < 0.005$). The corresponding results obtained from control animals that were injected with nontargeted nanoparticles are shown in Figure 2.11b. There was no discernible trend in the group, and the last three time points were not statistically different from zero. A comparison of the enhancement measured at 15 and 60 min yielded a p-value ≈ 0.27. The enhancement in H_f observed after 60 min for representative instances of targeted and nontargeted nanoparticles is shown in Figure 2.12. These images were generated by overlaying a 93% thresholded version of the H_f using a look-up table (LUT) on top of the conventional grayscale B-mode image.

For comparison, and to illustrate the potential value of the entropy-based analysis, the corresponding results obtained using the $\log[E_f]$ analysis are shown in Figure 2.13a for same data used to generate Figure 2.11. These data were obtained by computing the mean value of pixels lying below the 93% threshold at each time point (0, 15, 30, 45 and 60 min) for each animal (four injected with targeted nanoparticles, and four with nontargeted nanoparticles), as discussed above. Unlike the entropy case, the values at 15 and 60 min were not statistically different ($p = 0.10$). Figure 2.13b shows the corresponding result obtained from the control group of animals that were injected with nontargeted nanoparticles. There was no discernible trend in the group, and the last three time points were not statistically different from zero.

2.4.4
Targeting of MDA-435 Tumors

In a separate investigation, nascent MDA 435 tumors implanted in athymic nude mice reproducibly stimulated neovascular growth and the expresses ion of $\alpha_v\beta_3$ integrin [75]. Human MDA 435 cancer cells were implanted, by injection, in the left hindquarters of athymic nude mice between 10 and 22 days prior to acquisition of the data. Figure 2.14 shows an immunohistologically stained (β_3 staining) section of an excised MDA 435 tumor. β_3 expression, a marker for $\alpha_v\beta_3$-integrin, was found in

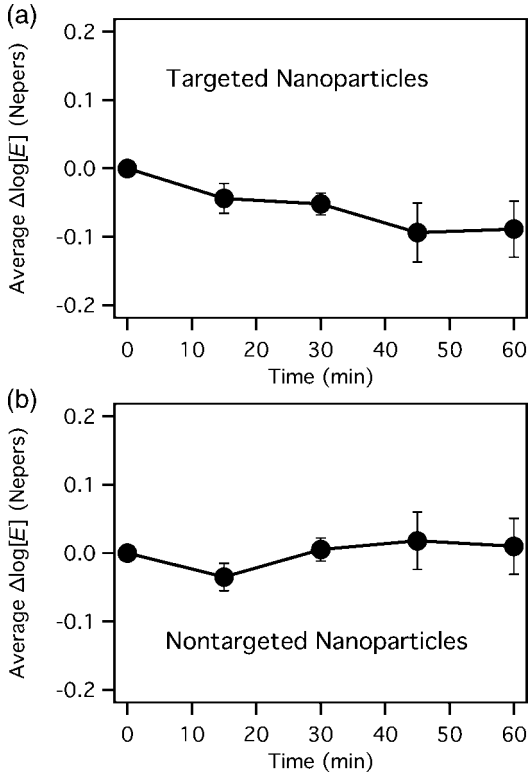


Figure 2.13 Plots of average enhancement obtained by analysis of log [E_f] images from (a) four HPV mice injected with $\alpha_v\beta_3$-targeted nanoparticles and (b) four control HPV mice given nontargeted nanoparticles.

abundance (red regions), although not exclusively, between the skin and the tumor capsule. The close proximity of these binding sites to the skin–transducer interface is one of the primary obstacles that must be overcome by any quantitative detection scheme intended to determine the extent of this region. Accordingly, the acoustic portion of the experiment was designed to maximize system sensitivity near this interface. This was carried out in order to maximize the opportunity to detect nanoparticles targeted towards angiogenic neovasculature. It also provided a stringent test of the H_f entropy-based metric's ability to separate signals near the confounding skin–tissue interface, which was one of the primary goals of this study.

Nine animals were injected with targeted nanoparticles, and seven with nontargeted nanoparticles to serve as controls. Each mouse was preanesthesitized with ketamine, after which an intravenous catheter was inserted into the right jugular vein to permit the injection of nanoparticles (either $\alpha_v\beta_3$-targeted or untargeted). The mouse was then placed on a heated platform maintained at 37 °C, and anesthesia administered continually with isoflurane gas through a nose cone.

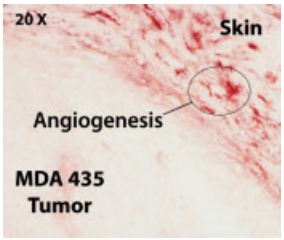


Figure 2.14 A histological specimen extracted from a MDA 435 mouse model, magnified 20-fold, to permit a better assessment of the thickness and architecture of the sites (red by β_3 staining) where $\alpha_v\beta_3$-targeted nanoparticle might attach. The skin and tumor are both visible in the image. The close proximity of neovasculature to the skin–transducer interface is one of the primary obstacles that must be overcome by any quantitative detection scheme intended to determine the extent of this region.

Subsequently, the mouse was injected with 0.030 ml of nanoparticle emulsion (equivalent to a whole-body dose of 1 mg kg^{-1} body mass). Ultrasound data were then acquired at 0, 15, 30, 60 and 120 min intervals.

No evidence of change was observed between the zero-minute and 120-min conventional grayscale B-mode images, while there was a slight (but nonsignificant) change with the color LUT (Figure 2.15). The lower part of Figure 2.15 shows images reconstructed from the same raw RF data, again, with a grayscale LUT in the top row and color LUT in the bottom row. Significant changes in the size of the brightest (red) region, located between the skin–tumor capsule boundary, were observed as expected. Unlike the conventional B-mode processing case, here the H_f processed data showed the region to comprise pixel values far brighter than the mean pixel value that occurred in the rest of the image.

A loss of spatial resolution between skin and tumor capsule was also observable in the image, as expected; this resulted from the smoothing effect of the moving window analysis. In view of this smoothing, which tends to reduce the variations in magnitude of a function, it was somewhat surprising that the image showed a greater separation between the background and enhanced regions. From the images, it could also be seen that the magnitude of values in the region between the skin and the tumor boundary increased with post-injection time. Moreover, the shape and location of the regions were consistent with a brightening effect due to an accumulation of nanoparticles in the angiogenic neovasculature [77, 78].

The data in Figure 2.16 compare the change in mean value of the 3% thresholded region (i.e. enhancement) for the B-mode images (logarithm of signal envelope) as a function of time post injection for controls versus targeted. The plots show that the conventional B-mode images cannot be used to distinguish between 0 and 120 min post injection.

Corresponding results, obtained using entropy imaging, are shown in Figure 2.17, using the same vertical and horizontal scales as in Figure 2.16, for ease of comparison. The plots show that only the entropy-based receiver was able to distinguish between 0 and 120 min post injection (paired t-test produces $p < 0.05$). Moreover,

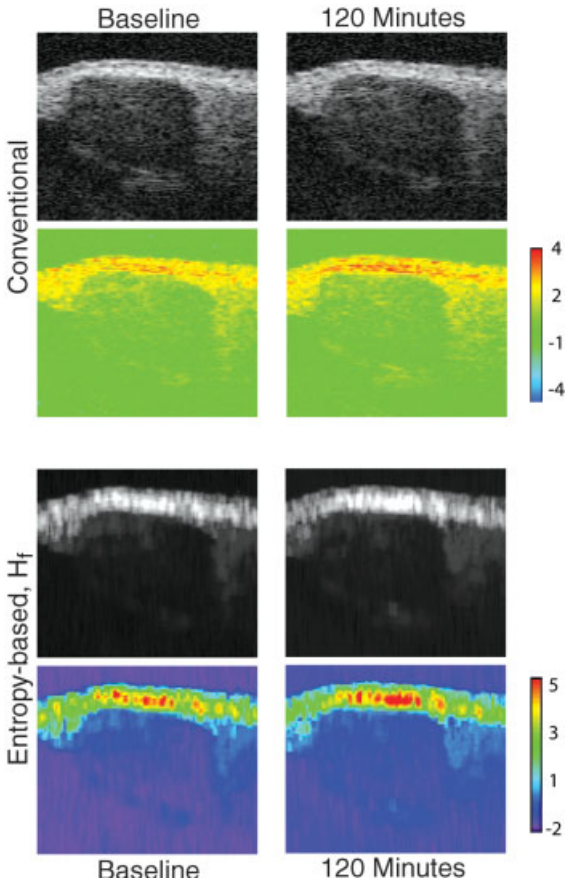


Figure 2.15 A comparison of 0-min and 120-min (post-injection) images obtained from conventional energy-based signal processing (upper part of figure) and the entropy-based H_f metrics (lower part of figure). Grayscale images of before and after data are presented for both types of signal processing. No change is evident in the B-mode images, and at most there is a slight change in the images. However, the application of a color look-up table (LUT) to the images revealed more detail. The same color LUT mapping was applied to both B-mode images to facilitate comparisons. The calibration bar of the mapping is shown to the right of the images in both cases. The images show a greater change with time and a greater separation of the neovascular region from the rest of the image than the B-mode images. Replication of this experiment in nine animals showed the change to be statistically significant only for the H_f-processed images.

the mean values increased in an approximately linear fashion versus time. The plot of control experiments showed there was no significant change in enhancement with time in these animals. However, a careful visual inspection of the image sequence revealed measurable changes in tumor shape and position that most likely were induced by respiration and relaxation of the animal over the 2-h experiment.

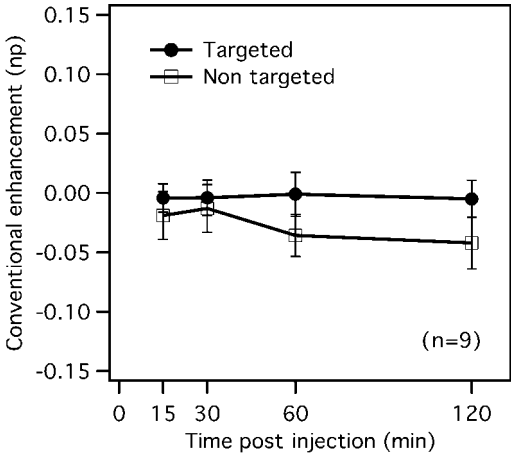


Figure 2.16 Quantitative comparison of enhancement and B-mode images obtained using targeted nanoparticles. No significant changes were observed with either untargeted or targeted nanoparticles. All plots have vertical units of nepers, as the image analysis was performed on noise-scaled or normalized images (which are unitless). As explained in the text, this does not alter the quantitative conclusion presented in these plots.

2.4.5
In Vivo Tumor Imaging at Clinical Frequencies

A separate ultrasound study utilizing targeted nanoparticles was designed to assess the feasibility of image-based angiogenic neovasculature using backscattered

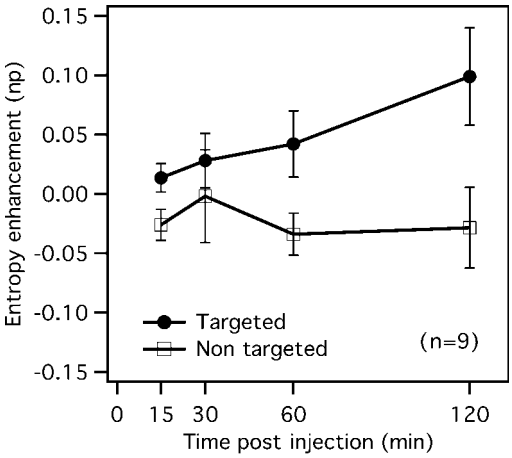


Figure 2.17 Quantitative comparison of enhancement for a thermodynamic receiver (H_f) obtained using targeted and untargeted nanoparticles. Only the targeted nanoparticles produced a significant change in enhancement ($p < 0.05$).

ultrasound in the frequency range between 7 and 15 MHz [79]. These investigators employed a liquid–PFC nanoparticle conjugation of an $\alpha_v\beta_3$ peptidomimetic to target the expression of $\alpha_v\beta_3$ in Vx-2 tumors implanted in the hindquarters of New Zealand White rabbits ($n = 9$). Anesthesia was administered continually with isoflurane gas, the model was injected with a whole-body dose of 0.66 ml kg^{-1} of nanoparticle emulsion, after which the ultrasound data were acquired at 0, 15, 30, 60 and 120 min. Six control rabbits were also imaged using the same methodology, but were not injected with nanoparticles. Beam-formed RF data were acquired using a modified research version of a clinical ultrasound system (Philips HDI-5000). Data were analyzed for all rabbits at all times post injection, using three different techniques: (i) conventional grayscale; (ii) H_f (an entropy-based quantity); and (iii) $\log[E_f]$ (i.e. the logarithm of the signal energy, E_f). Representative image data are shown in Figure 2.18, depicting the tumor in cross-section. A paired t-test comparing H_f image enhancement obtained at 0 and 120 min for the rabbits injected with targeted nanoparticles indicated a significant difference ($p < 0.005$). For control rabbits there was no significant difference between 0 and 120 min ($p = 0.54$). Conventional grayscale imaging at the fundamental frequency and $\log[E_f]$ imaging failed to detect a coherent signal, and did not show any systematic pattern of signal change.

2.5
Contact-Facilitated Drug Delivery and Radiation Forces

2.5.1
Primary and Secondary Radiation Forces

Acoustic radiation force is a phenomenon associated with the propagation of acoustic waves through a dissipative medium. It is caused by a transfer of momentum from the wave to the medium, arising either from absorption or reflection of the wave [80, 81]. For particles suspended in a liquid medium, these forces manifest themselves in two ways. The first way, which is referred to as the *primary radiation force*, tends to accelerate the suspended particles away from the source. The second way, referred to as the *secondary force*, is an interparticle force that can be completely attractive, if the particles lie in contours perpendicular to the incident field and can be completely repulsive if the particles are oriented parallel to the incident field. One very useful form for these primary and interparticle (secondary) forces is given by [82]

$$F_{\text{primary}} = \frac{V_0 P_A^2}{4\rho_0 c_0^2} k \sin(2k) f\left(\frac{\rho}{\rho_0}\right) \tag{2.11}$$

where

$$f\left(\frac{\rho}{\rho_0}\right) = \frac{\rho_0 c_0^2}{\rho c^2} - \frac{5\rho - 2\rho_0}{3\rho + \rho_0}, \tag{2.12}$$

Here, V_0 is the sphere volume, P_A is the acoustic pressure amplitude, ρ and ρ_0 are the sphere and fluid densities, respectively, and c and c_0 are the sound velocity in the

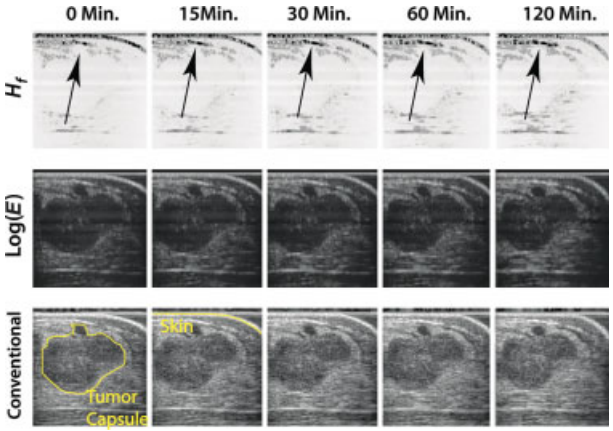


Figure 2.18 Images produced from beam-formed RF acquired from a rabbit injected with $\alpha_v\beta_3$-targeted nanoparticles. The three rows show composite images formed by the application of three different signal processing techniques: H_f, log[E_f] signal receiver (both applied with a moving window), and conventional image processing. Each composite image is comprised of five sub-images reconstructed from beam-formed RF acquired at 0, 15, 30, 60 and 120 min post injection, as indicated by the labels. Only the H_f composite image showed any evidence of change after injection (black arrows).

sphere and fluid, respectively. The interparticle force given by

$$F_{i.p.} = \xi r^{-4} f(\theta) \tag{2.13}$$

where

$$\xi = \frac{2\pi}{3}\frac{(\rho - \rho_0)^2}{\rho_0} a^3 b^3 u_0^2, \tag{2.14}$$

and a, b are the sphere radii, r is their separation distance, and u_0 is the velocity amplitude of the suspending medium.

The action of these forces *in vivo* is to concentrate the suspended nanoparticles and push them away from the acoustic source (i.e. away from the center of arterial flow and onto the capillary wall), as shown in Figure 2.19. This effect increases the potential to increase their therapeutic efficacy.

2.5.2 *In Vitro* Results

Besides detecting sparse epitopes for noninvasive imaging, PFC nanoparticles are capable of specifically and locally delivering drugs and other therapeutic agents through a novel process known as *contact-facilitated drug delivery* [12]. The direct transfer of lipids and drugs from the nanoparticles' surfactant layer to the cell membrane of the targeted cell is usually a slow and inefficient process. However, through ligand-directed targeting this process can be accelerated by minimizing the separation of the lipids and surfaces, and increasing the frequency and duration

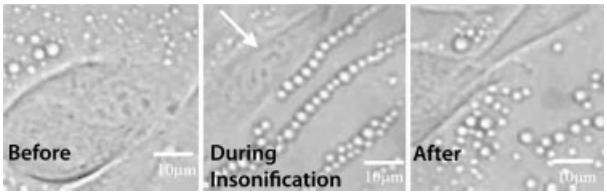


Figure 2.19 Still images of C32 melanoma cells with nanoparticles before, during and after insonification. These show primary and secondary radiation forces acting on the perfluorocarbon nanoparticles. The direction of acoustic insonification, as indicated by the arrow in the center panel, is the same in all three cases. Left panel: pre-insonification; the particles are arranged randomly. Center panel: during insonification, the particles line up on an axis perpendicular to the direction of insonification and move away from the source. Right panel: after insonification, the particle configuration is re-randomized by Brownian motion. During ultrasound exposure, the alignment of nanoparticles relative to the acoustic field (arrow) demonstrates conclusively that acoustic radiation forces (primary and secondary) influence the nanoparticles. This mechanism was observed to be a reversible and safe process; after ultrasound treatment the nanoparticles were no longer aligned, but had been neither destroyed nor altered.

of the lipid–surface interactions (see Figures 2.20 and 2.21). Spatial localization (via high-resolution ^{19}F-enhanced MRI) and quantification of the nanoparticles (via ^{19}F spectroscopy) permits the local therapeutic concentrations to be estimated. Thus, PFC nanoparticles can be used for detection, therapy and treatment monitoring.

As an example, *in vitro* vascular smooth muscle cells were treated with tissue factor-targeted PFC nanoparticles containing 0, 0.2 or 2.0 mol% doxorubicin or paclitaxel, or an equivalent amount of drug in buffer solution alone [83]. After targeting for only for 30 min, proliferation was inhibited for three days, while *in vitro* dissolution studies revealed that the nanoparticles' drug release persisted for more than one week. High-resolution MRI with a 4.7 Tesla field strength showed that the image intensity of the targeted vascular smooth muscle cells was twofold higher compared to nontargeted cells. In addition, the fluorine signal amplitude at 4.7 Tesla was unaffected by the presence of surface gadolinium, and was linearly

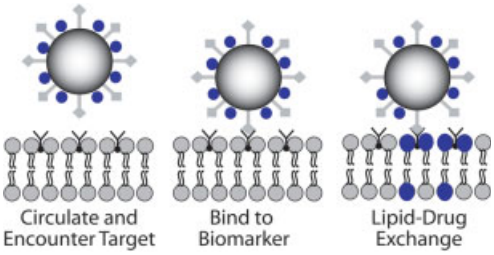


Figure 2.20 Schematic representation illustrating contact-facilitated drug delivery. The phospholipids and drugs within the nanoparticles surface-exchange with the lipids of the target membrane through a convection process, rather than diffusion, as is common among other targeted systems.

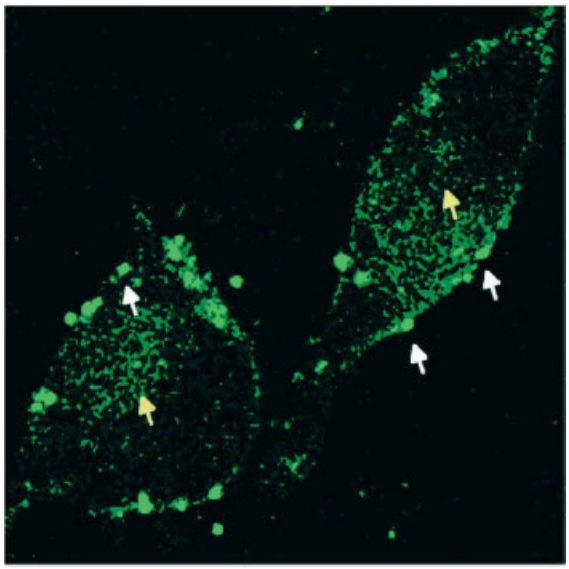


Figure 2.21 *In vitro* targeting of fluoroscein isothiocyanate (FITC)-labeled nanoparticles (white arrows) targeted to $\alpha_v\beta_3$ integrin expressed by C-32 melanoma cells. This illustrates the delivery of FITC-labeled surfactant lipids into target cell membranes (yellow arrows).

correlated to the PFC concentrations which, by direct inference could be related to the nanoparticles' number.

2.6 Conclusions

Targeted liquid PFC nanoparticles represent an extremely versatile platform which has been successfully employed in conjunction with ultrasound, single positron-emission tomography and MRI. These nanoparticles are capable of aiding in the detection of sparse biomarkers, such as integrins in angiogenesis, as well as high-density epitopes such as fibrin. They are unique in that they can be used to diagnose, treat and monitory therapy in a one-step process. Hence, their ongoing and future impact in the fields of cardiology and oncology are predicted to be substantial.

The field of engineered contrast agents continues to grow steadily and advance in line with the rapid developments in nanotechnology. At research centers worldwide, multidisciplinary teams have been assembled which combine expertise in the areas of physics, chemistry, biology, engineering and medicine, to focus on the challenges of creating this next generation of agents. Today, this field is progressing along a path that embraces the prediction summarized within the oft-quoted title of Richard Feynman's presentation, on the world of the nano-scale, to the American Physical Society on December 29th 1959: "There's plenty of room at the bottom."

The potential for significant contributions to paradigms of patient care have been reinforced in recent years via specific funding mechanisms from granting agencies, such as the US National Institutes of Health (NIH) initiatives creating the Programs of Excellence in Nanotechnology (NHLBI-PEN) and the Centers of Cancer Nanotechnology Excellence (NCI-CCNE). Currently, a number of agents are moving towards clinical trials under this aegis. In order to gain acceptance, the approval process by the US Food and Drug Administration is typically on the order of four to eight years for imaging agents (slightly longer for therapeutics). At the time of this writing, some agents have already advanced to Phase 1 and 2 clinical trials, and several new biotechnology start-up companies have also been launched that are devoted to the same goal.

References

1 Cyrus, T. (2006) *Nanomaterials for Cancer Therapy and Diagnosis* (eds S. Challa and S. Kumar), Wiley-VCH, Weinheim, p. 121.

2 Lanza, G.M., Lorenz, C.H., Fischer, S.E., Scott, M.J., Cacheris, W.P., Kaufmann, R.J., Gaffney, P.J. and Wickline, S.A. (1998) *Academic Radiology*, **5**, S173.

3 Marsh, J.N., Partlow, K.C., Abendschein, D.R., Scott, M.J., Lanza, G.M. and Wickline, S.A. (2007) *Ultrasound in Medicine and Biology*, **33**, 950.

4 Hughes, M.S., Marsh, J.N., Hall, C.S., Savery, D., Scott, M.J., Allen, J.S., Lacy, E.K., Carradine, C., Lanza, G.M. and Wickline, S.A. (2004) Proceedings IEEE Ultrasonics Symposium, 04CH37553, p. 1106.

5 Hughes, M.S., Marsh, J.N., Zhang, H., Woodson, A.K., Allen, J.S., Lacy, E.K., Carradine, C., Lanza, G.M. and Wickline, S.A. (2006) *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency, Control*, **53**, 1609.

6 Hughes, M.S., McCarthy, J., Marsh, J.N., Arbeit, J., Neumann, R., Fuhrhop, R., Wallace, K., Znidersic, D., Maurizi, B., Baldwin, S., Lanza, G.M. and Wickline, S.A. (2007) *Journal of the Acoustical Society of America*, **121**, 3542.

7 Flacke, S., Fischer, S., Scott, M.J., Fuhrhop, R.J., Allen, J.S., McLean, M., Winter, P., Sicard, G.A., Gaffney, P.J., Wickline, S.A. and Lanza, G.M. (2001) *Circulation*, **104**, 1280.

8 Moghimi, S. and Patel, H. (1989) *Biochimica et Biophysica Acta*, **984**, 384.

9 Harisinghani, M., Saini, S., Weisleder, R., Hahn, P., Yantiss, R., Tempany, C., Wood, B. and Mueller, P. (1999) *American Journal of Roentgenology*, **172**, 1347.

10 Keeler J. (2005) *Understanding NMR Spectroscopy*, J.W. Wiley & Sons, Chichester.

11 Slichter, Charles P. (1996) *Principles of Magnetic Resonance: Springer Series in Solid State Sciences*, Springer-Verlag, New York.

12 Lanza, G., Winter, P., Caruthers, S., Morawski, A., Schmieder, A., Crowder, K. and Wickline, S. (2004) *Journal of Nuclear Cardiology*, **11**, 733.

13 Nelson, K. and Runge, V. (1995) *Topics in Magnetic Resonance Imaging*, **7**, 124.

14 Kirsch, J.E. (1991) *Topics in Magnetic Resonance Imaging*, **3**, 1.

15 Ahrens, E., Rothbacher, U., Jacobs, R. and Fraser, S. (1998) *Proceedings of the National Academy of Sciences of the United States of America*, **5**, 8443.

16 Morawski, A.M., Winter, P.M., Crowder, K.C., Caruthers, S.D., Fuhrhop, R.W., Scott, M.J., Robertson, J.D., Abendschein, D.R., Lanza, G.M. and Wickline, S.A. (2004) *Magnetic Resonance in Medicine*, **51**, 480.

17 Rose, A. (1948) *Journal of the Optical Society of America*, **38**, 196.

18 Winter, P.M., Caruthers, S.D., Yu, X., Song, S.K., Chen, J.J., Miller, B., Bulte, J.W.M., Robertson, J.D., Gaffney, P.J., Wickline, S.A. and Lanza, G.M. (2003) *Magnetic Resonance in Medicine*, **50**, 411.

19 Gupta, H. and Weissleder, R. (1996) *Magnetic Resonance Imaging Clinics of North America*, **4**, 171.

20 Stanisz, G. and Henkelman, R. (2000) *Magnetic Resonance in Medicine*, **44**, 665.

21 Bushong, S. (2003) *Magnetic Resonance Imaging: Physical and Biological Principles*, 3rd edn, St. Louis, Mosby.

22 Longmaid, H., Adams, D., Neirinckx, R., Harrison, G., Brunner, C.P., Seltzer, S., Davis, M., Neuringer, L. and Geyer, R. (1985) *Investigative Radiology*, **20**, 141.

23 McFarland, E., Koutcher, J., Rosen, B., Teicher, B. and Brady, T. (1985) *Journal of Computer-Assisted Tomography*, **9**, 8.

24 Sloviter, H. and Mukherji, B. (1983) *Progress in Clinical and Biological Research*, **122**, 181.

25 Joseph, P., Yuasa, Y., Kundel, H., Mukherji, B. and Sloviter, H. (1985) *Investigative Radiology*, **20**, 504.

26 Zheng, Z., Croft, J., Giles, W. and Mensah, G. (2001) *Circulation*, **104**, 2158.

27 Kuller, L. and Lilienfeld, A. and Fisher, R. (1966) *Circulation*, **34**, 1056.

28 Naghavi, M., Libby, P., Erling, F., Casscells, S., Litovsky, S., Rumberger, J., Badimon, J., Stefanadis, C., Moreno, P. and Pasterkamp, G. (2003) *Circulation*, **108**, 1664.

29 Davies, M. and Thomas, A. (1985) *British Heart Journal*, **53**, 363.

30 Ambrose, J., Tannenbaum, M., Alexopoulos, D., Hjemdahl-Monsen, C., Leavy, J., Weiss, M., Borrico, S., Gorlin, R. and Fuster, V. 1988 *Journal of the American College of Cardiology* **12**, 56.

31 Ojio, S., Takatsu, H., Tanaka, T., Ueno, K., Yokoya, K., Matsubara, T., Suzuki, T., Watanabe, S., Morita, N., Kawasaki, M., Nagano, T., Nishio, I., Sakai, K., Nishigaki, K., Takemura, G., Noda, T., Minatoguchi, S. and Fujiwara, H. (2000) *Circulation*, **102**, 2063.

32 Lanza, G.M., Wallace, K.D., Scott, M.J., Cacheris, W.P., Abendschein, D.R., Christy, D.H., Sharkey, A.M., Miller, J.G., Gaffney, P.J. and Wickline, S.A. (1996) *Circulation*, **94**, 3334.

33 Morawski, A.M., Winter, P.M., Crowder, K.C., Caruthers, S.D., Fuhrhop, R.W., Scott, M.J., Robertson, J.D., Abendschein, D.R., Lanza, G.M. and Wickline, S.A. (2004) *Magnetic Resonance in Medicine*, **52**, 1255.

34 Lanza, G. and Wickline, S. (2001) *Progress in Cardiovascular Diseases*, **44**, 13.

35 Sajid, M. and Stouffer, G. (2002) *Thrombosis and Haemostasis*, **87**, 187.

36 Brooks, P., Clark, R. and Cheresh, D. (1994) *Science*, **264**, 569.

37 Anderson, S., Randall, K., Westlin, W., Null, C., Jackson, D., Lanza, G., Wickline, S. and Kotyk, J. (2000) *Magnetic Resonance in Medicine*, **44**, 433.

38 Winter, P., Caruthers, S., Kassner, A., Harris, T., Chinen, L., Allen, J., Zhang, H., Robertson, J., Wickline, S. and Lanza, G. (2003) *Cancer Research*, **63**, 5838.

39 Schmieder, A.H., Winter, P.M., Caruthers, S.D., Harris, T.D., Williams, T.A., Allen, J.S., Lacy, E.K., Zhang, H., Scott, M.J., Hu, G., Robertson, J.D., Wickline, S.A. and Lanza, G.M. (2005) *Magnetic Resonance in Medicine*, **53**, 621.

40 Gramiak, R. and Shah, P. (1968) *Investigative Radiology*, **3**, 356.

41 Dalla Palma, L. and Bertolotto, M. (1999) *European Radiology*, **9**, S338.

42 Correas, J., Bridal, L., Lesavre, A., Mejean, A., Claudon, M. and Helenon, O. (2001) *European Radiology*, **11**, 1316.

43 McCulloch, M., Gresser, C., Moos, S., Odabashian, J., Jasper, S., Bednarz, J., Burgess, P., Carney, D., Moore, V., Sisk, E., Waggoner, A., Witt, S. and Adams, D. (2000) *Journal of the American Society of Echocardiography*, **13**, 959.

44 Szabo, T.L. (2004) *Diagnostic Ultrasound Imaging: Inside Out*, Elsevier Academic Press, Burlington, MA.

45 Goldberg, B.B. Raichlen, J.S. and Forsberg, F. (eds) (2001) *Ultrasound Contrast Agents: Basic Principles and Clinical Applications*, Martin Dunitz, London.

46 Leighton, T.G. (1997) *The Acoustic Bubble*, Academic Press, San Diego.

47 Klibanov, A.L., Rasche, P.T., Hughes, M.S., Wojdyla, J.K., Galen, K.P., Wible, J.H. and Brandenburger, G.H. (2004) *Investigative Radiology*, **39**, 187.

48 Blomley, M., Cooke, J., Unger, E., Monaghan, M. and Cosgrove, D. (2001) *British Medical Journal*, **322**, 1222.

49 Shohet, R., Chen, S., Zhou, Y., Wang, Z., Meidell, R., Unger, R. and Grayburn, P. (2000) *Circulation*, **101**, 2554.

50 Price, R., Skyba, D.M.P., Kaul, S.M. and Skalak, T.C.P. (1998) *Circulation*, **98**, 1264.

51 Cheng, T.D., S.C. and Feinstein, S. (1998) Contrast echocardiography: review and future directions. *American Journal of Cardiology*, **81**, 41.

52 Terasawa, A., Miyatake, K., Nakatani, S., Yamagishi, M., Matsuda, H. and Beppu, S. (1993) *Journal of the American College of Cardiology*, **21**, 737.

53 Harvey, C., Blomley, M., Eckersley, R., Cosgrove, D., Patel, N., Heckemann, R. and Butler-Barnes, J. (2000) *Radiology*, **216**, 903.

54 Leong-Poi, H., Christiansen, J., Klibanov, A., Kaul, S. and Lindner, J. (2003) *Circulation*, **107**, 455.

55 Schumann, P., Christiansen, J., Quigley, R., McCreery, T., Sweitzer, R., Unger, E., Lindner, J. and Matsunaga, T. (2002) *Investigative Radiology*, **37**, 587.

56 Hauff, P., Reinhardt, M., Briel, A., Debus, N. and Schirner, M. (2004) *Radiology*, **231**, 667.

57 Hughes, M.S., Marsh, J.N., Arbeit, J., Neumann, R., Fuhrhop, R.W., Lanza, G.M. and Wickline, S.A. (2005) Proceedings IEEE Ultrasonics Symposium, 05CH37716, p. 617.

58 Lanza, G.M., Trousil, R.L., Wallace, K.D., Rose, J.H., Hall, C.S., Scott, M.J., Miller, J.G., Eisenberg, P.R., Gaffney, P.J. and Wickline, S.A. (1998) *Journal of the Acoustical Society of America*, **104**, 3665.

59 Hall, C.S., Marsh, J.N., Scott, M.J., Gaffney, P.J., Wickline, S.A. and Lanza, G.M. (2000) *Journal of the Acoustical Society of America*, **108**, 3049.

60 Hughes, M.S. (1992) *Journal of the Acoustical Society of America*, **91**, 2272.

61 Hughes, M.S. (1992) Proceedings IEEE Ultrasonics Symposium, 92CH31187, p. 1205.

62 Hughes, M.S. (1993) *Journal of the Acoustical Society of America*, **93**, 892.

63 Hughes, M.S. (1993) Proceedings IEEE Ultrasonics Symposium, 93CH33019, p. 697.

64 Hughes, M.S. (1994) *Journal of the Acoustical Society of America*, **95**, 2582.

65 Hughes, M.S., Marsh, J.N., Hall, C.S., Savy, D., Scott, M.J., Allen, J.S., Lacy, E.K., Carradine, C., Lanza, G.M. and Wickline, S.A. (2005) *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency, Control*, **52**, 1555.

66 Bucy, R.S. and Joseph, P.D. (1987) *Filtering for Stochastic Processes with Applications to Guidance*, Chelsea Publishing Company, New York.

67 Weiner, N. (1949) *Extrapolation, Interpolation, and Smoothing of Stationary Time Series: with Engineering Applications*, M.I.T. Press, Cambridge, MA.

68 Grenander, U. and Rosenblatt, M. (1984) *Statistical Analysis of Stationary Time Series*, Chelsea Publishing Company, New York.

69 Wheeden, R.L. and Zygmund, A. (1977) *Measure and Integral: An Introduction to Real Analysis*, Marcel-Dekker, New York.

70 Cover, T.M. and Thomas, J.A. (1991) *Elements of Information Theory*, Wiley-Interscience, New York.

71 Kullback, S. (1997) *Information Theory and Statistics*, Dover, New York.

72 Reza, F.M. (1994) *An Introduction to Information Theory*, Dover, New York.

73 Arbeit, J.M., Riley, R.R., Huey, B., Porter, C., Kelloff, G., Lubet, R., Ward, J.M. and Pinkel, D. (1999) *Cancer Research*, **59**, 3610.

74 Arbeit, J.M., Manger, K., Howley, P.M. and Hanahan, D. (1994) *Journal of Virology*, **68**, 4358.

75 Hughes, M.S., Marsh, J.N., Woodson, A.K., Lacey, E.K., Carradine, C., Lanza, G.M. and Wickline, S.A. (2005) Proceedings IEEE Ultrasonics Symposium, 05CH37716 p. 373.

76 Hughes, M., Marsh, J., Hall, C., Fuhrhop, R.W., Lacy, E.K., Lanza, G.M. and Wickline, S.A. (2005) *Journal of the Acoustical Society of America*, **117**, 964.

77 Winter, P.M., Morawski, A.M., Caruthers, S., Harris, T., Allen, J.S., Zhang, H.Y., Fuhrhop, R.W., Lacy, E.K., Williams, T.A., Lanza, G.M. and Wickline, S.A. (2003) *Circulation*, **108**, 168.

78 Winter, P.M., Morawski, A.M., Caruthers, S.D., Fuhrhop, R.W., Zhang, H., Williams, T.A., Allen, J.S., Lacy, E.K., Robertson, J.D., Lanza, G.M. and Wickline, S.A. (2003) *Circulation*, **108**, 2270.

79 Hughes, M.S., Marsh, J.N., Allen, J., Brown, P.A., Lacy, E.K., Scott, M.J., Lanza, G.M., Wickline, S.A. and Hall, C.S. (2004) Proceedings of the IEEE Ultrasonics Symposium, 04CH37553, p. 1106.

80 Torr, G.R. (1984) The acoustic radiation force. *American Journal of Physics*, **52**, 402.

81 Nyborg, W.L. (1965) Acoustic streaming, in *Physical Acoustics* Vol. **IIB** (ed. W. Mason), Academic Press, p. 265.

82 Ter Haar, G. and Wyard, S. (1978) *Ultrasound in Medicine and Biology*, **4**, 111.

83 Lanza, G., Yu, X., Winter, P., Abendschein, D., Karukstis, K., Scott, M., Chinen, L., Fuhrhop, R., Scherrer, D. and Wickline, S. (2002) *Circulation*, **106**, 2842.

3
Nanoparticles for Cancer Detection and Therapy

Biana Godin, Rita E. Serda, Jason Sakamoto, Paolo Decuzzi, and Mauro Ferrari

3.1
Introduction

3.1.1
Cancer Physiology and Associated Biological Barriers

Cancer is a major public health problem in developed countries, accounting for nearly one-fourth of deaths in the United States, exceeded only by heart diseases. According to a 2008 report by the American Cancer Society, estimated numbers for US cancer cases are 745 and 692 thousands for men and women, respectively [1, 2], with the lifetime probability of developing cancer higher in men (45%) than in women (38%). *Cancer* is a general term used to define any disease characterized by the uncontrolled proliferation of abnormal cells. Due to a widely used generalization of the condition, it is easy to overlook the fact that cancer is not a single disease, but rather a conglomerate of many diseases. During the past five decades, the complexity of cancer has been rendered more tangible by a large body of knowledge accumulated on the common principles of pathogenesis. It is now clear that cancer is a complex ailment caused by accumulation of multiple molecular alterations in the genetic material.

The disease can be divided into two broad categories of hematological malignancies (which affect circulating cells) and solid tumors. Solid tumors can be considered as an organ, and are divisible into three main subcompartments: vascular; interstitial; and cellular [3, 4]. Each of these subcompartments accounts for several biological barriers (or 'biobarriers') that a therapeutic agent should bypass to treat the disease effectively [5, 6]. Later in this section, we will describe the tumor compartments as well as related and other intrinsic 'biobarriers', which severely impede the localization of chemicals, biomolecules and particulate systems at their intended site of action. Biobarriers are sequential in nature, and therefore the probability of an active agent of reaching its therapeutic goal is the interrelated result of the individual probabilities of overcoming each one of the challenges it faces [7, 8].

Nanotechnology, Volume 5: Nanomedicine. Edited by Viola Vogel

ISBN: 978-3-527-31736-3

The *tumor vasculature* is extremely heterogeneous, with necrotic and hemorrhagic areas neighboring regions with a dense vascular network formed as a result of angiogenesis triggered to sustain a sufficient supply of oxygen and nutrients necessary for tumor growth and progression [5, 9]. Tumor blood vessels are architecturally and structurally different from their normal counterparts. The vascular networks that are formed in response to tumor growth are not organized into definitive venules, arterioles and capillaries – as for the normal circulation – but rather share chaotic features of all of them. Furthermore, the blood flow in tumor vessels is irregular, sluggish, and sometimes oscillating. Angiogenic vessels possess several abnormal features such as a comparatively high percentage of proliferating endothelial cells, an insufficient number of pericytes, an enhanced tortuosity, and the formation of an atypical basal membrane. As a result, tumor vasculature is more permeable, with the pore cut-off size ranging from 380 to 780 nm in different tumor models [10, 11]. The hemoglobin in the erythrocytes is 'oxygen-starved', which makes the microenvironment profoundly hypoxic. The tumor environment is also nutrient-deficient (e.g. glucose), acidic (owing to lactate production from anaerobic glycolysis), and under oxidative stress [3, 9]. Although the molecular controls of the above abnormalities are not fully elucidated, these may be attributed to the imbalanced expression and function of angiogenic factors. Various mediators can affect angiogenesis as well as vascular permeability. Among these are vascular endothelial growth factor (VEGF), nitric oxide, prostaglandins and bradykinin. Macromolecules can traverse through neoplastic vessels using one of the following pathways: vasculature fenestrations; interendothelial junctions; transendothelial channels (open gaps); and vesicular vacuolar organelles [9]. The tumor vasculature, in being formed *de novo* during the angiogenic process, possesses a number of characteristic markers which are not seen on the surface of normal blood vessels, and can serve as therapeutic targets (these will be discussed later).

The *interstitial compartment* of solid tumors is mainly composed of a collagen and elastic fiber, crosslinked structure. Interstitial fluid and high-molecular-weight gelling constituents, such as hyaluronate and proteoglycans, are interdispersed within the above network. The characteristic feature of the interstitium, which distinguishes it from the majority of normal tissues, is the intrinsic high pressure resulting from the absence of an anatomically well-defined and operating lymphatic network, as well as an apparent convective interstitial fluid flow. These parameters present additional biobarriers towards the penetration of a therapeutic agent into the cancer cells, as the transport of an anticancer molecule or nanovector in this tumor subcompartment will be governed by physiological (pressure) and physico-chemical (charge, lipophilicity, composition, structure) properties of the interstitium and the agent itself [4, 5].

The *cellular subcompartment* accounts for the actual cancerous cell mass. The barriers directly related to the cellular compartment are generally categorized in terms of alterations in the biochemical mechanisms within the malignant cells making them resistant to anticancer medications. Among these biochemical shifts

are the P-glycoprotein efflux system, which is responsible for multidrug resistance and the impaired structure of specific enzymes (i.e. topoisomerase). Moreover, in order to efficiently treat the disease, a cytotoxic agent should be able to cross the cytoplasmic and nuclear membranes – a far from trivial deed for basic drugs that are ionizable within an acidic tumor environment [12, 13].

As mentioned above, following their administration, therapeutic agents encounter a multiplicity of biological barriers that adversely impact their ability to reach the intended target at the desired concentrations [5–8, 14]. This problem is considerably decoupled from the ability of agents to recognize and selectively bind to the target, that is, by the use of antibodies, aptamers or ligands. In other words, despite their high specificity these agents invariably present with concentrations at target sites that are vastly inferior to what is expected on the basis of molecular recognition alone. The biodistribution profiles for conventional chemotherapeutic agents are evenly adverse, if not worse, leading to a plethora of unwanted toxicities and collateral effects at the expense of the therapeutic action (i.e. a decreased 'therapeutic index'). The reticuloendothelial system (RES), which comprises immune cells and organs such as the liver and spleen, presents an important physiological biobarrier, causing an efficient clearance of the agent from the bloodstream. Other barriers of epithelial and endothelial nature, for example the blood–brain barrier, are based on tight-junctions, which significantly limit the paracellular transport of agents that owe their molecular discrimination to several mechanisms and proteins (occludin, claudin, desmosomes, zonula occludens).

To summarize, some of the most challenging biobarriers as the main cause for tumor resistance to therapeutic intervention, include physiological noncellular and cellular barriers, such as the RES, epithelial/endothelial membranes and drug extrusion mechanisms, and biophysical barriers, which include interstitial pressure gradients, transport across the extracellular matrix (ECM), and the expression and density of specific tumor receptors.

3.1.2
Currently Used Anticancer Agents

Since the pathology of cancer involves the dysregulation of endogenous and frequently essential cellular processes, the treatment of malignancies is extremely challenging. The vast majority of presently used therapeutics utilize the fact that cancer cells replicate faster than most healthy cells. Thus, most of these agents do not differentiate greatly between normal and tumor cells, thereby causing systemic toxicity and adverse side effects. More selective agents – which include monoclonal antibodies and anti-angiogenic agents – are now available, and the efficiency of these medications is still under evaluation in various types of tumor. Since cancer arising from certain tissues – including the mammary and prostate glands – may be inhibited or stimulated by appropriate changes in hormone balance, several malignancies may also respond to hormonal therapy. Various groups of anticancer therapeutics are exemplified below.

3.1.2.1 Chemotherapy

Chemotherapy, or the use of chemical agents to destroy cancer cells, is a mainstay in the treatment of malignancies. The modern era of cancer chemotherapy was launched during the 1940s, with the discovery by Louis S. Goodman and Alfred Gilman of nitrogen mustard, a chemical warfare agent, as an effective treatment for blood malignancies [15, 16].

Through a variety of mechanisms, chemotherapy affects cell division, DNA synthesis, or induces apoptosis. Consequently, more aggressive tumors with high growth fractions are more sensitive to chemotherapy, as a larger proportion of the targeted cells are undergoing cell division at any one time. A chemotherapy agent may function in only one phase (G_1, S, G_2 and M) of the cell cycle (when it is called cell cycle-specific), or be active in all phases (cell cycle-nonspecific). The majority of chemotherapeutic drugs can be categorized as alkylating agents (e.g. cisplatin, carboplatin, mechlorethamine, cyclophosphamide, chlorambucil), antimetabolites (e.g. azathioprine, mercaptopurine), anthracyclines (daunorubicin, doxorubicin, epirubicin, idarubicin, valrubicin), plant alkaloids (vinca alkaloids and taxanes) and topoisomerase inhibitors (irinotecan, topotecan, amsacrine, etoposide) [17–19].

The lack of any great selectivity by chemotherapeutic agents between cancer and normal cells is apparent when considering the adverse effect profiles of most chemotherapy drugs [18, 19]. Hair follicles, skin and the cells that line the gastrointestinal tract are some of the fastest growing cells in the human body, and therefore are most sensitive to the effects of chemotherapy. It is for this reason that patients may experience hair loss, rashes and diarrhea, respectively. As these agents do not possess favorable pharmacokinetic profiles to localize specifically into the tumor tissue, they become evenly distributed throughout the body, with resultant adverse side effects and other toxic reactions that greatly limit their dosage.

3.1.2.2 Anti-Angiogenic Therapeutics

The publication of Judah Folkman's imaginative hypothesis in 1971 launched the current research area of anti-angiogenic therapy for cancer [20], although more than three decades elapsed before the Food and Drug Administration (FDA) approved the first anti-angiogenic drug, bevacizumab (a humanized monoclonal antibody directed against VEGF) [21, 22]. The first clinical trials with this agent, when used in combination with standard chemotherapy, resulted in an enhanced survival of metastatic colorectal cancer and advanced non-small-cell lung cancers [23, 24]. Another group of anti-angiogenic therapeutics, also approved by the FDA, is based on small-molecule receptor tyrosine kinase inhibitors (RTKIs) which target VEGF receptors, platelet-derived growth factor (PDGF) and other tyrosine kinase-dependent receptors [25]. Examples of agents in this group are sorafenib and sunitinib; these orally administered medications have been shown to be effective in the treatment of metastatic renal cell cancer and hepatocellular carcinoma, when used as monotherapy [26–28]. When used as monotherapy, the survival benefits of these treatments are relatively modest (usually measured in months). Additionally, the treatments are also costly [29] and have toxic side effects [30–34].

3.1.2.3 Immunotherapy

While tumor cells are ultimately derived from normal progenitor cells, transformation to a malignant phenotype is often accompanied by changes in antigenicity. Antibodies are amazingly selective, possessing the natural ability to produce a cytotoxic effect on target cells. The immune system was first appreciated over 50 years ago for its ability to recognize malignant cells and defend against cancer, when Pressman and Korngold [35] showed that antibodies could distinguish efficiently between normal and tumor tissues. These results were confirmed by Burnet [36] during the 1960s, who also showed that neoplasms are actually formed only when lymphocytes lose the capability of differentiating between normal and malignant cells. These studies grounded the foundation for modern monoclonal antibody (mAb) -based cancer therapy. The expression of tumor-associated antigens can arise due to a variety of mechanisms, including alterations in glycosylation patterns [37], the expression of virally encoded genes [38], chromosomal translocations [39], or an overexpression of cellular oncogenes [40, 41]. The first challenge in the development of efficient mAb-based therapeutics is the detection of an appropriate and specific tumor-associated antigen. Some examples of mAbs used for cancer therapy are given below.

Hematologic malignancies, which possess fewer barriers capable of preventing mAbs from accessing their target antigens, are well suited for mAb therapy. Following intravenous injection and distribution throughout the vascular space, therapeutic antibodies may easily access their targets on the surface of blood malignant cells. Many of these B- and T-cell surface antigens, such as CD20, CD22, CD25, CD33 or CD52, are expressed only on a particular family of hematopoietic cells [42, 43]. These antigens are also expressed at high levels on the surface of various populations of malignant cells, but not on normal tissues or hematopoietic progenitor cells. The chimeric antibody which binds to CD20 B lymphocyte surface antigens, rituxan (rituximab; Genentech) was among the first of the mAbs to receive FDA approval for the treatment of nonHodgkin's lymphoma [44]. Alemtuzumab, which recognizes CD52 antigens present on normal B and T lymphocytes (Campath-1; Ilex Oncology) has also received FDA approval for the treatment of patients suffering from chronic lymphocytic leukemia.

The successful treatment of solid tumors with mAb therapeutics has proved to be more elusive compared to hematological malignancies, although some significant therapeutic benefits have been achieved. The failure of mAbs in the treatment of these malignancies is primarily attributable to an insufficient level of injected mAb that actually reaches its target within a tumor mass. The results of several studies using radiolabeled mAbs have suggested that only a very small percentage of the original injected antibody dose (0.01–0.1% g^{-1} tumor tissue) is able to ever reach target antigens within a solid tumor [45–47]. These low *in vivo* concentrations are due to the series of biobarriers (see above) that an intravenously administered mAb encounters *en route* to its specific antigens on the surface of cancer cells. Herceptin (trastuzumab; Genentech) is a humanized antibody marketed for the treatment of metastatic breast cancer. This mAb recognizes an extracellular epitope of the HER-2 protein, which is highly overexpressed in approximately 25–30% of invasive

breast tumors [40, 41]. It is noteworthy that HER-2 expression on breast cancer cells can be as high as 100-fold in comparison to normal breast epithelial cells. Clinical trials with herceptin have shown it to be well tolerated, both as a single agent for second- or third-line therapy or in combination with chemotherapeutic agents as a first line of therapy. A combination therapy resulted in a 25% improvement in overall survival among patients with HER-2-overexpressing tumors that are refractory to other forms of treatment [48, 49].

The levels of prostate-specific membrane antigen (PSMA), a transmembrane protein expressed primarily on the plasma membrane of prostatic epithelial cells [50], are elevated in virtually all cases of prostatic adenocarcinoma, with maximum expression levels observed in metastatic disease and androgen-independent tumors [50–53]. Due to this behavior, PSMA has become an important biomarker for prostate cancer, and antibodies to PSMA are currently being developed for the diagnosis and imaging of recurrent and metastatic prostate cancer, as well as for the therapeutic management of malignant disease [53–56].

Another mAb used for the treatment of colorectal cancer is elecolomab (panorex; GlaxoSmith-Kline), the anti-epithelial cellular adhesion molecule. Today, many other immunotherapeutics are being used in the clinic or are undergoing various stages of clinical trials. Beyond their pronounced therapeutic potential, these agents can be efficiently combined with nanovectors to enhance targeting of the latter to cancer tissues.

3.1.2.4 **Issues and Challenges**

As mentioned above, currently used conventional cancer therapies have several drawbacks that result in a pronounced toxicity and poor treatment efficacy. On the other hand, current diagnostic techniques do not allow for the competent detection of various malignancies, and do not reflect the vast clinical heterogeneity of the condition. Targeted approaches will ultimately increase the treatment efficiency, while decreasing toxicity to normal cells and tissues; thus, specific drug delivery in cancer treatment is of prime importance. As opposed to cancers of the blood, solid malignancies possess several unique characteristics, such as extensive blood vessel growth (angiogenesis), damaged vascular architecture and enhanced permeability, and impaired lymphatic flow and drainage. All of the above can serve as effective therapeutic targeting mechanisms, as well as for the passive homing of agents into the tumor tissue by means of various delivery systems.

To summarize, current issues and unmet needs in translational oncology include:

- Improved strategies for early cancer diagnostics and imaging.
- Advanced technologies to overcome the toxicity and adverse effects of chemotherapeutic agents.
- An accumulation of new knowledge on cancer biology, allowing for the design of more efficient therapeutics for more aggressive and lethal cancer phenotypes.

Progress in the above listed fields will sculpt the major cornerstones for a yet-to-come 'personalized' tumor therapy and early and predictive diagnosis of the disease.

Later in this chapter we will describe the currently available and under-development carriers and vectors from a 'nano-toolbox', and critically discuss the benefits and weaknesses of these systems for the design of specific, personalized and targeted medications. The benefits of rational design of the nanovectors to efficiently negotiate biobarriers and various aspects of a preclinical characterization for the nanoscale systems will be argued.

3.2
Nanotechnology for Cancer Applications: Basic Definitions and Rationale for Use

Nanoscience involves investigations to learn new behaviors and properties of materials on a submicron scale. Various important functions of living organisms and biological processes take place at the nanoscale. As an example, a typical protein such as hemoglobin, which carries oxygen through the bloodstream, is 5 nm (i.e. five-billionths of a meter) in diameter, while gamma-globulin accounts for a diameter of about 10 nm. For a comparison, the diameter of an erythrocyte – the smallest cell in the human body – ranges from 5 to 7 μm.

Research on the nanoscale has been a missing dimensional link, among an atomic scale which provides the basics for chemistry and physics, and micro-scale technologies, such as electronics. This issue was addressed by a Nobel Laureate Richard Feynman in his legendary lecture, "There is a plenty room in the bottom," in 1959 [57]. Almost four decades later, Richard Smalley, who received his Nobel Prize in 1996 for the discovery of the foundational in nanoscience and nanotechnology carbon-60 molecules, said "...human health has always been determined on the nanometer scale; this is where the structure and properties of the machines of life work in every one of the cells in every living thing." Nowadays, nanotechnology is a rapidly growing multidisciplinary field involving support from scientists in academia, industry and regulatory as well as federal sectors. As an example, the National Nanotechnology Initiative (NNI) program was established in 2001 to coordinate Federal Nanotechnology Research and Development [58]. The 2009 budget request provides US$ 1.5 billion for the NNI, with major investment in nanotechnology research and development over the past decade, reflecting a broad support of the US Congress for this program. Nanotechnology can offer impressive resolutions, when applied to medical challenges such as cancer, diabetes, Parkinson's or Alzheimer's disease, cardiovascular problems and inflammatory or infectious diseases.

Nanotechnology is more than simply throwing together a batch of nanoscale materials – it requires the ability to manipulate and control them in a useful way. The definition of nanotechnology pertains to synthetic and engineerable objects which are nanoscale in dimensions, or have critical functioning components of such a size, and that therefore possess special emergent properties [59]. This is a general and operational definition involving the following interrelated constituents: nanoscale dimensions of the whole system or its vital components; man-made nature; and the unique characteristics of new material that arise due to its nanoscopic size, with each element in this three-part description being equally essential for an object to be

defined as 'nanotechnological'. Another vital component in this definition is that the unique features and emerging properties of the nanomaterial must be backed up by the correct mechanism of action (e.g. mathematical modeling). Other definitions of nanotechnology can be found in the literature and, according to some agencies, the word 'nanoscale' should be interpreted to encompass the range of 1 to 100 nm. For example, the National Cancer Institute defines nanotechnology as

> "*The field of research that deals with the engineering and creation of things from materials that are less than 100 nanometers (one-billionth of a meter) in size, especially single atoms or molecules.*"

Nanotechnology has already occupied its niche for quite a few years in medicine, being known as 'nanomedicine', particularly in oncology [60–62]. The most studied and commercially available drug-delivery nanoparticle is the liposome, with liposomal doxorubicin having been granted FDA approval since 1996 for use against Kaposi's sarcoma. Later, it was also approved for use in metastatic breast cancer and recurrent ovarian cancer. Cancer-related issues of nanomedicine are supported by major funding programs; for example in 2005, the National Cancer Institute launched a US$ 144 million Alliance for Nanotechnology in Cancer.

The use of nanoparticles as carriers for therapeutic and imaging contrast agents is based on the simultaneous, anticipated advantages of drug localization at cancer lesions, and the ability to circumvent the biological barriers encountered between the point of administration and the projected target. Although physical localization at the tumor site is frequently defined as 'targeting' among the drug-delivery community, this term has a different scientific connotation, referring to the preferential activity of the agent on tumor-associated biological pathways. Due to this discrepancy, we will here use the term 'targeting' only when referring to the specific recognition between particles and the lesion (e.g., due to the presence of mAb on the particle's surface), while referring to passive concentration governed by physical laws as 'localization' or 'direction'.

A '*nanovector*' is a nanoscale particle or system having nanoscale components for the delivery of therapeutic or contrast agent. Currently used and investigated nanovectors can be generally organized into three main categories or 'generations' as shown schematically in Figure 3.1 [6, 8, 63]:

- The first generation (Figure 3.1a) comprises a delivery system that homes into the action site governed by passive mechanisms. In the case of liposomes as a nanovector, the mode of tumor localization is based on the enhanced permeation and retention (EPR) effect, which drives the system to localize into tumor through fenestrations in the adjacent neovasculature. Some of these carriers are surface-modified with a stealth layer [e.g. polyethylene glycol (PEG)] which prevents their uptake by the RES, and thus substantially prolongs the particles' circulation time [63].
- The second generation in this classification is thus defined as having specific additional functionalities on each individual particle, allowing for molecular recognition of target tissue (Figure 3.1b), or for the active or triggered release of the payload at the disease site. The best examples of the first subclass of nanovectors

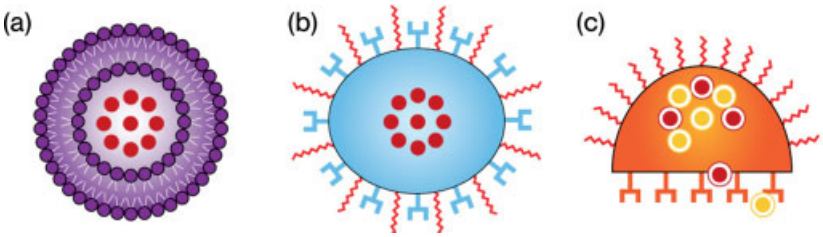


Figure 3.1 (a) First-generation nanovectors (e.g. clinical liposomes) comprise a container and an active principle, and localize in the tumor by enhanced permeation and retention (EPR), or the enhanced permeability of the tumor neovasculature; (b) Second-generation nanovectors further possess the ability to targeting their therapeutic action via antibodies and other biomolecules, remote activation, or responsiveness to environment; (c) Third-generation nanovectors (e.g. multistage agents) are capable of more complex functions, such as time-controlled deployment of multiple waves of active nanoparticles, deployed across different biological barriers and with different subcellular targets.

in this category are antibody-targeted nanoparticles, such as mAb-conjugated liposomes [64–69].

- Third-generation nanovectors, such as multistage agents, are capable of more complex functions which enable sequential overcoming of multiple biobarriers. An example is the time-controlled release of multiple payloads of active nanoparticles, negotiating different biological barriers and with different subcellular targets [7].

Later in this chapter we will focus on each of the three generations of nanovectors, discussing the 'pros' and 'cons', and presenting various examples of these technologies.

3.3
First-Generation Nanovectors and their History of Clinical Use

Today, the first-generation nanovectors that passively localize into tumor sites represents the only generation of nanomedicines broadly represented in the clinical situation. These systems are generally designed to achieve long circulation times for therapeutics and an enhanced accumulation of the drug into the target tissue. This is achieved through a pronounced extravasation of the carrier-associated therapeutic agent into the interstitial fluid at the tumor site, exploiting the locally increased vascular permeability, the EPR effect (Figure 3.2). An additional physiological factor which contributes to the EPR effect is that of impaired lymphatic function impeding clearance of the nanocarriers from their site of action [69–71]. The localization in this case is driven only by the particles' nanodimensions, and is not related to any specific recognition of the tumor or neovascular targets.

In order to prolong their circulation time, these systems are generally decorated on their surface by a 'stealth' layer (e.g. PEG) which prevents their uptake by phagocytic blood cells and organs of the RES system [63, 72, 73]. The most pronounced

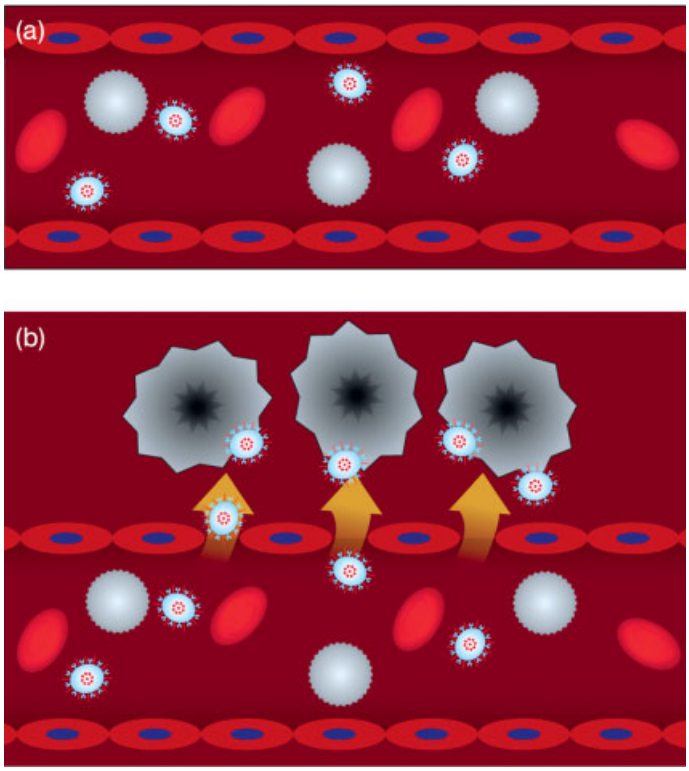


Figure 3.2 Mechanism of passive tumor targeting by enhanced permeation and retention (EPR).

representatives of this generation in clinical use are *liposomes*, which are the leaders among nanocarriers used in clinics. These self-assembling structures, which were first discovered by Bangham in 1965 [74], are composed of one or several lipid bilayers surrounding an aqueous core. This structure imparts an ability to encapsulate molecules that possess different degrees of lipophilicity; lipophilic and amphiphilic drugs will be localized in the bilayers while water-soluble molecules will concentrate into the hydrophilic core. The first drug to benefit from being encapsulated within this delivery system was doxorubicin. As of today, various companies market doxorubicin liposomal formulations, but Myocet (non-PEGylated liposomes) and Doxil (PEGylated liposomes) were among the first systems in clinical use [71, 75]. The pronounced advantages of liposomally encapsulated doxorubicin can be illustrated in its pharmacokinetic performance: an elimination half-life for the free drug is only 0.2 h, but this increases to 2.5 and 55 h, respectively, when non-PEGylated and PEGylated liposomal formulations are administered. Moreover, the area under the time–plasma concentration profile (the AUC), which indicates the bioavailability of an agent following its administration, is increased 11- and 200-fold for Myocet and Doxil, respectively, compared to the free drug [76]. Encapsulation into the liposomal carrier also causes a significant reduction in the most significant adverse side effect of doxorubicin, namely cardiotoxicity, as demonstrated in clinical trials [71, 75–77].

Liposomal doxorubicin is currently approved for the treatment of various malignancies, including Kaposi's sarcoma, metastatic breast cancer, advanced ovarian cancer and multiple myeloma. Other liposomal drugs which are either currently in use or are being evaluated in clinical trials include non-PEGylated liposomal daunorubicin (DaunoXome) and vincristine (Onco-TCS), PEGylated liposomal cisplatin (SPI-77) and lurtotecan (OSI-211) [78, 79].

Other systems in this category include metal nanoparticles for use in diagnostics, albumin paclitaxel nanoparticles approved for use in metastatic breast cancer, and drug–polymer constructs.

Nanoscale particles can act as contrast agents for all radiological imaging approaches. Iron-oxide particles provide a T_2-mode negative contrast for magnetic resonance imaging (MRI), while gold nanoparticles can be used to enhance the contrast in X-ray and computed tomography (CT) imaging, in a manner which is essentially proportional to their atomic number. Mechanical impendence disparity is the origin for the contrast in ultrasound imaging provided by the materials that are either more rigid (metals, ceramics) or much softer (microbubbles) than the surrounding tissue. The very existence of better contrast agents can drive the development of new imaging modalities. The emergence of nanocrystalline quantum dots has generated great interest in novel optical imaging technologies. The architecture and composition of quantum dots provide tunable emission properties that resist photobleaching. By concentrating preferentially at tumor sites following an EPR mechanism, nanoparticles which comprise a contrast material can provide an enhanced definition of anatomical contours and location, as well as the extent of disease. In addition, if coupled with a biological recognition moiety they can further offer molecular distribution information for the diagnostician [78].

Albumin-bound paclitaxel (Abraxane) was granted FDA approval in 2005. Paclitaxel is a highly lipophilic molecule that was previously formulated for injection with Cremophor, a toxic surfactant, under the trade name Taxol®. In a multicenter Phase II clinical trial involving 4400 women with metastatic breast cancer, Abraxane (30-min infusion, 260 mg m^{-2}) was proven to be more beneficial in terms of treatment efficiency and reduction in side effects than the free drug (3-h infusion, 175 mg m^{-2}) [80]. Albumin-bound methotrexate is currently being evaluated in the clinical situation.

Although the next group to be discussed does not have a particulate nature, these agents – drug–polymeric cleavable constructs – have also been considered as nanoengineered objects. In 1975, Ringdorf proposed a new concept of drug–polymer constructs that could be conjugated by using a linker with a certain degree of selectivity, and which would be stable in blood but cleaved in an acidic or enzymatic environment of a tumor site, or within an acidic intracellular compartment (e.g. endosomes) [81]. Some 20 years later, in 1994, doxorubicin conjugated to poly(*N*-(2-(hydroxypropyl)methacrylamide) (PHPMA), through an enzymatically cleavable tetrapeptide spacer (GFLG), was the first polymeric construct to enter clinical trials [70, 78]. This system significantly improved the therapeutic index of the drug, as indicated by a 45-fold higher maximum tolerated dose of the drug–polymer conjugate when compared to doxorubicin alone [82].

Other examples in this subcategory include PEG-L-asparaginase for lymphoblastic leukemia, PSMA-bound neocarcinostatin (which has been approved in Japan for the treatment of liver cancer), and PLGA-conjugated paclitaxel, which is currently undergoing Phase III evaluation for ovarian and non-small-cell lung cancer. In addition to such conventional polymer–drug, polymer–protein and protein–drug conjugates, several novel types of polymeric nanomedicines have also recently entered clinical trials, including cationic polyplexes for DNA and small interfering RNA (siRNA) delivery, dendrimers and polymeric micelles [71, 78, 79, 83].

3.4 Second-Generation Nanovectors: Achieving Multiple Functionality at the Single Particle Level

As defined above, the second generation of nanoparticles has specific additional functionalities on each individual particle, thus allowing for the molecular recognition of target tissues or for the active or triggered release of a payload at the disease site. The best examples of the targeting moieties utilized for homing the first subclass of nanovectors in this category (e.g. liposomes and other nanoparticles) are antibodies [64, 83–88]. Another example is the targeting through folate-receptors overexpressed on the membrane of some cancer cells.

Currently, a variety of targeting moieties besides antibodies are under investigation worldwide. These include ligands, aptamers and small peptides binding to specific target cell-surface markers or surface markers expressed in the disease microenvironment [89, 90].

By using active targeting, ligands can be attached to drugs to act as homing devices for binding to receptor structures expressed at the target site. Antibody–drug conjugates targeted to, for example, CD20, CD25 and CD33, which are (over) expressed in non-Hodgkin's lymphoma, T-cell lymphoma and acute myeloid leukemia, respectively, have been successfully used to deliver radionuclides (Zevalin), immunotoxins (Ontak) and antitumor antibiotics (Mylotarg) more selectively to tumor cells. Three platforms – immunoconstructs, immunoliposomes and immunopolymers – that utilize immune functional groups for targeting are presented schematically in Figure 3.3 [83].

A still unresolved question when targeting the solid tumor is the choice between high or low binding affinity of the ligand for its antigen or receptor. When the binding affinity is high, there is some evidence that the penetration of targeted therapeutics into a solid tumor is decreased due to the 'binding-site barrier'. In this case the targeted therapeutics binds strongly to the first targets encountered, but fails to diffuse further into the tumor. On the other hand, for targets in which most of the cells are readily accessible to the delivery system – for example, tumor vasculature and certain hematological malignancies – a high binding affinity might be desirable.

Another recently reported new system achieves targeting and detection based on PEGylated gold nanoparticles and surface-enhanced Raman scattering (SERS) (Figure 3.4a) [91]. These pegylated SERS nanoparticles have a significantly higher

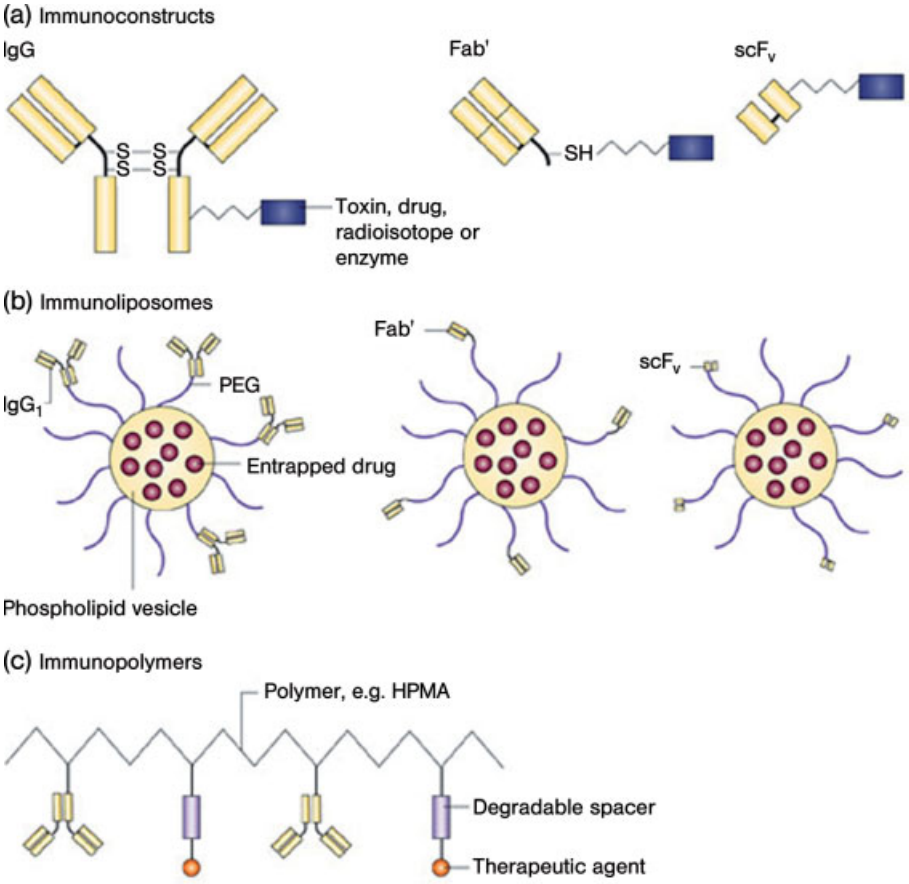


Figure 3.3 (a) Examples of targeted therapeutics constructs. Immunoconstructs are formed by the linking of antibodies, antibody fragments or nonantibody ligands to therapeutic molecules, such as toxins (immunotoxins), radioisotopes (radioimmunotherapy), drugs (immunoconjugates) or enzymes (ADEPT). Drug release, if required (immunotoxins and immunoconjugates), occurs through intracellular degradation of the peptide linker; (b) Immunoliposomes are formed by the attachment of multivalent arrays of antibodies, antibody fragments or nonantibody ligands to the liposome surface or, as in the example, to the terminus of hydrophilic polymers, such as polyethylene glycol (PEG), which are grafted at the liposome surface. The liposomes contain up to several million molecules of the therapeutic, the release of which occurs gradually by diffusion down its concentration gradient; (c) Immunopolymers are formed by linking both therapeutic agents and targeting ligands to separate sites on water-soluble, biodegradable polymers, such as hydroxypropylmethacrylamide (HPMA), with the use of appropriate degradable spacers to allow for drug release. ADEPT = antibody-directed enzyme–prodrug therapy; LTT = ligand-targeted therapeutic [83].

fluorescent intensity than quantum dots, with light emission in the near-infrared window, which is very appropriate for *in vivo* imaging. When conjugated to tumor-targeting ligands such as single-chain variable fragment (ScFv) antibodies, the conjugated nanoparticles were able to target tumor biomarkers such as epidermal

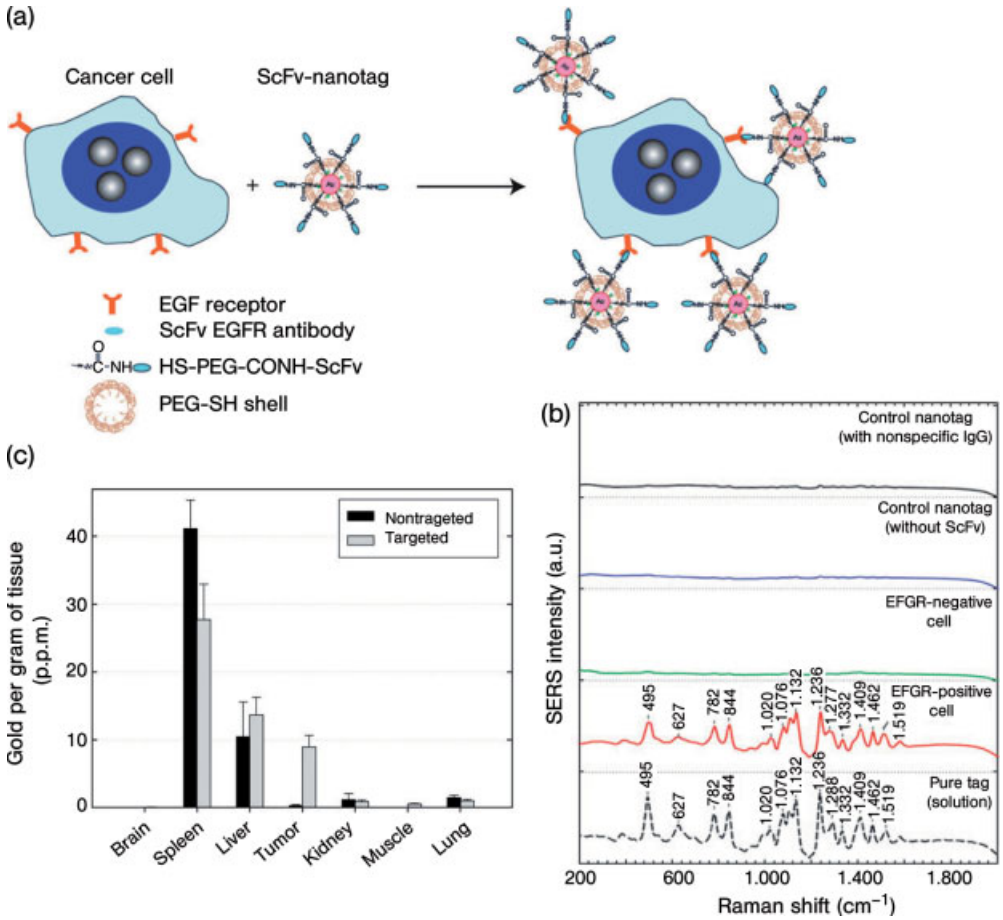


Figure 3.4 (a) Preparation of targeted SERS nanoparticles by using a mixture of SH-PEG and a heterofunctional PEG (SH-PEG-COOH). Covalent conjugation of an endothelial growth factor receptor (EGFR)-antibody fragment occurs at the exposed terminal of the heterofunctional PEG; (b) SERS spectra obtained from EGFR-positive cancer cells (Tu686) and EGFR-negative cancer cells (human non-small-cell lung carcinoma NCI-H520), together with control data and the standard tag spectrum. All spectra were taken in cell suspension with 785-nm laser excitation, and were corrected by subtracting the spectra of nanotag-stained cells by the spectra of unprocessed cells. The Raman reporter molecule is diethylthiatricarbocyanine (DTTC), and its distinct spectral signatures are indicated by wave numbers (cm^{-1}); (c) Biodistribution data of targeted and nontargeted gold nanoparticles in major organs at 5 h after injection, measured using inductively coupled plasma-mass spectrometry (ICP-MS). Reproduced with permission from Ref. [91].

growth factor (EGF) receptor on human cancer cells and in xenograft tumor models, with a 10-fold higher accumulation for targeted particles (see Figure 3.4b).

The nanovectors in the second subclass of this generation include responsive systems, such as pH-sensitive polymers or those activated by the disease site-specific enzymes, as well as a diverse group of remotely activated vectors. Among the most

interesting examples here are gold nanoshells that are activated by near-infrared light, or iron oxide nanoparticles triggered by oscillating magnetic fields [92, 93]. Other techniques used to remotely activate the second-generation particulates include ultrasound and radiofrequency (RF) [94–96]. Linking nanoshells to antibodies that recognize cancer cells enables these novel systems to seek out their cancerous targets prior to applying near-infrared light to heat them up. For example, in a mouse model of prostate cancer, nanoparticles activated with 2′-fluoropyrimidine RNA aptamers that recognized the extracellular domain of the PSMA, and loaded with docetaxel as a cytostatic drug, were used for targeting and destroying cancer cells [97, 98]. Another new approach is based on the coupling of nanoparticles to siRNA, used to silence specific genes responsible for malignancies. By using targeted nanoparticles, it was shown that the delivered siRNA can slow the growth of tumors in mice, without eliciting the side effects often associated with cancer therapies. Although the representatives of the second generation have not yet been approved by FDA, there are today numerous ongoing clinical trials involving targeted nanovectors, especially in cancer applications.

3.5
Third-Generation Nanoparticles: Achieving Collaborative Interactions Among Different Nanoparticle Families

The fundamental basis for the administration of drugs is to achieve a favorable therapeutic outcome in the treatment of a medical condition or disease, with minimal detrimental side effects. So far, we have described in detail the first- and second-generation nanoparticle therapeutic strategies. Although each generation has demonstrated incremental improvements relative to conventional intravenously administered chemotherapies, 'blockbuster' drug status has yet to be achieved by any nanobased construct. The second-generation nanoparticles offered new degrees of sophistication compared to their predecessors by employing additional complexities such as targeting moieties, remote activation, and environmentally sensitive components. However, these improvements predominantly represent simply a progressive evolution of the first-generation vectors; these subtle, yet augmenting, particle improvements do not fully address the primary challenge – or set of challenges – presented in the form of sequential biological barriers that continue to impair the efficacy of first- and second-generation nanoparticulates. This fundamental problem has given rise to a nanoparticle paradigm shift with the emergence of a third generation of particle that is specifically engineered to avoid biological barriers and to codeliver multiple nanovector payloads with tumor specificity.

The human body presents a robust bodily defense system that is extremely effective in preventing injected chemicals, biomolecules, nanoparticles and any other foreign agent(s) of therapeutic action from reaching their intended destinations. In addition to these natural biologic defenses, tumor-associated obstacles also exist. As a demonstration of the efficacy of these combined biological barriers, it has been calculated that only one out of every 100 000 molecules of drug successfully

reaches the intended site, permitting the overwhelming majority of the highly toxic, nondiscriminating, systemically disbursed poison to manifest in a number of undesirable side effects associated with cancer chemotherapy. This familiar scenario was quantitated in a study of Kaposi's sarcoma study that showed the percentage concentration of doxorubicin in Kaposi's sarcoma lesions to be ~0.001% [99]. This therapeutic phenomenon does not appear to be a tumor-specific challenge, however, and is therefore applicable to the lion's share of malignancies and tumor types [5, 100–102].

Some of the above-mentioned and most notable challenges include physiological barriers (i.e. the RES, epithelial/endothelial membranes, cellular drug extrusion mechanisms) and biophysical barriers (i.e. interstitial pressure gradients, transport across the ECM, expression and density of specific tumor receptors, and ionic molecular pumps). Biobarriers are sequential in nature, and therefore the probability of reaching the therapeutic objective is the product of individual probabilities of overcoming each and every one of them [8]. The requirement for a therapeutic agent to be provided with a sufficient collection of weaponry to conquer all barriers, yet still be small enough for safe vascular injection, is the major challenge faced by nanotechnology [14]. Once injected, nanoscale drug delivery systems (or 'nanovectors') are ideal candidates for the time-honored problem of optimizing the therapeutic index for treatment – that is, to maximize efficacy, while reducing adverse side effects.

The ideal injected chemotherapeutic strategy is envisioned to be capable of navigating through the vasculature after intravenous administration, to reach the desired tumor site at full concentration, and to selectively kill cancer cells with a cocktail of agents with minimal harmful side effects. Third-generation nanoparticle strategies represent the first wave of next-generation nanotherapeutics that are specifically equipped to address biological barriers to improve payload delivery at the tumor site. By definition, third-generation nanoparticles have the ability to perform a time sequence of functions which involve the cooperative coordination of multiple nanoparticles and/or nanocomponents. This novel generation of nanotherapeutics is exemplified through the employment of multiple nanobased products that synergistically provide distinct functionalities. In this chapter, the nanocomponents will include any engineered or artificially synthesized nanoproducts, including peptides, oligonucleotides (e.g. thioaptamers, siRNA) and phage with targeting peptides. Naturally existing biological molecules, such as antibodies, will be excluded from this designation, despite their ability to be synthesized.

Third-generation approaches have been developed to address the numerous challenges responsible for reducing the chemotherapeutic efficacy of earlier strategies. For example, surface modification of the exterior of nanoparticles with PEG has proven to be effective in increasing the circulation time within the bloodstream; however, this preservation tactic proves detrimental to the biological recognition and targeting ability of the nanovector [103]. In order to avoid such paradoxical approaches of employing debilitating improvements to therapeutic delivery systems, many research groups are combining multiple nanotechnologies to exploit the additive contributions of the constituent components. One example of third-generation nanoparticles is the biologically active molecular networks known as

'nanoshuttles'. These self-assemblies of gold nanoparticles within a bacteriophage matrix combine the hyperthermic response to near-infrared radiation of the gold with the biological targeting capabilities of the 4C-RGD sequence presented by the phage [104]. The nanoshuttles also collectively accommodate enhanced fluorescence, dark-field microscopy and surface-enhanced Raman scattering detection.

The next example of third-generation nanoparticles is the disease-inspired approach of the 'nanocell'. Newly emerging chemotherapeutic models utilize combinational therapies that are intended to inhibit tumor neovasculature growth and kill cancer cells. The coadministration of anti-angiogenic agents with conventional cytotoxic agents is a novel concept, but this practice has faced two critical problems. First, it has been shown that anti-angiogenic agents are capable of depleting blood flow to the tumor by interrupting new vessel growth. Unfortunately, however, this shutdown of tumor blood vessels has resulted in the prevention of chemotherapeutic agents from reaching the tumor site at sufficient concentrations. Furthermore, the decreased blood flow elicits intra-tumoral hypoxia which in turn increases the expression of hypoxia-inducible factor-1α (HIF1-α). HIF1-α overexpression is correlated to increased tumor invasiveness and chemotherapy resistance [105]. By using the same combinatorial chemotherapy approach, researchers have developed a nested nanoparticle construct that comprises a lipid-based nanoparticle enveloping a polymeric nanoparticle core called a 'nanocell'. Here, a conventional chemotherapeutic drug (doxorubicin) is conjugated to a polymer core and an anti-angiogenic agent (combretastatin) is then trapped within the lipid envelope. When the nanocells are accumulated within the tumor through the EPR effect, the sequential time release of the anti-angiogenic agent, followed by the cytotoxic drug, causes an initial disruption of tumor vascular growth and effectively traps the drug-conjugated nanoparticle core within the tumor to allow an eventual delivery of the cancer cell-killing agent.

The final example of third-generation nanoparticle technology utilizes a multistage approach that addresses many biological barriers experienced by an injec therapy. Currently, research groups are developing nanoporous silicon microparticles that utilize their unique particle size, shape and other physical characteristics in concert with active tumor biological targeting moieties to efficiently deliver payloads of nanoparticles to the tumor site. The ability to deliver a therapeutic agent to a tumor is analogous to the lunar voyage embarked upon by the Apollo 11 crew. This epic feat was not achieved simply by piloting a single vehicle to the moon and back; instead, it required a sequential execution of numerous steps, which included three stages of the Saturn V rocket to escape the Earth's atmosphere, a lunar module to descend and ascend to and from the lunar landscape, and finally a command module for re-entry and splashdown back to Earth. Similar mission-critical issues must also be addressed in a sequential manner when developing drug delivery systems to fight cancer.

The multistage drug delivery system is predicated upon a Stage 1 nanoporous silicon microparticle that is specifically designed (through mathematical modeling) to exhibit superior margination and adhesion properties during its negotiation through the systemic blood flow en route to the tumor site. Particle characteristics such as size, shape, porosity and charge can be exquisitely controlled with precise reproducibility through semiconductor fabrication techniques. In addition to its

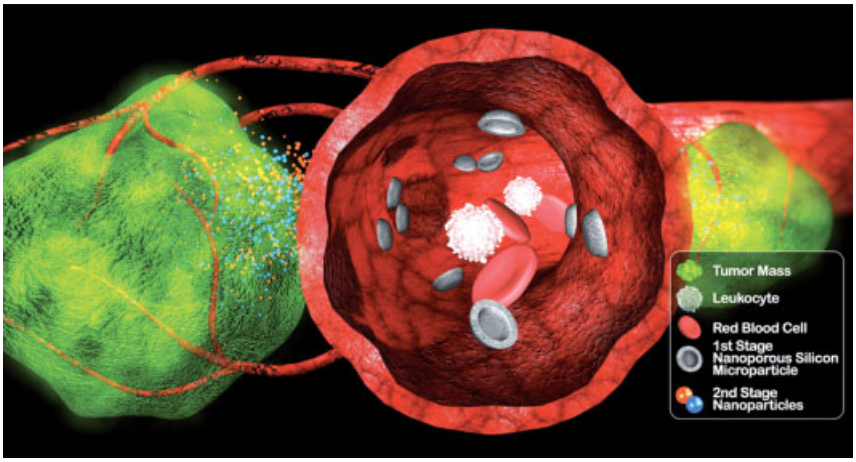


Figure 3.5 Multistage nanovectors. Stage 1 nanoporous silicon microparticles are engineered to exhibit an enhanced ability to marginate within blood vessels and adhere to tumor-associated endothelium. Once positioned at the tumor site, the Stage 1 particle can release its nanoparticle payload to achieve the desired therapeutic effect, prior to complete biodegradation of the carrier particle. The therapeutic outcome can be determined by selection of the nanoparticles loaded within the nanoporous structure of the Stage 1 particle.

favorable physical characteristics, the Stage 1 particle can be surface-treated with such modifications as PEG for RES avoidance and also equipped with biologically active targeting moieties (e.g. aptamers, peptides, phage, antibodies) to enhance the tumor-targeting specificity. This approach decouples the challenges of: (i) transporting therapeutic agents to the tumor-associated vasculature; and (ii) delivering therapeutic agents to cancer cells. The Stage 1 particles shoulder the burden of efficiently transporting a nanoparticle payload to the tumor site within the nanoporous structures of its interior (Figure 3.5). The nanoparticles called Stage 2 particles, generically represent any nanovector construct within the approximate diameter range of 5 to 100 nm. The Stage 1 particles have demonstrated the ability to rapidly load (within seconds) and gradually release (within hours) multiple nanoparticles (i.e. single-walled carbon nanotubes and quantum dots) during *in vitro* experiments, with complete biodegradation within 24–48 h, depending on the pore density [106]. Furthermore, unpublished preliminary data have demonstrated the ability to deliver liposomes and other nanovectors, as well as indications of the successful *in vivo* delivery of Stage 2 nanoparticles to tumor masses in xenograft murine models.

The multistage drug delivery system is emblematic of third-generation nanoparticle technology, since the strategy combines numerous nanocomponents to deliver multiple nanovectors to a tumor lesion. The Stage 1 particle is rationally designed to have a hemispherical shape to enhance particle margination within blood vessels, and to increase particle/endothelium interaction to maximize the probability of active tumor targeting and adhesion [107]. In addition to improved hemodynamic physical properties and active biological targeting by utilizing nanocomponents such as aptamers and phage, the Stage 1 particle can also present with specific surface

modifications in order to avoid RES uptake and exhibit degradation rates predetermined by nanopore density. Upon tumor recognition and vascular adhesion, a series of nanoparticle payloads may be released in a sequential order predicated upon Stage 1 particle degradation rates and payload conjugation strategies (e.g. environmentally sensitive crosslinking techniques, pH, temperature, enzymatic triggers). The versatility of this platform nanovector multistage delivery particle allows for a multiplicity of applications. Depending upon the nanoparticle 'cocktail' loaded within the Stage 1 particle, this third-generation nanoparticle system can provide for the delivery not only of cytotoxic drugs but also of remotely activated hyperthermic nanoparticles, contrast agents and future nanoparticle technologies.

3.6
Nanovector Mathematics and Engineering

Third-generation particles are transported by the blood flow and interact with the blood vessel walls, both specifically – through the formation of stable ligand–receptor bonds – and nonspecifically, by means of short-ranged van der Waals, electrostatic and steric interactions. If suitable conditions are met in terms of a sufficiently high expression of vascular receptors and sufficiently low hydrodynamic shear stresses at the wall, particles may adhere firmly to the blood vessel walls and control cell uptake, either by avoiding or favoring, based on their final objective. Such an intravascular 'journey' can be broken down into three fundamental events which form the cornerstone of the rational design, namely: the margination dynamics; the firm adhesion; and the control of internalization. The rational design of particles has the aim of identifying the dominating governing parameters in each of the above-cited events in order to propose the optimal design strategy as a function of the biological target (diseased cell or environment).

In physiology, the term 'margination' is conventionally used to describe the lateral drift of leukocytes and platelets from the core of the blood vessels towards the endothelial walls. This event is of fundamental importance as it allows an intimate contact between the circulating cells and the vessel walls, and in the case of leukocytes it is required for diapedesis. Similarly, the rational particle design should aim at generating a marginating particle, that can spontaneously move preferentially in close proximity to the blood vessel walls. Accumulating the particles in close proximity to the blood vessel walls is highly desirable both in vascular targeting and when the delivery strategy relies on the EPR approach. This occurs for two main reasons:

- The particles can 'sense' the vessel walls for biological and biophysical diversities, as for instance the overexpression of specific vascular markers (vascular targeting) or the presence of sufficiently large fenestrations through which they extravasate (EPR-based strategy).
- The particles can more easily leave the larger blood vessels in favor of the smaller ones, thus accumulating in larger numbers within the microcirculation (Figure 3.6) [108].

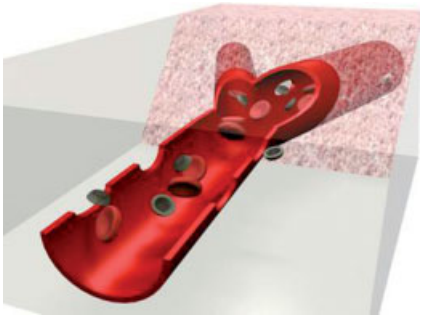


Figure 3.6 Marginating particles can more likely 'sense' the vessel walls for biological and biophysical diversity and more easily leave the vascular compartments through openings along the endothelium.

While leukocyte and platelet margination is an active process requiring an interaction with red blood cells (RBCs) and the dilatation of inflamed vessels with blood flow reduction [109], particle margination can only be achieved by proper rational design.

It should be noted that the RBCs – the most abundant blood-borne cell population – have a behavior opposite to margination, with an accumulation that occurs preferentially within the core of the vessels. This has long been described by Fahraeus and Lindqvist [110], and is referred to as the *plasma skimming* effect. An immediate consequence of this phenomenon is the formation of a 'cell-free layer' in the proximity of the wall, which varies in thickness with the size of the channel and mean blood velocity. For example, it may be as large as few tens of microns in arterioles ($\geq$100 μm in diameter) and a few microns in capillaries ($\geq$10 μm in diameter [111]). Particles designed to marginate should accumulate and move in a cell-free layer, which is also characterized by an almost linear laminar flow.

The motion of spherical particles in a linear laminar flow has been described by Goldmann *et al.* [112], who showed that the exerted hydrodynamic forces grow with the particle radius, and that no lateral drift would be observed unless an external forces such as gravitational or magnetic, or short-ranged van der Waals and electrostatic interactions were applied [113]. In other words, a neutrally buoyant spherical particle moving in close proximity to a wall can drift laterally only if an external force is applied. Here, it is important to recall that the gravitational force has been shown to be relevant even for submicrometer polystyrene beads (relative density to water of 0.05 g cm^{-3}), and that margination dynamics can be effectively controlled in horizontal channels by changing the size of the nonbuoyant nanoparticles [114]. On the other hand, nonspherical particles exhibit more complex motions with tumbling and rolling that can be exploited to control their margination dynamics, without any need for lateral external forces. The longitudinal (drag) and lateral (lift) forces, as well as the torque exerted by the flowing blood, depend on the size, shape and orientation of the particle to the stream direction, and change over time as the particle is transported. Considering an ellipsoidal particle with an aspect ratio of 2 (Figure 3.7a) in a linear laminar flow, the particle trajectory and its separation distance from the wall are shown in Figure 3.7b. Clearly, the particle motion is very complex,

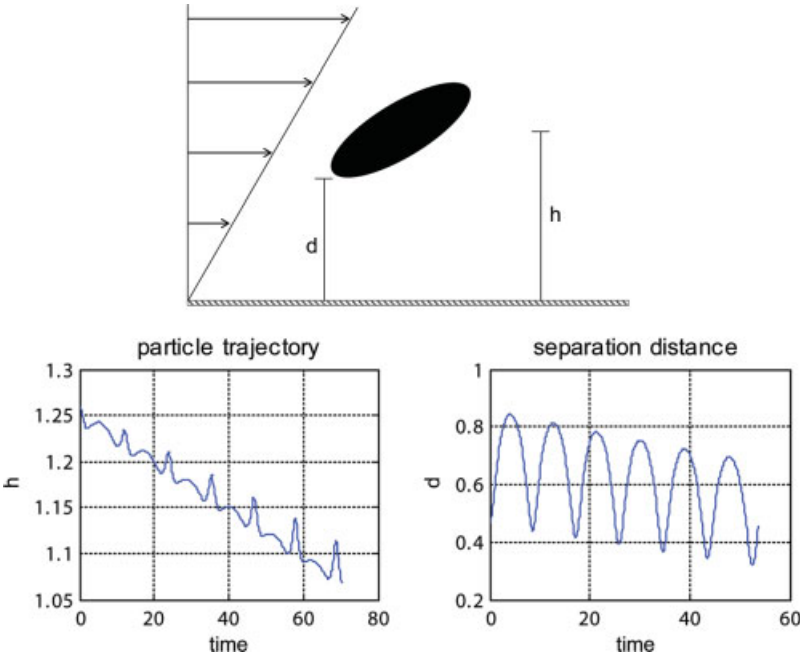


Figure 3.7 (a) An ellipsoidal particle transported within a linear laminar flow at a distance *d* from a rigid wall, as it would be in close proximity to the vessel walls; (b) The trajectory of the ellipsoidal particle with its characteristic oscillatory motion, and its separation distance from the wall reducing with time.

with periodic oscillations towards and away from the wall. Overall, however, the particle would approach the wall and interact with its surface.

For nonspherical particles, it has been shown that the lateral drifting velocity is directly related to their aspect ratio [115, 116], with a maximum between the two extremes: a sphere, with aspect ratio unity, and a disk, with aspect ratio infinity.

More recently, *in vitro* experiments have been conducted using spherical, discoidal and quasi-hemispherical particles with the same weight injected into a parallel plate flow chamber under controlled hydrodynamic conditions [117]. The experiments have shown that discoidal particles tend to marginate more than quasi-hemispherical and more than spherical particles in a gravitational field. Notably, these observations neglect the interaction of the particles with blood cells, in particular RBCs. However, this is a reasonable assumption as long as the particles are sufficiently smaller than RBCs and tend to accumulate within the cell-free layer.

Therefore, with regards to what concerns the design of marginating particles their size and shape, their geometric properties are of fundamental importance.

The marginating particle moving in close proximity to the blood vessels can interact both specifically and nonspecifically with the endothelial cells, and eventually adhere firmly to it. Firm and stable adhesion is ensured as long as the dislodging forces (hydrodynamic forces and any other force acting to release the particle from

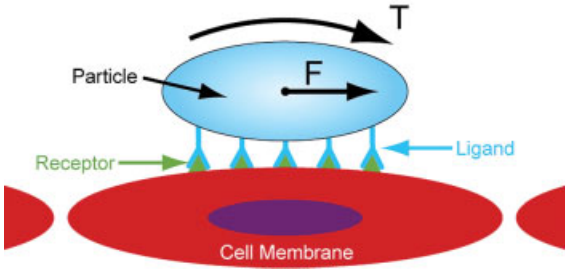


Figure 3.8 The longitudinal (drag) force (F) and the torque (T) exerted over a particle adhering to a cell layer under flow.

the target cell) are balanced by specific ligand–receptor interactions and nonspecific adhesion forces arising at the cell–particle interface (Figure 3.8).

The strength of adhesion must be expressed in terms of an adhesion probability factor, P_a, defined as the probability of having at least one ligand–receptor bond formed under the action of the dislodging forces. The probability of adhesion is decreased as the shear stress at the blood vessel wall μS and as the characteristic size of the particle increase; and grows as the surface density of ligand molecules m_l distributed over the particle surface and of receptor molecules m_r expressed at the cell membrane increases. However, for a fixed volume particle – that is to say, for a fixed payload – oblate particles with an aspect ratio γ larger than unity would have a larger strength of adhesion having fixed all other parameters [118]. Interestingly, for each particle shape, a characteristic size can be identified for which the probability of adhesion has a maximum, as shown in Figure 3.9. For small particles, the hydrodynamic forces are small but the area of interaction at the particle–cell interface is also smaller, leading consequently to a small number of ligand–receptor bonds involved which cannot withstand even a small dislodging force. For large particles, the number

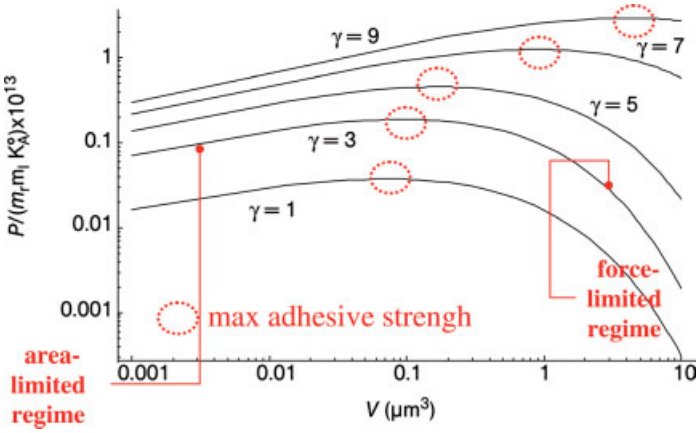


Figure 3.9 Variation of the normalized adhesive probability factor (P) with the volume V of the particle for different values of the aspect ratio γ [118].

of ligand–receptor bonds that can be formed grows, but the hydrodynamic forces grow even more. The optimal size for adhesion – that is, the size for which P_a has a maximum – falls between these two limiting conditions.

As an example, when considering a capillary with a shear stress at the wall of $\mu S = 1$ Pa and a surface density of receptors $m_r = 100\,\mu m^2$, the optimal radius for a spherical particle would be about 500 nm with a total volume of $0.05\,\mu m^3$, whereas the optimal volume for an oblate spheroidal particle with an aspect ratio $\gamma = 2$ would be more than 50 times larger ($3.5\,\mu m^2$) [118].

In particle adhesion, rational design should focus on the shape of the particle and the type and surface density of ligand molecules decorating the particle surface.

Once the particle has adhered to the target cell, it should be internalized if the aim is to release drugs or therapeutic agents within the cytosol or at the nuclear level (gene delivery). Alternatively, it should resist internalization if the target cell is used just as a docking site (vascular targeting) from which are released second-stage particles. The internalization rate is affected by the geometry of the particle and the ligand–receptor bonds involved.

Freund and colleagues [119] developed a mathematical model for receptor-mediated endocytosis based on an energetic analysis. This showed that a threshold particle radius R_{th} exists, below which endocytosis could never occur and, that an optimal particle radius R_{opt} exists, slightly larger than R_{th}, for which internalization is favored with the maximum internalization rate, thus confirming (in theory) the above-cited experimental observations. This analysis was then generalized to account for the contribution of the surface physico-chemical properties that may dramatically affect the internalization process, changing significantly both R_{opt} and R_{th} (Figure 3.10), as shown by Decuzzi and Ferrari [114].

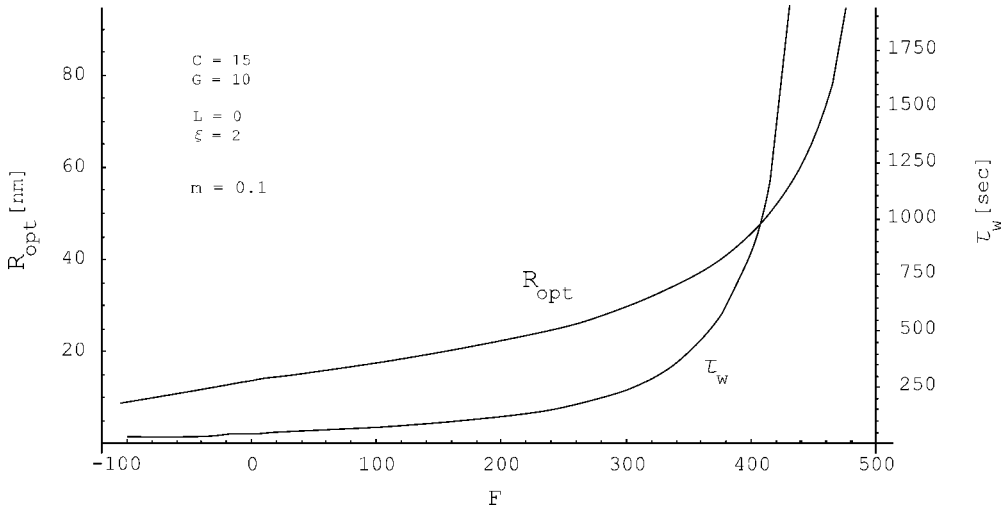


Figure 3.10 The optimal radius R_{opt} and the wrapping time τ_w as a function of the nonspecific parameter F, growing with the repulsive nonspecific interaction at the cell–particle interface.

A more recent theoretical model has been developed by Decuzzi and Ferrari for the receptor-mediated endocytosis of nonspherical particles [120]. This shows how elongated particles laying parallel to the cell membrane are less prone to internalization compared to spherical particles or particles laying normal to the cell membrane. The results show clearly how particle size and shape can be used to control the internalization process effectively, and that particles that deviate slightly from the spherical shape are more easily internalized compared to elongated particles that deviate severely from the classical spherical shape.

Even in the case of particle internalization, a judicious combination of surface physico-chemical properties and particle geometry would lead to a particle with optimized internalization rates, depending on the final biological applications.

Finally, a mathematical model has been recently developed [121] that allows one to predict the adhesive and endocytotic performances of particulate systems based on three different categories of governing parameters: (i) geometric (radius of the particle); (ii) biophysical (ligand-to-receptor surface density ratio; nonspecific interaction parameter; hydrodynamic force parameter); and (iii) biological (ligand–receptor binding affinity). This finding has led to the definition of *Design Maps* through which the three different states of the particulate system can be predicted: (i) no adhesion at the blood vessel walls; (ii) firm adhesion with no internalization by the endothelial cells; or (iii) firm adhesion and internalization (Figure 3.11) [121].

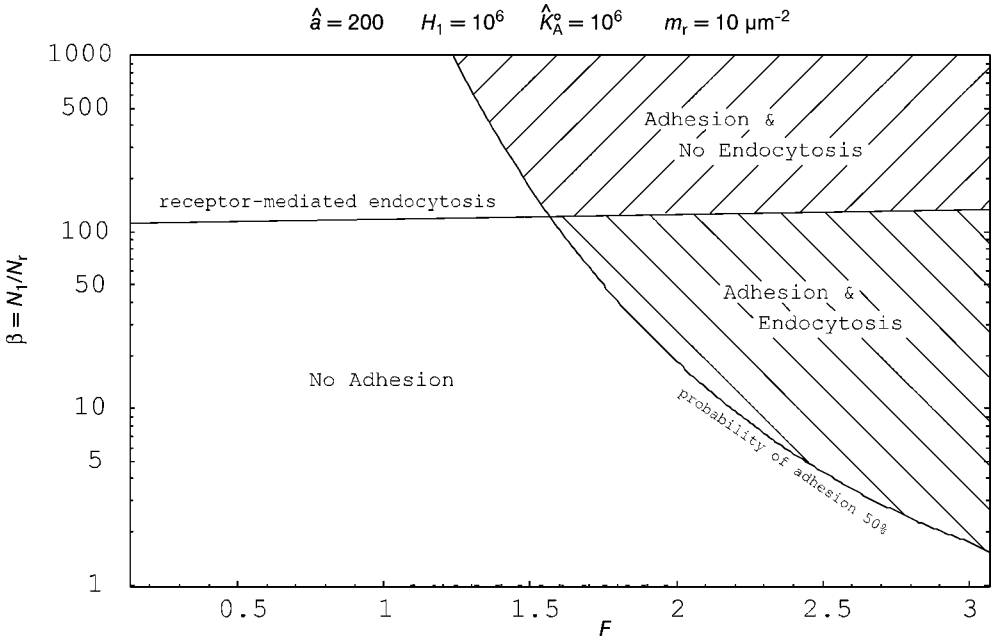


Figure 3.11 A typical design map showing areas for no adhesion; adhesion and no endocytosis; and adhesion and endocytosis [121].

3.7
The Biology, Chemistry and Physics of Nanovector Characterization

The small size, unique physico-chemical properties, and biological activity of nanoparticles create the need for extensive characterization prior to their use in biomedical applications. The National Cancer Institute has established the Nanotechnology Characterization Laboratory (NCL) to standardize and perform preclinical characterizations of nanomaterials designed for cancer therapeutics and diagnostics [122]. The objectives of the NCL are to speed the development of nanotechnology-based products for cancer patients, while reducing the risk to inventors, as well as encouraging private-sector investment. A further aim is to establish an analytical cascade of protocols for nanomaterial characterization. Challenges to creating standardized characterization techniques include the wide variety of materials used to construct nanomedicines. Thus, the characterization strategy is broad and includes physico-chemical characterization, sterility and pyrogenicity assessment, biodistribution (absorption, distribution, metabolism, excretion) and toxicity, both *in vitro* and *in vivo* in animal models [123]. An examination of the biological and functional characteristics of multicomponent/combinatorial platforms is also addressed.

NCL's standardized analytical cascade includes tests for preclinical toxicology, pharmacology and efficacy. The protocols include assays for physical attributes; *in vitro* testing for toxicity or biocompatibility; and *in vivo* testing for safety, efficacy and toxicokinetic properties in animal models (Figure 3.12).

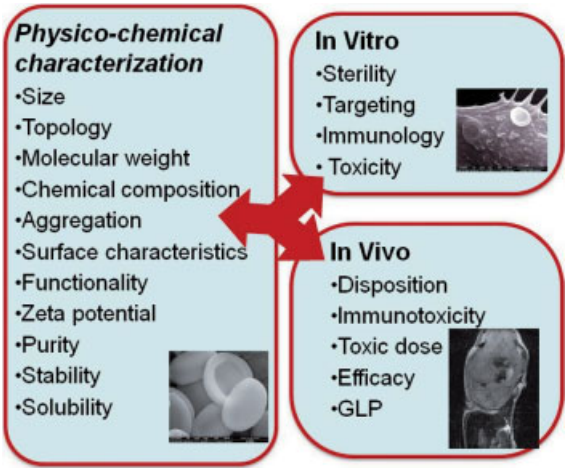


Figure 3.12 Preclinical characterization of nanoparticles involves physico-chemical, *in vitro* and *in vivo* characterization. The list includes assays outlined by the National Cancer Institute Nanotechnology Characterization Laboratory.

3.7.1 Physical Characterization

Physical characterization includes assays for particle size, size distribution, molecular weight, density, surface area, porosity, solubility, surface charge density, purity, sterility, surface chemistry and stability. The mean particle size – that is, the hydrodynamic diameter – is determined by batch-mode dynamic light scattering (DLS) in aqueous suspensions. Care must be taken with these measurements, because they can be affected by other parameters. For example, precautions include the cleaning of cuvettes with filtered, demineralized water; media filtering with 0.1 μm pore size membranes and pre-rinsing cuvettes multiple times; scattering contributions by media in the absence of analyte; optimized sample concentration; and filtering samples in conjunction with loading into the cuvette. The sample concentration should be optimized to avoid signal-to-noise ratio (SNR) deterioration at low concentration and particle interactions and scattering effects at high concentration. Another precaution is to add only small amounts of monovalent electrolyte in order to avoid salt effects on the electrical double-layer surrounding the particles in the media. Again, concentration optimization is necessary for optimal measurements. In order to evaluate instrument performance, latex size standards are commercially available. When analyzing these data, the absolute viscosity and refractive index for the suspending media is required to calculate the hydrodynamic diameter.

3.7.2 *In Vitro* Testing

In vitro testing includes the assessment of sterility, targeting, *in vitro* immunology and toxicity testing. Sterility testing for contaminates includes monitoring for the presence of endotoxins, bacteria, yeast, molds and mycoplasm. As an example, the LAL (limulus amoebocyte lysate) assay is commonly used to test for the presence of bacterial endotoxin. Although standard immunological *in vitro* assays exist, the preclinical immunotoxicity testing of nanoparticles has been hampered due to interference by the nanoparticles within the assay. Whilst a variety of mechanisms of interference exists, the most common occurrences are light absorbance by nanoparticles (which interfere with colorimetric methods) and the catalytic properties of nanoparticles creating false-positive effects in enzyme assays [122].

In vitro targeting assays measure cell binding and the internalization of particles (Figure 3.13). This is particularly relevant for drug delivery systems, as the route of internalization dictates the subcellular localization of nanoparticles. As an example, caveolar-mediated uptake leads to nanoparticles being localized into organelles with a nonacidic pH [124], whereas clathrin-mediated uptake favors their lysosomal entrapment [125], the latter leading to drug degradation. The uptake of larger particles (typically >500 nm) generally occurs by phagocytosis [126, 127]. Phagosomes typically fuse with endosomes, leading to lysosomal accumulation [126]. Common targeting assays include confocal microscopy, transmission and scanning electron microscopy, flow cytometric analysis and quantitative assessment assays (e.g. the BCA protein

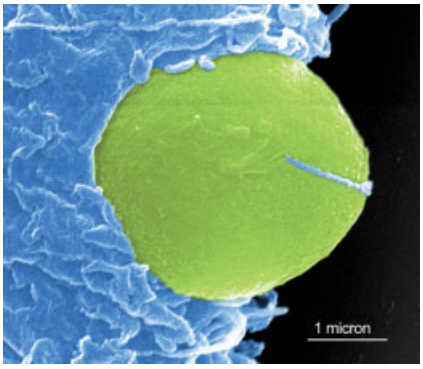


Figure 3.13 Pseudocolored scanning electron microscopy image of an endothelial cell internalizing a nanoporous silicon particle.

assay to assess PLGA uptake [128]). Both, fluorescence microscopy and flow cytometry rely on the attachment of fluorescent probes to the nanoparticle; alternatively, the latter technique may rely on changes in light scattering caused by the presence of internalized nanoparticles [127]. Controls are always essential to ensure that any intracellular fluorescence is not due to the uptake of dye that might have been released from the particles [128]. Factors affecting nanoparticle uptake include nanoparticle concentration, incubation time, nanoparticle size and shape and culture media.

For multicomponent systems, targeting may be difficult to access *in vitro*, especially for systems composed of nested particles, where each particle is targeting a specific and discrete population. Additional problems arise due to the modification of particles with imaging agents. For example, the conjugation of fluorescent probes to the surface of particles alters the surface charge density of the particle, and may also mask the binding of ligands on the particle surface. This in turn alters the ability of particles to bind to the cell-surface receptors that are responsible for their uptake. *In vitro* targeting assays also need to emphasize the impact of serum opsonization on particle uptake [127]. Serum components are signals for immune cells, and may either activate cells or serve as bridges attaching particles to cells. For example, antibodies found in serum may bind to particles and mediate their uptake via Fc receptors found on specific cell populations. The end result is dramatic, however, and may even completely alter which cell populations are able to internalize the particles.

To date, research investigations have shown that nanoparticles can stimulate and/or suppress the immune response [129]. Compatibility with the immune system is affected to a large degree by the surface chemistry. The cellular interaction of nanoparticles is dependent on either their direct binding to surface receptors or binding through the absorption of serum components to particles and their subsequent interaction with cell receptors [127]. Blood-contact properties of the nanomaterial and cell-based assays are used to determine the immunological compatibility of the device.

The blood-contact properties of nanoparticles are characterized by plasma protein binding, hemolysis, platelet aggregation, coagulation and complement activation.

- Plasma protein binding is achieved using two-dimensional gel electrophoresis, with individual proteins being evaluated by mass spectrometry. A drawback here is the need for 1 mg of nanoparticles to complete the assay.
- Hemolysis is assayed by a quantitative colorimetric determination of hemoglobin in whole blood, with the end result expressed as percentage hemolysis.
- Platelet aggregation is expressed as the percentage of active platelets per sample compared to a control baseline sample determined using a Z2 Coulter counter for the analysis of platelet-rich plasma.
- Multiple tests are used to assess the effect of nanoparticles on plasma coagulation, including prothrombin time.
- Complement activation is measured initially by the qualitative determination of total complement activation by Western blot. Anti-C3 specific antibodies recognize both native C3 and its cleaved product, which is a common product of all three complement activation pathways. Positive results elicit additional assays aimed at determining the specific complement activation pathway.

In vitro immunology assays also include cell-based assays including colony-forming units-granular macrophages (CFU-GM), leukocyte proliferation, macrophage/neutrophil function and cytotoxic activity of natural killer (NK) cells. The effect of nanoparticles on the proliferation and differentiation of murine bone marrow hematopoietic stem cells (HSC) is monitored by measuring the number of colony forming units (CFU) in the presence and absence of nanoparticles. The effect of nanoparticles on lymphocyte proliferation is determined in similar manner. The ability of nanoparticles to either induce or suppress proliferation is measured and compared to control induction by phytohemaglutinin. Macrophage/neutrophil function is measured by the analysis of phagocytosis, cytokine induction, chemotaxis and oxidative burst. Similar to earlier targeting studies, nanoparticle internalization is measured, but with respect to classical phagocytic cells rather than to the target populations. A current phagocytosis assay utilizes luminol-dependent chemiluminescence, although alternative detection dyes must be used for nanoparticles that interfere with measurements. Cytokine production induced by nanoparticles is measured using white blood cells isolated from human blood. Following particle incubation, the cell culture supernatants are collected and analyzed for the presence of cytokines, using cytometry beads. The chemoattractant capacity of nanoparticles is measured used a cell migration assay; here, cell migration through a 3 μm filter towards test nanoparticles is quantitated using a fluorescent dye. The final measure of macrophage activation is a measure of nitric oxide production using the Greiss reagent. NK-mediated cytotoxicity can be measured by radioactive release assays, in which labeled target cells release radioactivity upon cytolysis by NK cells. A new label-free assay known as 'xCELLigence' (available from Roche) is used to measure the electrical impedance of cells attached to the bottom of a microtiter plate containing cell sensor arrays. In this system, any changes in cell morphology, number, size or attachment are detected in real time.

3.7.2.1 *In Vitro* Toxicity Testing

Standard assays for toxicity assess oxidative stress, necrosis, apoptosis and metabolic stability. Oxidative stress is quantified as a measure of glutathione (GSH) reduction, lipid peroxidation and reactive oxygen species (ROS) in cells treated with nanoparticles. These are measured using colorimetric and fluorescence assays. Cytotoxicity can be measured by using two assays: reduction of 3-(4,5-dimethyl-2-thiazolyl)-2,5-diphenyl-2*H*-tetrazolium bromide (MTT); and lactate dehydrogenase (LDH) release. The degree of caspase-3 activation is also used as a measure of cytotoxicity as it is an indicator of apoptosis. Assays for metabolic stability include cytochrome P450 (CYP450) and glucuronidation.

3.7.3 *In Vivo* Animal Testing

The final category of assays relies on *in vivo* animal testing. Under this umbrella are included disposition studies, immunotoxicity, dose-range-dependent toxicity and efficacy. The initial disposition of nanoparticles is dependent upon tissue distribution, clearance, half-life and systemic exposure. In the NCL regime, immunotoxicity is measured as a 28-day screen and by immunogenicity testing (repeat-dosing). Dose-dependent toxicity can be evaluated by monitoring blood chemistry, hematology, histopathology and gross pathology. Depending on the nature of the delivery system, the efficacy is measured either by imaging or by therapeutic impact.

One possible route of nanoparticle exposure within the work environment is that of *inhalation*, which in turn creates a need for additional studies that include animal inhalation and intratracheal instillation assays [130, 131]. These additional studies also illicit the need for even more characterization studies, such as determining the dispersion properties of nanoparticles. Hence, new methods to determine not only hazard and risk assessments but also therapeutic efficacy continue to be developed as new areas of concern arise. The careful characterization and optimized bioengineering of both nanoparticles and microparticles represent key contributors to the generation of nanomedical devices with optimal delivery and cellular interaction features.

3.8 A Compendium of Unresolved Issues

Unresolved issues and opportunities live in symbiosis. Programmatically, we welcome even the most daunting challenges, as their mere identification as such – not a simple task in most cases, and invariably one that requires the right timing and knowledge maturation – frequently happens when solutions are conceivable, or well within the reach of the scientific community. With this essentially positive outlook we will list in this section some questions that appear daunting at this time, but are starting to present themselves with finer detail and resolution, indicating in our mind that readers in a few years, if any, will find them to be essentially resolved, and the

value of this section, if any, to be basically that of a message in a bottle across the seas of time – and reading patience.

1. The key issue for all systemically administered drugs (nanotechnological, biological, and chemotherapeutical alike) is the management of biological barriers. Biological targeting is always helpful, under the assumption that a sufficient amount of the bioactive agent successfully navigates the sequential presentation of biological barriers. This is a very stringent and daunting assumption – essentially the success stories of the pharmaceutical world correspond to the largely serendipitous negotiation of a subset of biological barriers, for a given indication, and in a sufficiently large subset of the population. The third-generation nanosystems described above are but a first step toward the development of a general, modular system that can systematically address the biological barriers in their sequential totality. We certainly expect that novel generations, and refinements of nanovector generations 1–3 will be developed, to provide a general solution to the chief problem of biobarriers and biodistribution-by-design.

2. There is no expectation that any single, present or future biobarrier management vectoring system will be applicable to all, or even to most. Personalization of treatment is the focus of great emphasis worldwide – with overwhelming bias toward personalization by way of molecular biology. The vectoring problem, on the other hand, is a combination of biology, physics, engineering, mathematics, and chemistry – with substantial prevalence of the non-biological components of vector design. The evolution of nanotechnologies makes it conceivable, that personalization of treatment will develop as combination of biological methods, and vector design based on non-biological sciences. Foundational elements of mathematics-based methods of rational design of vectors were disclosed in the preceding chapters. The missing link to personalized therapy at this time is the refinement of imaging technologies that can be used to identify the characteristics of the target pathologies – lesion-by-lesion, at any given time, and with the expectation that a time evolution will occur – providing the basis for the synthesis of personalized vectors, which may then carry bioactive agents that may be further personalized for added therapeutic optimization. The word 'personalization' does not begin to capture the substance of this proposition; perhaps 'individualization' is a better term, with the understanding that treatment would individualize at the lesion level (or deeper) in a time-dynamic fashion, rather than at the much coarser level of a individual patient at a given time.

3. Hippocrates left no doubt that safety is first and foremost. The conjoined twin of personalized treatment by biodistribution design is the adverse collateral event by drug concentration at unintended location. Safety – and the regulatory approval pathways that are intended to ensure it – in a more advanced sense that the current observation of macroscopic damage requires an accurate determination of the biodistribution of the administered agents. This would be an ideal objective for all drugs, to be sure, but arguably an impossible one in general. Here is where a challenge turns into an opportunity for the nanomedicine: The ability of nanovectors

to be or carry active agents of therapy, while at the same time be or carry contrast agents that allow the tracking and monitoring of their biodistribution in real time provides the nanopharmaceutical world with a unique advantage. Alas reality at this time smiles much less than this vision would entail: In general it proves very difficult at this time to comprise or conjugate nanovectors with contrast agents or nuclear tracers in a manner that is stable in-vivo. Forming a construct that will not separate in their components once systemically administered is a difficult general conjugation chemistry proposition. Less than total success at it means that what is being tracked may be the label rather than the vector or the drug. The problem becomes combinatorially more complex with increasing numbers of nanovector components. Another facet of the same problem is that frequently the conjugation of labels, contrast agents or tracers dramatically alters biodistribution with respect to the construct that is intended for therapeutic applications. One strategic recommendation that naturally emerges for the immediate future is to prioritize nanovectors that are themselves easily traceable with current radiological imaging modalities.

4. Again in common with all drugs, the development and clinical deployment of nanomedicines would greatly benefit from the development of methods for the determination of toxicity and efficacy indicators from non-invasive or minimally invasive procedures such as blood draws. Serum or plasma proteomics and peptidomics are a promising direction toward this elusive goal. The challenge-turned-into-competitive advantage for nanomedicine is in the ability of nanovectors to carry reporters of location and interaction, which can be released into the blood stream and collected therefrom, to provide indications of toxicity and therapeutic effect.

5. Individualization by rational design of carriers together with biological optimization of drug – both informed by imaging and biological profiling – are a dimension of progress toward optimal therapeutic index for all. Another dimension is in the time dynamics of release: the right drug at the right place at the right time is the final objective. With their exquisite control of size, shape, surface chemistry and overall deign parameters, nanovectors are outstanding candidates for controlled release by implanted (nano)devices or (nano)materials; yet another case of challenge turned opportunity, and of synergistic application of multiple nanotechnologies to form a higher generation nanosystem.

6. The last and perhaps most important challenge ahead – and a wide, extraordinarily exciting prairie of opportunities for rides of discovery – is the generation of novel biological hypotheses. With the higher-order nanotechnologies in development, it is possible to reach subcellular target A with nanoparticle species X at time T, and in the same cell and from the same platform then reach subcellular target B with nanoparticle species Y at subsequent time T′. What therapeutic advantages that may bring, for the possible combinations of A, B, X, Y, T and T′ (and extensions-by-induction of the concept) is absolutely impossible to fathom at this time. There is basically little is any science on it – and of course that is the case, since the technology that permits the validating experiment is in its infancy at this time. The

growth of the infant must be accompanied by the co-development of the biological sciences that frame the missing hypotheses and turn their investigation into science. We respectfully suggest that this would change the course of medicine.

References

1 American Cancer Society website, Statistics 08 (2008) http://www.cancer.org/docroot/STT/stt_0.asp (accessed 21 August 2008).

2 Jemal, A., Siegel, R., Ward, E., Murray, T., Xu, J. and Thun, M.J. (2007) Cancer statistics, 2007. *Cancer Journal for Clinicians*, **57**, 43.

3 Brigger, I. Dubernet, C. and Couvreur, P. (2002) Nanoparticles in cancer therapy and diagnosis. *Advanced Drug Delivery Reviews*, **54**, 631.

4 Jain, R.K. (1987) Transport of molecules in the tumor interstitium: a review. *Cancer Research*, **47**, 3039–3051.

5 Jain, R.K. (1999) Transport of molecules, particles, and cells in solid tumors. *Annual Review of Biomedical Engineering*, **1**, 241.

6 Sanhai, W.R., Sakamoto, J.H., Canady, R. and Ferrari, M. (2008) Seven challenges for nanomedicine. *Nature Nanotechnology*, **3**, 242.

7 Sakamoto, J., Annapragada, A., Decuzzi, P. and Ferrari, M. (2007) Antibiological barrier nanovector technology for cancer applications. *Expert Opinion on Drug Delivery*, **4**, 359.

8 Ferrari, M. (2005) Nanovector therapeutics. *Current Opinion in Chemical Biology*, **9**, 343.

9 Kerbel, R.S. (2008) Tumor angiogenesis. *The New England Journal of Medicine*, **358**, 2039.

10 Hobbs, S.K., Monsky, W.L., Yuan, F., Roberts, W.G., Griffith, L., Torchilin, V.P. and Jain, R.K. (1998) Regulation of transport pathways in tumor vessels: role of tumor type and microenvironment. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 4607.

11 Yuan, F., Dellian, M., Fukumura, D., Leuning, M., Berk, D.D., Yorchilin, P. and Jain, R.K. (1995) *Cancer Research*, **55**, 3752.

12 Links, M. and Brown, R. (1999) Vascular permeability in a human tumor xenograft: molecular size dependence and cutoff size. *Expert Reviews in Molecular Medicine*, **1**, 1.

13 Krishna, R. and Mayer, L.D. (2000) Multidrug resistance (MDR) in cancer. Mechanisms, reversal using modulators of MDR and the role of MDR modulators in influencing the pharmacokinetics of anticancer drugs. *European Journal of Cancer Science*, **11**, 265.

14 Ferrari, M. (2005) Cancer nanotechnology: opportunities and challenges. *Nature Reviews Cancer*, **5**, 161.

15 Goodman, L.S., Wintrobe, M.M., Dameshek, W., Goodman, M.J., Gilman, A. and McLennan, M.T. (1946) Landmark article 21 September 1946: Nitrogen mustard therapy. Use of methyl-bis(beta-chloroethyl)amine hydrochloride and tris(beta-chloroethyl)amine hydrochloride for Hodgkin's disease, lymphosarcoma, leukemia and certain allied and miscellaneous disorders. *Journal of the American Medical Association*, **105**, 475, Reprinted in, *Journal of the American Medical Association*, **1984**, 251, 2255.

16 Gilman, A. (1963) The initial clinical trial of nitrogen mustard. *American Journal of Surgery*, **105**, 574.

17 Baxevanis, C.N., Perez, S.A. and Papamichail, M. (2008) Combinatorial treatments including vaccines, chemotherapy and monoclonal antibodies for cancer therapy. *Cancer Immunology, Immunotherapy*, Epub ahead of print, DOI 10.1007/s00262-008-0576-4.

18 Zitvogel, L., Apetoh, L., Ghiringhelli, F. and Kroemer, G. (2008) Immunological aspects of cancer chemotherapy. *Nature Reviews Immunology*, **8**, 59.

19 National Cancer Institute website (2008) http://www.cancer.gov/ (accessed 12 October 2008).

20 Folkman, J. (1971) Tumor angiogenesis: therapeutic implications. *The New England Journal of Medicine*, **285**, 1182.

21 Folkman, J. (2007) Angiogenesis: an organizing principle for drug discovery? *Nature Reviews Drug Discovery*, **6**, 273.

22 Ferrara, N., Hillan, K.J., Gerber, H.P. and Novotny, W. (2004) Discovery and development of bevacizumab, an anti-VEGF antibody for treating cancer. *Nature Reviews Drug Discovery*, **3**, 391.

23 Hurwitz, H., Fehrenbacher, L., Novotny, W. *et al.* (2004) Bevacizumab plus irinotecan, fluorouracil, and leucovorin for metastatic colorectal cancer. *The New England Journal of Medicine*, **350**, 2335.

24 Sandler, A., Gray, R., Perry, M.C. *et al.* (2006) Paclitaxel-carboplatin alone or with bevacizumab for non-small-cell lung cancer. *The New England Journal of Medicine*, **355**, 2542.

25 Faivre, S., Demetri, G., Sargent, W. and Raymond, E. (2007) Molecular basis for sunitinib efficacy and future clinical development. *Nature Reviews Drug Discovery*, **6**, 734.

26 Motzer, R.J., Michaelson, M.D., Redman, B.G. *et al.* (2006) Activity of SU11248, a multitargeted inhibitor of vascular endothelial growth factor receptor and platelet-derived growth factor receptor, in patients with metastatic renal cell carcinoma. *Journal of Clinical Oncology*, **24**, 16.

27 Escudier, B., Eisen, T., Stadler, W.M. *et al.* (2007) Sorafenib in advanced clear-cell renal-cell carcinoma. *The New England Journal of Medicine*, **356**, 125.

28 Llovet, J., Ricci, S., Mazzaferro, V. *et al.* (2008) Sorafenib in advanced hepatocellular carcinoma. *The New England Journal of Medicine*, **359**, 378.

29 Berenson, A. (2006) (15 February) New York Times.

30 Eskens, F.A. and Verweij, J. (2006) The clinical toxicity profile of vascular endothelial growth factor (VEGF) and vascular endothelial growth factor receptor (VEGFR) targeting angiogenesis inhibitors; a review. *European Journal of Cancer (Oxford, England: 1990)*, **42**, 3127.

31 Verheul, H.M. and Pinedo, H.M. (2007) Possible molecular mechanisms involved in the toxicity of angiogenesis inhibition. *Nature Reviews Cancer*, **7**, 475.

32 Jain, R.K., Duda, D.G., Clark, J.W. and Loeffler, J.S. (2006) Lessons from phase III clinical trials on anti-VEGF therapy for cancer. *Nature Clinical Practice Oncology*, **3**, 24.

33 Kerbel, R.S. (2006) Antiangiogenic therapy: a universal chemosensitization strategy for cancer? *Science*, **312**, 1171.

34 Bergers, G. and Benjamin, L.E. (2003) Tumorigenesis and the angiogenic switch. *Nature Reviews Cancer*, **3**, 401.

35 Pressman, D. and Korngold, L. (1953) The in vivo localization of anti-Wagner-osteogenic-sarcoma antibodies. *Cancer*, **6**, 619.

36 Burnet, F.M. (1967) Immunological aspects of malignant disease. *Lancet*, **1**, 1171.

37 Krzeslak, A., Pomorski, L., Gaj, Z. and Lipinska, A. (2003) Differences in glycosylation of intracellular proteins between benign and malignant thyroid neoplasms. *Cancer Letters*, **196**, 101.

38 Mehl, A.M., Fischer, N., Rowe, M. *et al.* (1998) Isolation and analysis of two strongly transforming isoforms of the Epstein-Barr virus (EBV)-encoded latent membrane protein-1 (LMP1) from a single Hodgkin's lymphoma. *International Journal of Cancer*, **76**, 194.

39 Clark, S.S., McLaughlin, J., Timmons, M. *et al.* (1988) Expression of a distinctive BCR-ABL oncogene in Ph1-positive acute lymphocytic leukemia (ALL). *Science*, **239**, 775.

40 Slamon, D.J., Godolphin, W., Jones, L.A. *et al.* (1989) Studies of the HER2/neu proto-oncogene in human breast and ovarian cancer. *Science*, **244**, 707.

41 Slamon, D.J., Clark, G.M., Wong, S.G., Levin, W.J., Ullrich, A. and McGuire, W.L. (1987) Human breast cancer: Correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science*, **235**, 177.

42 Dillman, R.O. (2001) Monoclonal antibody therapy for lymphoma: an update. *Cancer Practice*, **9**, 71.

43 Countouriotis, A., Moore, T.B. and Sakamoto, K.M. (2002) Cell surface antigen and molecular targeting in the treatment of hematologic malignancies. *Stem Cells (Dayton, Ohio)*, **20**, 215.

44 Leget, G.A. and Czuczman, M.S. (1998) Use of Rituximab, the new FDA-approved antibody. *Current Opinion in Oncology*, **10**, 548.

45 Khawli, L.A., Miller, G.K. and Epstein, A.L. (1994) Effect of seven new vasoactive immunoconjugates on the enhancement of monoclonal antibody uptake in tumors. *Cancer*, **73**, 824.

46 Goldenberg, D.M. (1988) Targeting of cancer with radiolabeled antibodies. Prospects for imaging and therapy. *Archives of Pathology and Laboratory Medicine*, **112**, 580.

47 Epenetos, A.A., Snook, D., Durbin, H., Johnson, P.M. and Taylor-Papadimitriou, J. (1986) Limitations of radiolabeled monoclonal antibodies for localization of human neoplasms. *Cancer Research*, **46**, 3183.

48 Cobleigh, M.A., Vogel, C.L., Tripathy, D. *et al.* (1999) Multinational study of the efficacy and safety of humanized anti-HER2 monoclonal antibody in women who have HER2-overexpressing metastatic breast cancer that has progressed after chemotherapy for metastatic disease. *Journal of Clinical Oncology*, **17**, 2639.

49 Baselga, J., Tripathy, D., Mendelsohn, J. *et al.* (1996) Phase II study of weekly intravenous recombinant humanized anti-p185HER2 monoclonal antibody in patients with HER2/neu-overexpressing metastatic breast cancer. *Journal of Clinical Oncology*, **14**, 737.

50 Silver, D.A., Pellicer, I., Fair, W.R., Heston, W.D. and Cordon-Cardo, C. (1997) Prostate-specific membrane antigen expression in normal and malignant human tissues. *Clinical Cancer Research*, **3**, 81.

51 Sweat, S.D., Pacelli, A., Murphy, G.P. and Bostwick, D.G. (1998) Prostate-specific membrane antigen expression is greatest in prostate adenocarcinoma and lymph node metastases. *Urology*, **52**, 637.

52 Wright, G.L., Grob, B.M., Haley, C. *et al.* (1996) Upregulation of prostate-specific membrane antigen after androgen-deprivation therapy. *Urology*, **48**, 326.

53 Murphy, G.P., Elgamal, A.A., Su, S.L., Bostwick, D.G. and Holmes, E.H. (1998) Current evaluation of the tissue localization and diagnostic utility of prostate specific membrane antigen. *Cancer*, **83**, 2259.

54 Smith-Jones, P.M., Vallabhajosula, S., Navarro, V., Bastidas, D., Goldsmith, S.J. and Bander, N.H. (2003) Radiolabeled monoclonal antibodies specific to the extracellular domain of prostate-specific membrane antigen: preclinical studies in nude mice bearing LNCaP human prostate tumor. *Journal of Nuclear Medicine*, **44**, 610.

55 McDevitt, M.R., Barendswaard, E., Ma, D. *et al.* (2000) An alpha-particle emitting antibody ([213Bi]J591) for radioimmunotherapy of prostate cancer. *Cancer Research*, **60**, 6095.

56 Bander, N.H., Nanus, D.M., Milowsky, M.I., Kostakoglu, L., Vallabahajosula, S. and Goldsmith, S.J. (2003) Targeted systemic therapy of prostate cancer with a monoclonal antibody to prostate-specific membrane antigen. *Seminars in Oncology*, **30**, 667.

57 Feynman, R. (1960) There's plenty of room at the bottom. *Engineering and Science*, **23**, 22.

58 National Nanotechnology Initiative program . (2008) http://www.nano.gov/NNI_FY09_budget_summary.pdf (accessed 12 October 2008).

59 Theis, T., Parr, D., Binks, P., Ying, J., Drexler, K.E., Schepers, E., Mullis, K., Bai, C., Boland, J.J., Langer, R., Dobson, P., Rao, C.N. and Ferrari, M. (2006) nan' o.tech.nol' o.gy n. *Nature Nanotechnology*, **1**, 8.

60 Heath, J.R. and Davis, M.E. (2008) Nanotechnology and cancer. *Annual Review of Medicine*, **59**, 251.

61 Nie, S., Kim, G.J., Xing, Y. and Simons, J.W. (2007) Nanotechnology applications in cancer. *Annual Review of Biomedical Engineering*, **9**, 257.

62 Riehemann, K., Schneider, S.W., Luger, T.A., Godin, B., Ferrari, M. and Fuchs, H. (2008) Nanomedicine - Developments and perspectives. *Angewandte Chemie - International Edition*, in press, DOI : 10.1002/ange. 200802585.

63 Harris, J.M. and Chess, R.B. (2003) Effect of pegylation on pharmaceuticals. *Nature Reviews Drug Discovery*, **2**, 214.

64 Brannon-Peppas, L. and Blanchette, J.O. (2004) Nanoparticle and targeted systems for cancer therapy. *Advanced Drug Delivery Reviews*, **56**, 1649.

65 Torchilin, V.P. (2007) Targeted pharmaceutical nanocarriers for cancer therapy and imaging. *The APS Journal*, **9**, E128.

66 Saul, J.M., Annapragada, A.V. and Bellamkonda, R.V. (2006) A dual-ligand approach for enhancing targeting selectivity of therapeutic nanocarriers. *Journal of Controlled Release*, **114**, 277.

67 Yang, X., Wang, H., Beasley, D.W. *et al.* (2006) Selection of thioaptamers for diagnostics and therapeutics. *Annals of the New York Academy of Sciences*, **116**, 1082.

68 Souza, G.R., Christianson, D.R., Staquicini, F.I. *et al.* (2006) Networks of gold nanoparticles and bacteriophage as biological sensors and cell-targeting agents. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 1215.

69 Maeda, H., Wu, J., Sawa, T., Matsumura, Y. and Hori, K. (2000) Tumor vascular permeability and the EPR effect in macromolecular therapeutics: a review. *Journal of Controlled Release*, **65**, 271.

70 Duncan, R. (2006) Polymer conjugates as anticancer nanomedicines. *Nature Reviews Cancer*, **6**, 688.

71 Torchilin, V.P. (2005) Recent advances with liposomes as pharmaceutical carriers. *Nature Reviews Drug Discovery*, **4**, 145.

72 Romberg, B., Hennink, W.E. and Storm, G. (2008) Sheddable coatings for long-circulating nanoparticles. *Pharmaceutical Research*, **25**, 55.

73 Gabizon, A. and Martin, F. (1997) Polyethylene glycol-coated (pegylated) liposomal doxorubicin. Rationale for use in solid tumours. *Drugs*, **54**, 15.

74 Bangham, A.D., Standish, M.M. and Watkins, J.C. (1965) The action of steroids and streptolysin S on the permeability of phospholipid structures to cations. *Journal of Molecular Biology*, **13**, 238.

75 Drummond, D.C., Meyer, O., Hong, K., Kirpotin, D.B. and Papahadjopoulos, D. (1999) Optimizing liposomes for delivery of chemotherapeutic agents to solid tumors. *Pharmacological Reviews*, **51**, 691.

76 Hofheinz, R.D., Gnad-Vogt, S.U., Beyer, U. and Hochhaus, A. (2005) Liposomal encapsulated anti-cancer drugs. *Anti-Cancer Drugs*, **16**, 691.

77 Parveen, S. and Sahoo, S.K. (2006) Nanomedicine: clinical applications of polyethylene glycol conjugated proteins and drugs. *Clinical Pharmacokinetics*, **45**, 965.

78 Zhang, L., Gu, F.X., Chan, J.M., Wang, A.Z., Langer, R.S. and Farokhzad, O.C. (2007) Nanoparticles in medicine: therapeutic applications and developments. *Clinical Pharmacology and Therapeutics*, **83**, 761.

79 Peer, D., Karp, J.M., Hong, S.Y., Farokhzad, O., Margalit, R. and Langer, R.

(2007) Nanocarriers as an emerging platform for cancer therapy. *Nature Nanotechnology*, **2**, 751.

80 Gradishar, W.J., Tjulandin, S., Davidson, N., Shaw, H., Desai, N., Bhar, P., Hawkins, M. and O'Shaughnessy, J. (2005) Phase III trial of nanoparticle albumin-bound paclitaxel compared with polyethylated castor oil-based paclitaxel in women with breast cancer. *Journal of Clinical Oncology*, **23**, 7794.

81 Ringsdorf H. (1975) Structure and properties of pharmacologically active polymers. *Journal of Polymer Science Polymer Symposium*, **51**, 135.

82 Vasey, P.A., Kaye, S.B., Morrison, R., Twelves, C., Wilson, P., Duncan, R., Thomson, A.H., Murray, L.S., Hilditch, T.E. and Murray, T. (1999) Phase I clinical and pharmacokinetic study of PK1 [*N*-(2-hydroxypropyl)methacrylamide copolymer doxorubicin]: first member of a new class of chemotherapeutic agents-drug-polymer conjugates. Cancer Research Campaign Phase I/II Committee. *Clinical Cancer Research*, **5**, 83.

83 Allen, T.M. (2002) Ligand-targeted therapeutics in anticancer therapy. *Nature Reviews Drug Discovery*, **2**, 750.

84 Juweid, M., Neumann, R., Paik, C., Perez-Bacete, M.J., Sato, J., van Osdol, W. and Weinstein, J.N. (1992) Micropharmacology of monoclonal antibodies in solid tumors: direct experimental evidence for a binding site barrier. *Cancer Research*, **52**, 5144.

85 Banerjee, R.K., van Osdol, W., Bungay, P.M., Sung, C. and Dedrick, R.L. (2001) Finite element model of antibody penetration in a prevascular tumor nodule embedded in normal tissue. *Journal of Controlled Release*, **74**, 193.

86 Adams, G.P., Schier, R., McCall, A.M., Simmons, H.H., Horak, E.M., Alpaugh, R.K., Marks, J.D. and Weiner, L.M. (2001) High affinity restricts the localization and tumor penetration of single-chain fv antibody molecules. *Cancer Research*, **61**, 4750.

87 Goren, D., Horowitz, A.T., Zalipsky, S., Woodle, M.C., Yarden, Y. and Gabizon, A. (1996) Targeting of stealth liposomes to erbB-2 (Her/2) receptor: in vitro and in vivo studies. *British Journal of Cancer*, **74**, 1749.

88 Langer, R. (1998) Drug delivery and targeting. *Nature*, **392**, 5.

89 Kang, J., Lee, M.S., Copland, J.A., III, Luxon, B.A. and Gorenstein, D.G. (2008) Combinatorial selection of a single stranded DNA thioaptamer targeting TGF-beta1 protein. *Bioorganic and Medicinal Chemistry Letters*, **18**, 1835.

90 Hajitou, A., Trepel, M., Lilley, C.E. *et al.* (2006) A hybrid vector for ligand-directed tumor targeting and molecular imaging. *Cell*, **125**, 385.

91 Qian, X., Peng, X.H., Ansari, D.O., Yin-Goen, Q., Chen, G.Z., Shin, D.M., Yang, L., Young, A.N., Wang, M.D. and Niel, S. (2008) In vivo tumor targeting and spectroscopic detection with surface-enhanced Raman nanoparticle tags. *Nature Biotechnology*, **26**, 83.

92 Duncan, R. (2003) The dawning era of polymer therapeutics. *Nature Reviews Drug Discovery*, **2**, 347.

93 Hirsch, L.R., Stafford, R.J., Bankson, J.A., Sershen, S.R., Rivera, B., Price, R.E., Hazle, J.D., Halas, N.J. and West, J.L. (2003) Nanoshell-mediated near-infrared thermal therapy of tumors under magnetic resonance guidance. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 13549.

94 Douziech-Eyrolles, L., Marchais, H., Herve, K., Munnier, E., Souce, M., Linassier, C., Dubois, P. and Chourpa, I. (2007) Nanovectors for anticancer agents based on superparamagnetic iron oxide nanoparticles. *International Journal of Nanomedicine*, **2**, 541.

95 Schroeder, A., Avnir, Y., Weisman, S., Najajreh, Y., Gabizon, A., Talmon, Y., Kost, J. and Barenholz, Y. (2007) Controlling liposomal drug release with low frequency ultrasound: mechanism and feasibility. *Langmuir*, **23**, 4019.

96 Monsky, W.L., Kruskal, J.B., Lukyanov, A.N., Girnun, G.D., Ahmed, M., Gazelle, G.S., Huertas, J.C., Stuart, K.E., Torchilin, V.P. and Goldberg, S.N. (2002) Radio-frequency ablation increases intratumoral liposomal doxorubicin accumulation in a rat breast tumor model. *Radiology*, **224**, 823.

97 Farokhzad, O.C., Cheng, J., Teply, B.A., Sherifi, I., Jon, S., Kantoff, P.W., Richie, J.P. and Langer, R. (2006) Targeted nanoparticle-aptamer bioconjugates for cancer chemotherapy in vivo. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 6315.

98 Farokhzad, O.C., Karp, J.M. and Langer, R. (2006) Nanoparticle-aptamer bioconjugates for cancer targeting. *Expert Opinion on Drug Delivery*, **3**, 311.

99 Northfelt, D.W., Martin, F.J., Working, P., Volberding, P.A., Russell, J., Newman, M., Amantea, M.A. and Kaplan, L.D. (1996) Doxorubicin encapsulated in liposomes containing surface-bound polyethylene glycol: pharmacokinetics, tumor localization, and safety in patients with AIDS-related Kaposi's sarcoma. *Journal of Clinical Pharmacology*, **36**, 55.

100 Jang, S.H., Wientjes, M.G., Lu, D. and Au, J.L. (2003) Drug delivery and transport to solid tumors. *Pharmaceutical Research*, **20**, 1337.

101 Lankelma, J., Dekker, H., Luque, F.R., Luykx, S., Hoekman, K., van der Valk, P., van Diest, P.J. and Pinedo, H.M. (1999) Doxorubicin gradients in human breast cancer. *Clinical Cancer Research*, **5**, 1703.

102 Tannock, I.F., Lee, C.M., Tunggal, J.K., Cowan, D.S. and Egorin, M.J. (2002) Limited penetration of anticancer drugs through tumor tissue: a potential cause of resistance of solid tumors to chemotherapy. *Clinical Cancer Research*, **8**, 878.

103 Klibanov, A.L., Maruyama, K., Beckerleg, A.M., Torchilin, V.P., and Huang, L. (1991) Activity of amphipathic poly (ethylene glycol) 5000 to prolong the circulation time of liposomes depends on the liposome size and is unfavorable for immunoliposome binding to target. *Biochimica et Biophysica Acta*, **1062**, 142.

104 Souza, G.R., Yonel-Gumruk, E., Fan, D. *et al.* (2008) Bottom-up assembly of hydrogels from bacteriophage and Au nanoparticles: the effect of cis- and trans-acting factors. *PLoS ONE*, **3**, e2242.

105 Sengupta, S., Eavarone, D., Capila, I., Zhao, G., Watson, N., Kiziltepe, T. and Sasisekharan, R. (2005) Temporal targeting of tumour cells and neovasculature with a nanoscale delivery system. *Nature*, **436**, 568.

106 Tasciotti, E., Liu, X., Bhavane, R., Plant, K., Leonard, A.D., Price, B.K., Cheng, M.M., Decuzzi, P., Tour, J.M., Robertson, F.M., and Ferrari, M. (2008) Mesoporous silicon particles as a multistage delivery system for imaging and therapeutic applications. *Nature Nanotechnology*, **3**, 151.

107 Ferrari, M. (2008) Nanogeometry: beyond drug delivery. *Nature Nanotechnology*, **3**, 131.

108 Decuzzi, P., Pasqualini, R., Arap, W. and Ferrari, M. (2008) Intravascular delivery of particulate systems: Does geometry really matter? *Pharmaceutical Research*, 20 August, Epub ahead of print.

109 Goldsmith, H.L. and Spain, S. (1984) Margination of leukocytes in blood flow through small tubes. *Microvascular Research*, **27**, 204.

110 Fahraeus, R. and Lindqvist, T. (1931) The viscosity of the blood in narrow capillary tubes. *The American Journal of Physiology*, **96**, 562.

111 Sharan, M. and Popel, A.S. (2001) A two-phase model for flow of blood in narrow tubes with increased effective viscosity near the wall. *Biorheology*, **38**, 415.

112 Goldman, A.J., Cox, R.G. and Brenner, H. (1967) Slow viscous motion of a sphere parallel to a plane wall. II. Couette flow. *Chemical Engineering Science*, **22**, 653.

113 Decuzzi, P., Lee, S., Bhushan, B. and Ferrari, M. (2005) A theoretical model for the margination of particles within blood vessels. *Annals of Biomedical Engineering*, **33**, 179.

114 Decuzzi, P. and Ferrari, M. (2007) The role of specific and non-specific interactions in receptor-mediated endocytosis of nanoparticles. *Biomaterials*, **28**, 2915–2922.

115 Gavze, E. and Shapiro, M. (1998) Motion of inertial spheroidal particles in a shear flow near a solid wall with special application to aerosol transport in microgravity. *Journal of Fluid Mechanics*, **371**, 59.

116 Filipovic, N., Stojanovic, B., Kojic, N. and Kojic, M. (2008) *Computer Modeling in Bioengineering -Theoretical Background, Examples and Software*. John Wiley & Sons, Chichester, UK.

117 Gentile, F., Chiappini, C., Fine, D., Bhavane, R.C., Peluccio, M.S., Ming-Cheng Cheng, M., Liu, X., Ferrari, M. and Decuzzi, P. (2008) The effect of shape on the margination dynamics of non-neutrally buoyant particles in two-dimensional shear flows. *Journal of Biomechanics*, **41**, 2312.

118 Decuzzi, P. and Ferrari, M. (2006) The adhesive strength of non-spherical particles mediated by specific interactions. *Biomaterials*, **27**, 5307.

119 Gao, H., Shi, W. and Freund, L.B. (2005) Mechanics of receptor-mediated endocytosis. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 9469.

120 Decuzzi, P. and Ferrari, M. (2008) The receptor-mediated endocytosis of nonspherical particles. *Biophysical Journal*, **94**, 3790.

121 Decuzzi, P. and Ferrari, M. (2008) Design maps for nanoparticles targeting the diseased microvasculature. *Biomaterials*, **29**, 377.

122 The National Cancer Institute, the Nanotechnology Characterization Laboratory (NCL) (2008) http://ncl.cancer.gov (accessed 12 October 2008).

123 Hall, J.B., Dobrovolskaia, M.A., Patri, A.K. and McNeil, S.E. (2007) Characterization of nanoparticles for therapeutics. *Nanomedicine*, **2**, 789.

124 Pelkmans, L., Kartenbeck, J. and Helenius, A. (2001) Caveolar endocytosis of simian virus 40 reveals a new two-step vesicular-transport pathway to the ER. *Nature Cell Biology*, **3**, 473.

125 Serda, R.E., Adolphi, N.L., Bisoffi, M. and Sillerud, L.O. (2007) Targeting and cellular trafficking of magnetic nanoparticles for prostate cancer imaging. *Molecular Imaging*, **6**, 277.

126 Serda, R.E. Gu J, J., Bhavane, R.C., Liu, W., Chiappini, C., Robertson, F., Decuzzi, P. and Ferrari, M. (2008) Microengineering delivery vectors to target inflamed vascular endothelium and reduce RES uptake. (submitted).

127 Tjelle, T.E., Lovdal, T. and Berg, T. (2000) Phagosome dynamics and function. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, **22**, 255.

128 Davda, J. and Labhasetwar, V. (2002) Characterization of nanoparticle uptake by endothelial cells. *International Journal of Pharmaceutics*, **233**, 51.

129 Dobrovolskaia, M.A. and McNeil, S.E. (2007) Immunological properties of engineered nanomaterials. *Nature Nanotechnology*, **2**, 469.

130 Morimoto, Y. and Tanaka, I. (2008) Effects of nanoparticles on humans. *Sangyo Eiseigaku Zasshi*, **50**, 37.

131 Chen, J., Tan, M., Nemmar, A., Song, W., Dong, M., Zhang, G. and Li, Y. (2006) Quantification of extrapulmonary translocation of intratracheal-instilled particles in vivo in rats: effect of lipopolysaccharide. *Toxicology*, **222**, 195.

Part Three:
Imaging and Probing the Inner World of Cells

Nanotechnology, Volume 5: Nanomedicine. Edited by Viola Vogel

ISBN: 978-3-527-31736-3

4
Electron Cryomicroscopy of Molecular Nanomachines and Cells

Matthew L. Baker, Michael P. Marsh, and Wah Chiu

4.1
Introduction

Genome-sequencing projects continue to provide complete genetic descriptions of an ever-increasing number of model organisms [1–4]. Based on our current knowledge, it has been estimated that life depends on 200–300 core biological processes [5]. Individual gene products rarely function independently; to the contrary, large multicomponent protein assemblies are more often responsible for complex cellular functions. These assemblies are often dynamic and, in many cases, transient. As such, these assemblies are often termed 'molecular nanomachines', capable of carrying out a wide range of functions through often specific and highly intricate interactions [6–10].

Equally as complex as the nanomachines themselves, the individual components can adopt a wide variety of morphologies, functions and interactions. In addressing these complexities, structural genomics seeks to provide a description of all protein folds, where a fold is defined as the three-dimensional (3-D) structure of protein that relates the spatial arrangements and connectivity of secondary structure elements, such as α-helices and β-sheets. As such, the protein fold represents the basic 'building block' of much larger and more complex assemblies that carry out biochemical and cellular processes. To date, more than 51 000 individual protein structures are known [11]; however, far fewer unique folds are recognized.

It is generally accepted that the primary structure of a protein – the amino acid sequence – dictates its 3-D structure, or fold. As such, proteins with a similar primary sequence likely assume similar folds. However, the converse is not necessarily true; proteins with vastly dissimilar sequences can assume similar folds. Nonetheless, the protein fold is ultimately responsible for the necessary 3-D environment for protein function and intermolecular and intramolecular interactions. The description of these folds – and, in particular, their interactions within cellular complexes – is therefore paramount to the understanding of all molecular nanomachines and biological processes.

Nanotechnology, Volume 5: Nanomedicine. Edited by Viola Vogel

ISBN: 978-3-527-31736-3

Electron cryomicroscopy (cryo-EM) is an emerging methodology that is particularly well suited for studying molecular nanomachines at near-native or chemically defined conditions. Cryo-EM can be used to study nanomachines of various sizes, shapes and symmetries, including two-dimensional (2-D) arrays, helical arrays and single particles [12]. With recent advances, cryo-EM can now not only reveal the gross morphology of these nanomachines but also provide highly detailed models of protein folds approaching atomic resolutions [13–17]. In this chapter, we will present the methodology of single-particle cryo-EM, as well as its potential biomedical applications and future prospects.

Complementary to structural studies of nanomachines with cryo-EM, the application of cryo-tomography (cryo-ET) can depict the locations and low-resolution structures of nanomachines in a 3-D cellular environment. The power of cryo-ET comes from its unique ability to observe directly biological nanomachines *in situ*, without the need for isolation and purification. This approach has the potential to capture the structural diversity of nanomachines in their milieu of interacting partners and surrounding cellular context.

4.2
Structure Determination of Nanomachines and Cells

Figure 4.1 shows a series of typical steps in imaging nanomachines using cryo-EM or cryo-ET. The first steps are common to both techniques; biochemical preparation, specimen preservation via rapid freezing; and imaging the frozen, hydrated specimens by low-dose electron microscopy. Although the subsequent steps differ for the two techniques, they both include image processing to generate a 3-D reconstruction, interpreting the 3-D volume density together with other biological data and archiving the density maps and models. In this chapter we will not address how to perform each of the aforementioned mentioned steps, as numerous technical reports and books exist that describe them in detail [12, 18]. Rather, we will briefly summarize these steps and their applications to a few examples of molecular nanomachines and cells.

4.2.1
Experimental Procedures in Cryo-EM and Cryo-ET

In principle, most of these steps are rather straightforward, and the length of time taken to start from a highly purified nanomachine to obtaining a complete structure can range from a few days to months. However, as with any experimental method, various hurdles may be encountered that require further optimization before a reliable structure can be determined.

4.2.1.1 Specimen Preparation for Nanomachines and Cells

Specimen preparation is a critical step in single-particle cryo-EM, which necessarily requires high conformational uniformity while preserving functional activities. In X-ray crystallography, crystallization is a selective process through which only

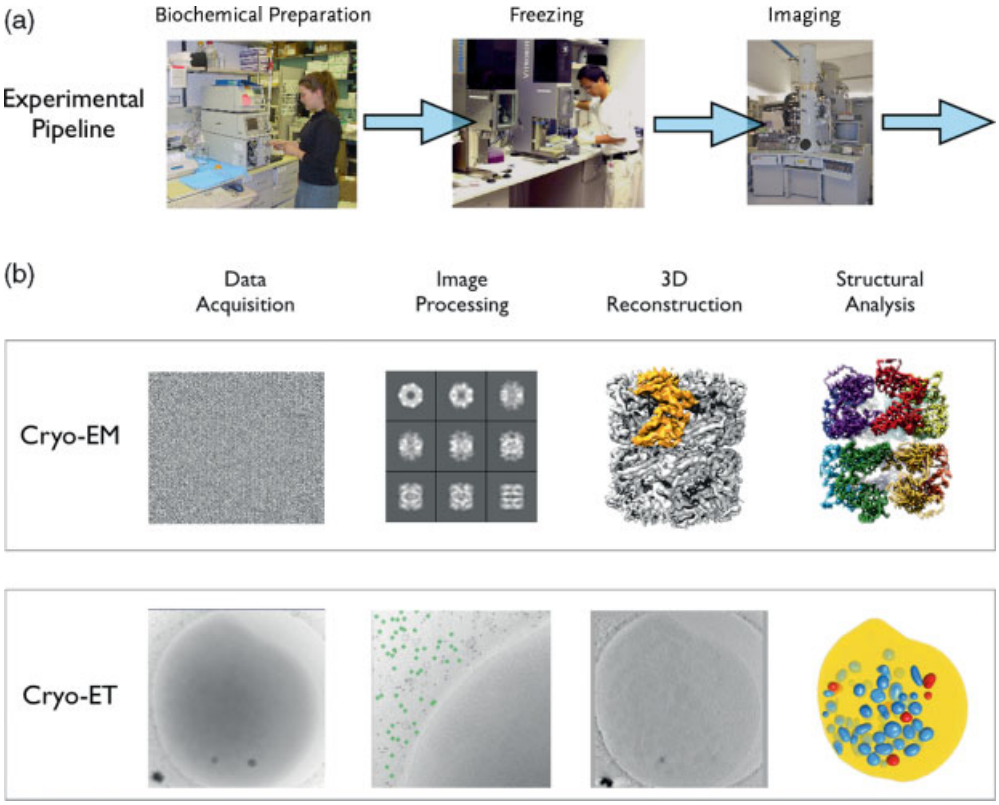


Figure 4.1 The experimental pipeline for cryo-imaging experiments. (a) The first steps of the experiment – specimen preparation, specimen freezing and microcopy – are common to both cryo-EM and cryo-ET; (b) The subsequent steps diverge and differ between the two types of experiments. For cryo-EM, these steps are illustrated with examples of the biological nanomachine GroEL. For cryo-ET, they are illustrated with a cell, the human platelet.

molecules of the same conformations nucleate and crystallize to form a diffracting object. In addition to chemical purification, crystallization also forces the molecules into specific, uniform spatial organization such that diffraction data can be averaged from over billions of molecules in identical conformations. However, cryo-EM experiments image one set of molecules at a time, regardless of their conformations, and thus possibly represents an ensemble of conformations of the molecules in a single micrograph. In order to obtain the highest possible resolution structures by cryo-EM, it is still necessary to computationally average from several hundreds to tens of thousands of a conformationally homogeneous set of particle images recorded in multiple micrographs. Nevertheless, computational methods are being developed to sort out images of particles with different conformations.

The nature of cryo-ET experiments differs substantially from single-particle cryo-EM experiments, and the resolution of the reconstructions is much lower. In contrast to the cryo-EM approach, where images of many conformationally

uniform particles are merged to yield a 3-D model, cryo-ET merges many images of the same specimen target, collected at different angles. With this approach, a reconstruction can be computed from the images of a single cell or nanomachine, and so conformational uniformity is not an issue in the most general case. The merging of whole cells or organelles such as single particles is not a reasonable goal, as uniformity can never realistically be expected; however, some subcelluar structures may be sufficiently uniform in conformation to warrant merging and averaging from the 3-D tomogram. Below, we consider such an example when discussing the bacterial flagella motor.

4.2.1.2 **Cryo-Specimen Preservation**

Following biochemical isolation and purification, the first step in a cryo-imaging experiment is to embed a biochemically purified nanomachine or cell under well-defined chemical conditions in ice on a cryo-EM grid [19]. This freezing process is extremely quick in order to prevent the formation of crystalline ice, and thus produces a matrix of vitreous ice in which the water molecules remain relatively unordered. The spread of the nanomachines on the grid should be neither too crowded, such that they would contact each other, nor too dilute as to only have a few nanomachines recorded in each micrograph. For cryo-EM, it is preferable to have the nanomachines situated in random orientations to allow sufficient angular sampling needed for the subsequent 3-D reconstruction procedure. The ideal thickness of the embedding ice is slightly greater than the size of the nanomachine or cell. Excessive ice thickness is detrimental because it diminishes the signal-to-noise ratio (SNR) of the images that can be acquired. Ice that is too shallow can be a problem for cryo-ET experiments, whereby flattening of the specimen can occur. The capillary forces of the solvent, in the fluid phase just prior to vitrification, can compress the sample; this has been reported in vesicles [20] as well as real cells where a 1 μm-thick cell can be reduced to 600 nm [21, 22].

Some specimens are very easy to prepare, while others are more difficult, which necessarily means optimization of the specimen preparation is a trial-and-error process. In general, this step – the preparation of the frozen, hydrated specimens, preserved in vitreous ice with an optimal ice thickness – is often a bottleneck. Analogous to the crystallization process in X-ray crystallography, there is no foolproof recipe for optimal specimen preservation. However, a computer-driven freezing apparatus has made this step more reproducible and tractable in finding optimal conditions for freezing a given specimen [23].

In principle, the frozen, hydrated specimens represent native conformations as they are maintained in an aqueous buffer. Fixation of the nanomachines in a specific orientation can occur prior to freezing. Specimen freezing can also be coordinated with a time-resolved chemical mixing reaction; prototype apparatuses have been built to perform such a time-resolved reaction [24, 25]. It is conceivable that a more sophisticated instrument can be built to allow all sorts of chemical reactions, including those that can be light-activated. Such an approach would allow cryo-EM to follow the structure variations in a chemical process with a temporal resolution of milliseconds [25].

4.2.1.3 Low-Dose Imaging

Once the sample has been frozen, the entire cryo-EM grid can be inserted into the electron cryo-microscope and imaged with electrons ranging from 100 to 400 keV. Electrons at these energies will damage the molecules during imaging [12]. Therefore, low-dose imaging is necessary to minimize the damage to the specimen before the image is recorded. To maintain the frozen, hydrated specimen in vitreous ice inside the electron microscope vacuum, the specimen is kept at low temperature, typically at or below liquid nitrogen temperature. If the specimen temperature is higher than $-160\,^{\circ}C$, the vitreous ice undergoes a phase transition to crystalline ice and denatures the nanomachines [19].

From a radiation damage perspective, the advantages and disadvantages of keeping biological specimens at different low temperatures have been studied [26–28]. High-quality images have been obtained using liquid helium temperature, and have resulted in high-resolution structures of 2-D crystals [29, 30], helical arrays [31, 32] and single particles [13, 14] where protein backbone traces were feasible. Imaging specimens at liquid nitrogen temperature has also been used successfully for the similar high-resolution structure determination of a broad spectrum of specimens [15, 17, 33, 34]. In the case of cellular cryo-ET, it has been suggested that liquid nitrogen is a preferred temperature [27, 35] because of a significant loss of contrast at liquid helium temperature [36].

4.2.1.4 Image Acquisition

Data collection differs significantly for cryo-EM and cryo-ET. For cryo-EM, images of a field containing multiple, randomly oriented specimens are recorded. Individual particles are recorded only once because of the radiation damage constraints for obtaining the highest possible resolution information. For cryo-ET, a series of images of the same specimen is acquired as the stage is iteratively tilted over an interval spanning approximately 130°. A typical tilt-series might include one image collected every 2° between -65° and $+65^{\circ}$. The resolution of the tomographic data is much lower because of the effects of cumulative radiation damage to the specimen throughout the data collection.

For cryo-EM and cryo-ET experiments, images have been traditionally collected on photographic film and subsequently digitized using a high-resolution film scanner. Recent advances in CCD cameras for electron microscopes have made direct digital recording feasible [37–39]. With a modern electron microscope equipped with specialized software for low-dose imaging, data collection is relatively simple and can be either partially or fully automated [40–43].

4.2.2 Computational Procedures in Cryo-EM and Cryo-ET

The recorded image of a nanomachine is essentially a projection (2-D) of its mass density along the path of the irradiating electrons. In order to retrieve its 3-D structure, the particle must be sampled in different angular views [44]. For cryo-ET, this sampling is carried out systematically whereby each image in the tilt-series

constitutes a separate angular view. With cryo-EM, the varied orientation of particles in the ice naturally provides an angular distribution of views. The number of views required for the reconstruction is proportional to the diameter of the particle and is inversely proportional to the desired resolution [45]. Because of the noisy nature of the image and the uneven angular distribution (in the case of cryo-EM) of the views, the actual number of the particles used to calculate a reconstruction at a certain resolution is much higher than the theoretical minimum [46, 47]. Ideally, the particles embedded in the vitreous ice are oriented randomly. However in some cases, the particles tend to assume a preferred orientation with respect to the surface of the embedding ice. This can often be overcome by varying the buffer or solvent by adding a small amount of detergent.

4.2.2.1 Image Processing and Reconstruction

During the image-processing phase, individual specimen images are aligned with respect to each other and then combined to form a 3-D density map [18, 48, 49]. For cryo-EM studies, the image processing is an iterative process. Several image-processing packages are available for single-particle cryo-EM, such as EMAN [50], SPIDER [51], IMAGIC [52] and Frealign [53]; these are multi-step procedures that can generally be broken into the following steps: (i) identify the locations of each particle; (ii) determine and correct the contrast transfer function and damping function for the particle images; (iii) classify the images according to their conformational identity and orientation parameters; and (iv) average the particle images in each classes and 3-D reconstruct to produce the final 3-D map. The specimen classification, particle averaging and reconstruction of the density map are iterated using the previous iteration as a reference. Iteration of these steps continues until no improvement in the 3-D density maps is made over the previous cycle of refinement. The final resolution of the map is typically assessed by a parameter referred to as Fourier shell correlation (FSC), in which two maps derived from two independent sets of image data are compared [54]. The FSC essentially measures a similarity and reproducibility of two structures in Fourier space; the final resolution is often determined using the 0.5 criterion.

Alignment and reconstruction differ for cryo-ET experiments. For cryo-ET, the nominal angular assignment is known for each image because the tilt-angle of the stage was recorded for each image of the tilt-series. Higher-precision angular assignments must be determined for reconstruction. Alignment processing is frequently simplified by including gold particles in the specimen; these particles have a strong contrast, even under low-dose imaging conditions, and serve as landmarks for registering images with respect to each other [55–57]. Once aligned, the images are then recombined directly by a reconstruction algorithm such as weighted back-projection [58, 59]. Many academically developed processing packages are available that will compute the alignment and reconstruction [60–63]. Although there is no community-accepted convention for assessing the resolution of tomographic reconstructions, a number of statistical approaches have been proposed [64].

4.2.2.2 Structure Analysis and Data Mining

In order to analyze the cryo-EM density maps of large, complex nanomachines, a number of tools have been developed that range from feature detection to domain localization. Perhaps the most well-developed set of tools for the analysis of cryo-EM density maps are those aimed at fitting known crystal or NMR structures to density maps. These tools range from simple rigid-body fitting to complex and dynamic flexible fitting algorithms (for reviews, see Refs [65, 66]). Regardless, each of these tools requires that the structures of a known domain or closely related domain are known.

Recent studies have also shown that cryo-EM density is sufficient for discriminating good models from a gallery of potential structures [67, 68]. In particular, cryo-EM density has been incorporated as a scoring function in a constrained homology modeling approach [69]. As with the aforementioned fitting routines, this approach relies on the availability of a known structure from which a sequence/structure alignment is produced. In the case where a suitable structural template is not known, a constrained *ab initio* modeling approach has also been developed in which the cryo-EM density can be used directly to screen a large gallery of potential models [67]. While no structural template is needed, this approach is restricted to relatively small (<200 amino acids), single-domain proteins.

At subnanometer resolutions, secondary structure elements become visible; α-helices appear as long density rods, while β-sheets appear as thin surfaces [70–73]. By using a variety of feature detection and computational geometry algorithms, secondary structure elements can be reliably identified and quantified [71, 72, 74, 75]. The spatial description of such elements has been used not only to describe protein structure, but also to infer structure and function of individual protein domains [70, 76].

Until recently, the resolution of cryo-EM density necessitated the use of the aforementioned approaches for understanding macromolecular structure and function. Several cryo-EM structures have now achieved resolutions better than 4.5 Å resolution, at which point the pitch of α-helices, separation of β-strands, as well as the densities that connect them, can be seen unambiguously with no reference to crystal structure [13, 14, 17]. In addition to these features, many of the bulky side chains could also be seen. However, it should be noted that these structures still do not have the resolution to utilize standard X-ray crystallographic methods for model construction. However, several *de novo* models have now been constructed directly from these high-resolution cryo-EM density maps using mostly manual assignment and visualization tools [13, 14, 17].

The annotation of cryo-ET maps is a different process because the goals of tomographic imaging are substantially different. Rather than trying to determine high-resolution protein structures, cryo-ET experiments are often focused on how the components of cells or nanomachines are spatially organized. *Segmentation* is a process by which the salient features of the reconstruction of an individual cell or nanomachine are traced. A segmented map hides the noisy data and highlights the structural findings. Tools for segmentation of tomograms include the academic IMOD package [60] and the commercial package Amira (Mercury Systems, Chelmsford, MA, USA). Unfortunately, this is a laborious manual process for which no suitable general-purpose automated routine has been advanced (for a review, see

Ref. [77]). Manual segmentation is applied by annotating one cell or nanomachine at a time. As a research study might require the annotation of many cells or nanomachines, this can drastically increase the time required to annotate the results. In cases where there are uniform nanomachines (e.g. ribosomes) present, these can be computationally identified, extracted and merged to improve the resolution (for a review, see Ref. [78]).

4.2.3 Data Archival

The result of a cryo-EM experiment is typically a 3-D density map with multiple domains and/or models used to annotate the structure and function of the molecular nanomachine. In reaching this model, multiple intermediate data sets and image processing workflows are produced. Databases, such as EMEN [79] and others [80, 81], function on a laboratory scale and can house the final 3-D density maps and model, as well as the original specimen images and all of the intermediate data and processes. The final density map, models and associated metadata can also be deposited in public repositories such as the electron microscopy databank (EMDB) and the protein databank (PDB) [82]. Individual cryo-EM structures are easily retrieved through accession numbers or IDs directly from publicly accessible websites.

4.3 Biological Examples

Cryo-EM is a powerful technique in that it can be used to image a wide variety of specimens under an equally wide array of conditions. Despite the lack of atomic resolution, these structures can provide unprecedented views of the structure and function of molecular nanomachines. In the following sections, we describe two very different samples, and how cryo-EM has provided us with a unique glimpse into their organization. It should also be noted that each of these samples are complex nanomachines that undergo dynamic structural and functional processes in carrying out their intended functions.

The resolution of cryo-EM models is considerably better than that of cryo-ET models, mainly because the heterogeneous cryo-ET particles are rarely suitable for averaging. Even when they are, the average may be the sum of tens to hundreds from extracted tomograms rather than the thousands to millions of asymmetric units in a single-particle reconstruction that contribute to a cryo-EM model. Despite the lower resolution, tremendous insight may still be gained from the cellular context, as evidenced by the two cryo-ET examples presented here.

Both, cryo-EM and cryo-ET offer unique views of nanomachines, and as such integrating the two approaches can generate multiresolution models where a tomogram establishes a low-resolution survey of a cell, and the individual machines in that model are the product of cryo-EM studies. This integrated approached is demonstrated in our last example.

4.3.1
Skeletal Muscle Calcium Release Channel

Ryanodine receptor (RyR1) is a 2.3 MDa homotetramer that regulates the release of Ca^{2+} from the sarcoplasmic reticulum to initiate muscle contraction (for a review, see Ref. [49]). Figure 4.2a shows the 9.6 Å resolution cryo-EM density map of RyR1 reconstructed from ~28 000 particle images (Figure 4.2b) [83]. In this map, the structural organization, including the transmembrane and cytoplasmic regions for each monomer, as well as domains within individual monomers can be clearly seen.

A structural analysis of the RyR1 map using SSEHunter [71] revealed 41 α-helices, 36 in the cytoplasmic region and five in the transmembrane region, as well as seven β-sheets in the cytoplasmic region of a RyR1 monomeric subunit (Figure 4.2c). Interestingly, a kinked inner, pore-lining helix and a pore helix in the transmembrane region bears a remarkable similarity to those of the MthK channel [84]. β-Sheets located in the constricted part that connect the transmembrane and cytoplasmic regions have been seen in the crystal structures of inward rectifier K^+ channels (Kir channels) [85, 86] and a cyclic nucleotide-modulated (HCN2) channel [87]. In Kir channels, this β-sheet has been proposed to form part of the cytoplasmic pore, which is connected to the inner pore. Therefore, this region in the RyR1 may play a role in regulating the ions by interacting with cellular regulators which are yet to be determined.

While there is no crystal structure from any domain or region of RyR1, a homologous domain from the IP3 receptor is known. Using the aforementioned cryo-EM constrained homology modeling approach [69], it was possible to derive three protein folds, based on the ligand-binding suppressor and IP_3-binding core domains from the type 1 IP_3 receptor, for the N-terminal portion (residues 12–565) of the RyR1 primary sequence [88] (Figure 4.2d). Interestingly, these models were localized to a region at the four corners of the RyR1 tetramer, a region that has also been implicated to interact with the dihydropyridine receptor (DHPR) during the

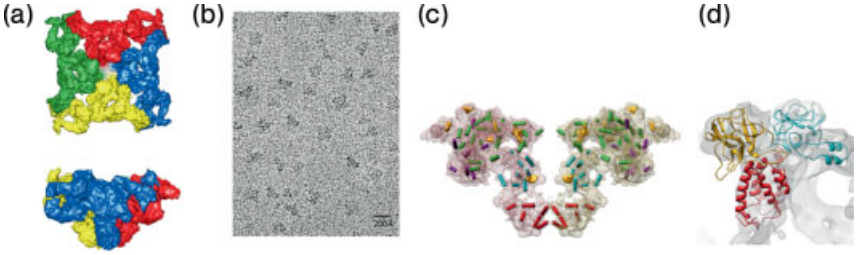


Figure 4.2 RyR1 at 9.6 Å resolution [88]. RyR1 at 9.6 Å resolution. The side and top views shown in (a) were reconstructed from ~28 000 individual particle images. The four subunits are annotated in different colors; (b) A representative view of the particle images; (c) The spatial dispositions of the α-helices (cylinders) and β-sheets (orange planes) in two of the homotetrameric subunits; (d) The three N-terminal models for RyR1 are shown fitted to the cryo-EM density. Model 1 is shown in cyan (residues Q12-S207), model 2 in yellow (residues G216-T407), and model 3 in red (residues A408-Y565). Models 2 and 3 are based on the aforementioned IP_3-binding core domain, while model 1 is based on the ligand-binding suppressor domain.

excitation–contraction coupling of the muscle [89–91]. Also of interest, several disease-related mutations in RyR1 occur within this region. As such, imaging of RyR1 with cryo-EM coupled to structural analysis has resulted into insight into the channels function and role in muscle contraction.

4.3.2
Bacteriophage Epsilon15

Epsilon15 is a 700 Å wide, 22 MDa nanomachine that infects *Salmonella* (Figure 4.3a) [92]. An icosahedral protein shell surrounds its dsDNA genome; at one vertex a large tail assembly protrudes from this shell. Without imposing any symmetry, a reconstruction of the native virion revealed at ~20 Å resolution all the molecular components of the virus, including the portal vertex (Figure 4.3b) [92]. When icosahedral symmetry was imposed during the reconstruction (effectively increasing the number of particle images, as there are 60 asymmetric units per particle), features of the nonicosahedral components such as the portal vertex were averaged out, but the icosahedral position shell proteins could be seen at a finer detail (4.5 Å resolution) [13]. In this high-resolution map, a complete annotation of the capsid components was possible.

While ~4.5 Å resolution is generally insufficient to construct a model by X-ray crystallographic standards, the aforementioned *de novo* model building tools for cryo-EM density maps make it possible with the existence of large α-helices which can be

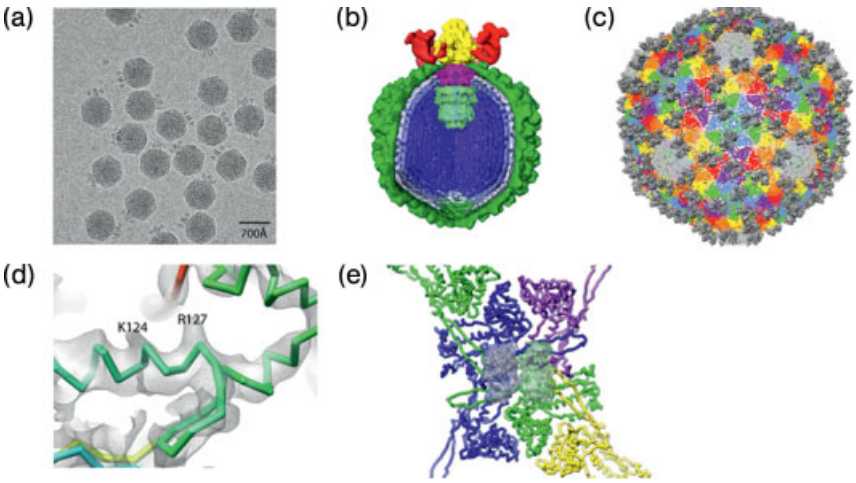


Figure 4.3 Bacteriophage epsilon15. (a) A 300-kV electron image of the phage particles embedded in vitreous ice; (b) A cut-away view of the 20 Å asymmetric reconstruction of epsilon15 [92] shows the molecular components of the portal vertex, the capsid shell protein and the viral DNA. The different molecular components are annotated in different colors; (c) 4.5 Å resolution structure [13], showing the backbone model of the Gp7 and the density of Gp10; (d) Side-chain density in the cryo-EM map; (e) A zoomed-in view of the capsid showing one gp10 spanning across four gp7 molecules, functioning as a 'molecular stapler'.

used to anchor the sequence-to-structure assignment. Gp7, a 420-amino acid protein that makes up the majority of the icosahedral capsid shell, was identified in the reconstruction and shown to have eight α-helices (>2.5 turns) and three β-sheets. Using these features, a complete *de novo* model was constructed (Figure 4.3c) to reveal a structure similar to that of the HK97 major capsid protein [93], despite the lack of detectable sequence similarity. In addition to the detection of a common fold, side-chain density could also be visualized in several regions throughout the map (Figure 4.3d).

Construction of the Gp7 model clearly revealed the presence of a previously undetected capsid protein. Biochemical analysis of the capsid later confirmed this protein to be Gp10 (12 kDa). In analyzing the sequence of this protein, potential structural homology to the PDZ domains [94] was identified. Taken together with the location of this protein of the capsid surface (along the icosahedral twofold symmetry axes), this small protein most likely acts like a molecular staple, bridging four adjacent gp10 molecules and thus assuring stability in the mature capsid (Figure 4.3e).

4.3.3
Bacterial Flagellum

The bacterial flagellum is an intricate biological nanomachine that transduces chemical energy into mechanical energy. The flagellar motor is a complicated assembly of approximately 25 unique polypeptide components that, when assembled correctly and operating under the proper electrochemical gradient, drives the rotary motor at speeds of nearly 300 Hz (for a review, see Ref. [95]). Models of the flagellar motor have been advanced through both single-particle and helical cryo-EM studies of the purified complexes from various spirochetes [96–98]. Cryo-ET has been used to complement these models; the *in situ* perspective derived from tomograms can reveal components of the structure that may be – and if fact, are – lost in the isolation and purification steps.

Recent studies have utilized cryo-ET to examine *in situ* the structure of the motor and the greater flagellar apparatus from the spirochete *Treponema primitia*. In their first report, Jensen and coworkers presented the structural details of the motor [99] in which 20 flagellar motors were computationally isolated from their positions in the cryo-ET reconstructions of 15 intact cells (each cell has two flagella, but not all tomograms captured both). These motors were subsequently merged to yield a ∼70 Å resolution model (Figure 4.4a). Like man-made motors, the flagellar motor consists of a rod attached to a rotor that rotates amidst an array of stationary stators; purified flagellar complexes lacked the stators altogether. This tomography-derived model was the first reconstruction to integrate the full motor with the stators, revealing the stators' 16-fold symmetry and their position with respect to the membrane. Besides accounting for the stator density, this model also revealed unexpected density above the stators in the peptidoglycan layer, including a new component termed the P-collar (Figure 4.4b). The P-collar and new findings about stator geometry raise new questions about the motor's mechanistic details.

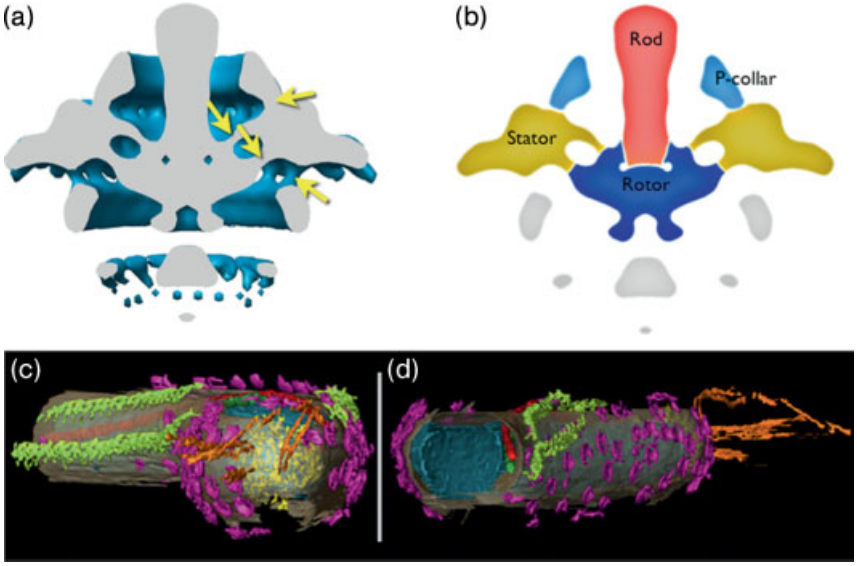


Figure 4.4 Cryo-ET findings on the flagellar motor and cellular features of spirochete *Treponema primitia*. (a) A cutaway surface rendering of the flagellar motor computed by averaging together 20 motors computationally isolated from tomograms. Novel findings included the surprising connectivity between the stator and other components of the motor, marked by yellow arrows; (b) Schematic of the motor organization as revealed through the analysis of the structure shown in (a). The position and geometry of the stator and the existence of the P-collar were missed by previous studies that examined biochemically purified flagellar motors; (c) Surface rendering of a *T. primitia* cell, showing many novel cellular structures, including surface bowls (magenta), the surface hook arcade (yellow) and tip fibrils (orange); the flagella are shown in green and red.

A subsequent cryo-ET study conducted by Jensen and coworkers revealed a plethora of novel structures in the *Treponema* cell, including outer-membrane bowls, polar fibrils, a polar cone and a surface hook arcade that sometimes tracks with the cellular position of flagella (Figure 4.4c and d) [100]. These findings beg new questions as to how, and if, these features relate to flagellar function. More importantly, two distinct periplasmic layers in *T. primitia* were revealed; this observation, when combined with video observations by light microscopy, affirms the rolling cylinder model of motility over the competing gyration model. The importance of this motility mechanism is underscored by the association between motility and pathogenicity in spirochetes such as those that cause Lyme disease and syphilis (for a review, see Ref. [101]).

4.3.4
Proteomic Atlas

Traditional cellular biology studies are frequently limited by carrying out experiments *in vitro* or investigating only fractions of cells. It is an obvious and tremendous

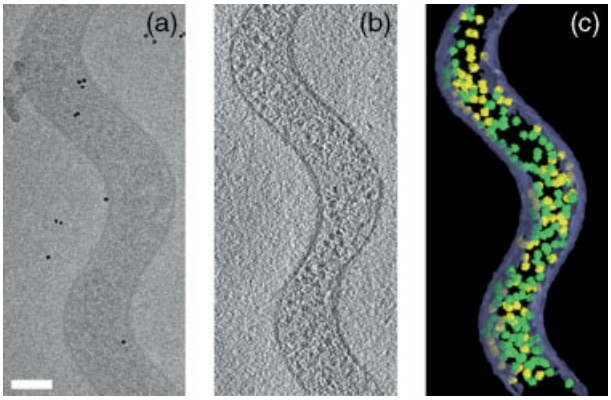


Figure 4.5 A map of ribosome location in the spirochete *Spiroplasma melliferum*. (a) The 0 (projection, an unprocessed cryo-ET image, showing a spirochete cell; (b) A slice of the reconstruction, which shows higher density at positions occupied by ribosomes; (c) A surface rendering of the reconstruction which has been filtered and segmented. High confidence ribosome matches are colored green; intermediate confidence matches are colored yellow.

advantage to integrate the structures and processes of all of the cellular space, enabling investigators to comprehend cells *in toto*. Today, Baumeister and colleagues continue to make strides towards this goal of visualizing a complete cell with all of its major nanomachines. Early proof-of-concept studies have shown that it is possible to identify and differentiate large complexes in the tomograms of synthetic cells [20]. Moreover, recent advances in data processing suggest that even similar assemblies with subtle differences in mass, such as GroEL and GroEL-GroES, can be differentiated [102]. The first application of mapping nanomachines in a cell showed that the total spatial distribution of ribosomes through an entire cell could be directly observed in the spirochete *Spiroplasma melliferum* (Figure 4.5) [103].

The archaebacteria *Thermoplasma acidophilum* is a relatively simple cell with only approximately 1507 open reading frames (ORFs) comprising considerably fewer subcellular assemblies [104]. As such, it is an attractive cryo-ET target for mapping the 3-D position of all major nanomachines – the 'proteomic atlas' – which, ultimately, will reveal unprecedented detail about the 3-D organization of protein–protein networks [105].

4.4 Future Prospects

Today, single-particle cryo-EM has reached the turning point where it is now possible to resolve relatively high-resolution structures of molecular nanomachines under conditions not generally possible with other high-resolution structure determination techniques. Due to the intrinsic nature of the cryo-EM experiment, it can also produce unique and biologically important information, even when a high-resolution

structure is already known. Cryo-EM structures of both the ribosome [106] and GroEL [14, 107, 108] have provided significant insight into structural and functional mechanisms, despite being extensively studied using X-ray crystallography.

One obvious challenge for cryo-EM is the pursuit of higher resolution (i.e. close to or better than 3.0 Å), at which point full, all-atom models could be constructed. On the other hand, cryo-EM is not aimed solely at high resolution. Rather, it offers the ability to resolve domains and/or components that are highly flexible at lower resolutions, as well as samples with multiple conformational states [108]. With further developments in the image-processing routines, both high-resolution structure determination and 'computational purification' of samples [108] will further allow for the exploration of complex molecular nanomachines in greater detail.

As with cryo-EM, improvements in data collection and image processing will allow cryo-ET to achieve more accurate and higher-resolution reconstructions of large nanomachines and cells. However – as alluded to in the proteomic atlas – the real strength of cryo-ET is its power to integrate known atomic structures and cryo-EM reconstructions to provide a complete model of *in vivo* protein function [105, 109]. Such integration will ultimately establish a true spatial and temporal view of functional nanomachines within the cell, which can systematically be investigated in either healthy or diseased states. In addition, there is a trend towards the integration of live cell observations made by light microscopy, followed by cryo-ET observations of the same specimens (e.g. [110]). Such hybrid approaches require not only new instrumentation to make sequential observations practical but also the computational tools to integrate the data. These integrated cellular views promise to enhance our understanding of cell structure and function relationships in normal and diseased states at higher spatial and temporal resolutions.

Acknowledgments

These studies were supported by grants from NIH (P41RR02250, 2PN2EY016525, R01GM079429) and NSF (IIS-0705474, IIS-0705644). We thank our collaborators Drs Irina Serysheva, Steve Ludtke, Yao Cong, Maya Topf, Andrej Sali and Susan Hamilton on the RYR1 project; Wen Jiang, Peter Weigele, Jonathan King, Joanita Jakana and Juan Chang on the epsilon15 phage project; and Jose Lopez on human platelet. We also thank Dr Grant Jensen at California Institute of Technology and Drs Wolfgang Baumeister and Julio Ortiz at the Max Planck Institute for providing the artwork for Figures 4.3 and 4.4, respectively.

References

1 Celniker, S.E. and Rubin, G.M. (2003) The *Drosophila melanogaster* genome. *Annual Review of Genomics and Human Genetics*, **4**, 89.

2 Olivier, M., Aggarwal, A., Allen, J., Almendras, A.A., Bajorek, E.S., Beasley, E.M., Brady, S.D., Bushard, J.M., Bustos, V.I., Chu, A., Chung, T.R., De Witte, A.,

Denys, M.E., Dominguez, R., Fang, N.Y., Foster, B.D., Freudenberg, R.W., Hadley, D., Hamilton, L.R., Jeffrey, T.J., Kelly, L., Lazzeroni, L., Levy, M.R., Lewis, S.C., Liu, X., Lopez, F.J., Louie, B., Marquis, J.P., Martinez, R.A., Matsuura, M.K., Misherghi, N.S., Norton, J.A., Olshen, A., Perkins, S.M., Perou, A.J., Piercy, C., Piercy, M., Qin, F., Reif, T., Sheppard, K., Shokoohi, V., Smick, G.A., Sun, W.L., Stewart, E.A., Fernando, J., Tejeda, Tran, N.M., Trejo, T., Vo, N.T., Yan, S.C., Zierten, D.L., Zhao, S., Sachidanandam, R., Trask, B.J., Myers, R.M. and Cox, D.R. (2001) A high-resolution radiation hybrid map of the human genome draft sequence. *Science*, **291**, 1298.

3 Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., Gocayne, J.D., Amanatides, P., Ballew, R.M., Huson, D.H., Wortman, J.R., Zhang, Q., Kodira, C.D., Zheng, X.H., Chen, L., Skupski, M., Subramanian, G., Thomas, P.D., Zhang, J., Gabor Miklos, G.L., Nelson, C., Broder, S., Clark, A.G., Nadeau, J., McKusick, V.A., Zinder, N., Levine, A.J., Roberts, R.J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A.E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T.J., Higgins, M.E., Ji, R.R., Ke, Z., Ketchum, K.A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G.V., Milshina, N., Moore, H.M., Naik, A.K., Narayan, V.A., Neelam, B., Nusskern, D., Rusch, D.B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304.

4 Yu, H., Peters, J.M., King, R.W., Page, A.M., Hieter, P. and Kirschner, M.W. (1998) Identification of a cullin homology region in a subunit of the anaphase-promoting complex. *Science*, **279**, 1219.

5 Martin, A.C. and Drubin, D.G. (2003) Impact of genome-wide functional analyses on cell biology research. *Current Opinion in Cell Biology*, **15**, 6.

6 Alberts, B. (1998) The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell*, **92**, 291.

7 Alberts, B. and Miake-Lye, R. (1992) Unscrambling the puzzle of biological machines: the importance of the details. *Cell*, **68**, 415.

8 Levchenko, A. (2001) Computational cell biology in the post-genomic era. *Molecular Biology Reports*, **28**, 83.

9 Sali, A. (2003) NIH workshop on structural proteomics of biological complexes. *Structure*, **11**, 1043.

10 Sali, A. and Chiu, W. (2005) Macromolecular assemblies highlighted. *Structure*, **13**, 339.

11 Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) The Protein Data Bank. *Nucleic Acids Research*, **28**, 235.

12 Glaeser, R.M., Downing, K.H., DeRosier, D.L., Chiu, W. and Frank, J. (2007) *Electron crystallography of biological macromolecules*, Oxford University Press, Oxford, UK. New York.

13 Jiang, W., Baker, M.L., Jakana, J., Weigele, P.R., King, J. and Chiu, W. (2008) Backbone structure of the infectious epsilon15 virus capsid revealed by electron cryomicroscopy. *Nature*, **451**, 1130.

14 Ludtke, S.J., Baker, M.L., Chen, D.H., Song, J.L., Chuang, D.T. and Chiu, W. (2008) De Novo backbone trace of GroEL from single particle electron cryomicroscopy. *Structure*, **16**, 441.

15 Zhang, X., Settembre, E., Xu, C., Dormitzer, P.R., Bellamy, R., Harrison, S.C. and Grigorieff, N. (2008) Near-atomic resolution using electron cryomicroscopy and single-particle reconstruction. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 1867.

16 Zhou, Z.H. (2008) Towards atomic resolution structural determination by single-particle cryo-electron microscopy. *Current Opinion in Structural Biology*, **18**, 218.

17 Yu, X., Jin, L. and Zhou, Z.H. (2008) 3.88 A structure of cytoplasmic polyhedrosis virus by cryo-electron microscopy. *Nature*, **453**, 415.

18 Frank, J. (2006) *Three-dimensional electron microscopy of macromolecular assemblies: visualization of biological molecules in their native state*, 2nd edn, Oxford University Press, New York.

19 Dubochet, J., Adrian, M., Chang, J.J., Homo, J.C., Lepault, J., McDowall, A.W. and Schultz, P. (1988) Cryo-electron microscopy of vitrified specimens. *Quarterly Reviews of Biophysics*, **21**, 129.

20 Frangakis, A.S., Bohm, J., Forster, F., Nickell, S., Nicastro, D., Typke, D., Hegerl, R. and Baumeister, W. (2002) Identification of macromolecular complexes in cryoelectron tomograms of phantom cells. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 14153.

21 Grimm, R., Singh, H., Rachel, R., Typke, D., Zillig, W. and Baumeister, W. (1998) Electron tomography of ice-embedded prokaryotic cells. *Biophysical Journal*, **74**, 1031.

22 Nickell, S., Hegerl, R., Baumeister, W. and Rachel, R. (2003) Pyrodictium cannulae enter the periplasmic space but do not enter the cytoplasm, as revealed by cryo-electron tomography. *Journal of Structural Biology*, **141**, 34.

23 Frederik, P.M. and Hubert, D.H. (2005) Cryoelectron microscopy of liposomes. *Methods in Enzymology*, **391**, 431.

24 Berriman, J. and Unwin, N. (1994) Analysis of transient structures by cryo-microscopy combined with rapid mixing of spray droplets. *Ultramicroscopy*, **56**, 241.

25 White, H.D., Walker, M.L. and Trinick, J. (1998) A computer-controlled spraying-freezing apparatus for millisecond time-resolution electron cryomicroscopy. *Journal of Structural Biology*, **121**, 306.

26 Chiu, W., Downing, K.H., Dubochet, J., Glaeser, R.M., Heide, H.G., Knapek, E., Kopf, D.A., Lamvik, M.K., Lepault, J., Robertson, J.D., Zeitler, E. and Zemlin, F. (1986) Cryoprotection in electron microscopy. *Journal of Microscopy*, **141**, 385.

27 Comolli, L.R. and Downing, K.H. (2005) Dose tolerance at helium and nitrogen temperatures for whole cell electron tomography. *Journal of Structural Biology*, **152**, 149.

28 Iancu, C.V., Wright, E.R., Heymann, J.B. and Jensen, G.J. (2006) A comparison of liquid nitrogen and liquid helium as cryogens for electron cryotomography. *Journal of Structural Biology*, **153**, 231.

29 Murata, K., Mitsuoka, K., Hirai, T., Walz, T., Agre, P., Heymann, J.B., Engel, A. and Fujiyoshi, Y. (2000) Structural determinants of water permeation through aquaporin-1. *Nature*, **407**, 599.

30 Henderson, R., Baldwin, J.M., Ceska, T.A., Zemlin, F., Beckmann, E. and Downing, K.H. (1990) An atomic model for the structure of bacteriorhodopsin. *Biochemical Society Transactions*, **18**, 844.

31 Miyazawa, A., Fujiyoshi, Y. and Unwin, N. (2003) Structure and gating mechanism of the acetylcholine receptor pore. *Nature*, **424**, 949.

32 Yonekura, K., Maki-Yonekura, S. and Namba, K. (2003) Complete atomic model of the bacterial flagellar filament by electron cryomicroscopy. *Nature*, **424**, 643.

33 Nogales, E., Wolf, S.G. and Downing, K.H. (1998) Structure of the alpha beta

tubulin dimer by electron crystallography. *Nature*, **391**, 199.

34 Sachse, C., Chen, J.Z., Coureux, P.D., Stroupe, M.E., Fandrich, M. and Grigorieff, N. (2007) High-resolution electron microscopy of helical specimens: a fresh look at tobacco mosaic virus. *Journal of Molecular Biology*, **371**, 812.

35 Iancu, C.V., Tivol, W.F., Schooler, J.B., Dias, D.P., Henderson, G.P., Murphy, G.E., Wright, E.R., Li, Z., Yu, Z., Briegel, A., Gan, L., He, Y. and Jensen, G.J. (2006) Electron cryotomography sample preparation using the Vitrobot. *Nature Protocols*, **1**, 2813.

36 Wright, E.R., Iancu, C.V., Tivol, W.F. and Jensen, G.J. (2006) Observations on the behavior of vitreous ice at approximately 82 and approximately 12K. *Journal of Structural Biology*, **153**, 241.

37 Booth, C.R., Jakana, J. and Chiu, W. (2006) Assessing the capabilities of a 4 k $\times$ 4 k CCD camera for electron cryo-microscopy at 300 kV. *Journal of Structural Biology*, **156**, 556.

38 Booth, C.R., Jiang, W., Baker, M.L., Hong Zhou, Z., Ludtke, S.J. and Chiu, W. (2004) A 9 Å single particle reconstruction from CCD captured images on a 200 kV electron cryomicroscope. *Journal of Structural Biology*, **147**, 116.

39 Chen, D.H., Jakana, J., Liu, X., Schmid, M.F. and Chiu, W. (2008) Achievable resolution from images of biological specimens acquired from a 4 k $\times$ 4 k CCD camera in a 300-kV electron cryomicroscope. *Journal of Structural Biology*, **163**, 45.

40 Carragher, B., Kisseberth, N., Kriegman, D., Milligan, R.A., Potter, C.S., Pulokas, J. and Reilein, A. (2000) Leginon: an automated system for acquisition of images from vitreous ice specimens. *Journal of Structural Biology*, **132**, 33.

41 Lei, J. and Frank, J. (2005) Automated acquisition of cryo-electron micrographs for single particle reconstruction on an FEI Tecnai electron microscope. *Journal of Structural Biology*, **150**, 69.

42 Zhang, P., Beatty, A., Milne, J.L.S. and Subramaniam, S. (2001) Automated data collection with a Tecnai 12 electron microscope: applications for molecular imaging by cryo-microscopy. *Journal of Structural Biology*, **135**, 251.

43 Marsh, M.P., Chang, J.T., Booth, C.R., Liang, N.L., Schmid, M.F. and Chiu, W. (2007) Modular software platform for low-dose electron microscopy and tomography. *Journal of Microscopy*, **228**, 384.

44 DeRosier, D.L. and Klug, A. (1968) Reconstruction of three-dimensional structures from electron micrographs. *Nature*, **217**, 130.

45 Crowther, R.A., DeRosier, D.J. and Klug, A. (1970) The reconstruction of a three-dimensional structure from projections and its application to electron microscopy. *Proceedings of the Royal Society of London*, **317**, 319.

46 Henderson, R. (1995) The potential and limitations of neutrons, electrons and X-rays for atomic resolution microscopy of unstained biological molecules. *Quarterly Reviews of Biophysics*, **28**, 171.

47 Liu, X., Jiang, W., Jakana, J. and Chiu, W. (2007) Averaging tens to hundreds of icosahedral particle images to resolve protein secondary structure elements using a Multi-Path Simulated Annealing optimization algorithm. *Journal of Structural Biology*, **160**, 11.

48 Jiang, W. and Chiu, W. (2007) Cryoelectron microscopy of icosahedral virus particles. *Methods in Molecular Biology (Clifton, NJ)*, **369**, 345.

49 Serysheva, I.I., Chiu, W. and Ludtke, S.J. (2007) Single-particle electron cryomicroscopy of the ion channels in the excitation-contraction coupling junction. *Methods in Cell Biology*, **79**, 407.

50 Ludtke, S.J., Baldwin, P.R. and Chiu, W. (1999) EMAN: Semi-automated software for high resolution single particle reconstructions. *Journal of Structural Biology*, **128**, 82.

51 Frank, J., Radermacher, M., Penczek, P., Zhu, J., Li, Y., Ladjadj, M. and Leith, A. (1996) SPIDER and WEB: processing and visualization of images in 3D electron microscopy and related fields. *Journal of Structural Biology*, **116**, 190.

52 van Heel, M., Harauz, G., Orlova, E.V., Schmidt, R. and Schatz, M. (1996) A new generation of the IMAGIC image processing system. *Journal of Structural Biology*, **116**, 17.

53 Grigorieff, N. (2007) FREALIGN: high-resolution refinement of single particle structures. *Journal of Structural Biology*, **157**, 117.

54 Saxton, W.O. and Baumeister, W. (1982) The correlation averaging of a regularly arranged bacterial cell envelope protein. *Journal of Microscopy*, **127**, 127.

55 Lawrence, M.C. (1992) *Electron Tomography: Three-dimensional Imaging with the Transmission Electron Microscope*, 1st edn (ed. J. Frank), Springer, p. 197.

56 Luther, P.K., Lawrence, M.C. and Crowther, R.A. (1988) A method for monitoring the collapse of plastic sections as a function of electron dose. *Ultramicroscopy*, **24**, 7.

57 Mastronarde, D.N. (2006) *Electron tomography: Methods for three-dimensional visualization of structures in the cell*, 2nd edn (ed. J. Frank), Springer, p. 187.

58 Radermacher, M. (2006) *Electron tomography: Methods for three-dimensional visualization of structures in the cell*, 2nd edn (ed. J. Frank), Springer, p. 245.

59 Sandberg, K., Mastronarde, D.N. and Beylkin, G. (2003) A fast reconstruction algorithm for electron microscope tomography. *Journal of Structural Biology*, **144**, 61.

60 Kremer, J.R., Mastronarde, D.N. and McIntosh, J.R. (1996) Computer visualization of three-dimensional image data using IMOD. *Journal of Structural Biology*, **116**, 71.

61 Nickell, S., Forster, F., Linaroudis, A., Net, W.D., Beck, F., Hegerl, R., Baumeister, W. and Plitzko, J.M. (2005) TOM software toolbox: acquisition and analysis for electron tomography. *Journal of Structural Biology*, **149**, 227.

62 Winkler, H. (2007) 3D reconstruction and processing of volumetric data in cryo-electron tomography. *Journal of Structural Biology*, **157**, 126.

63 Zheng, S.Q., Keszthelyi, B., Branlund, E., Lyle, J.M., Braunfeld, M.B., Sedat, J.W. and Agard, D.A. (2007) UCSF tomography: an integrated software suite for real-time electron microscopic tomographic data collection, alignment, and reconstruction. *Journal of Structural Biology*, **157**, 138.

64 Cardone, G., Grunewald, K. and Steven, A.C. (2005) A resolution criterion for electron tomography based on cross-validation. *Journal of Structural Biology*, **151**, 117.

65 Rossmann, M.G., Morais, M.C., Leiman, P.G. and Zhang, W. (2005) Combining X-ray crystallography and electron microscopy. *Structure*, **13**, 355.

66 Chiu, W., Baker, M.L., Jiang, W., Dougherty, M. and Schmid, M.F. (2005) Electron cryomicroscopy of biological machines at subnanometer resolution. *Structure*, **13**, 363.

67 Baker, M.L., Jiang, W., Wedemeyer, W.J., Rixon, F.J., Baker, D. and Chiu, W. (2006) Ab initio modeling of the herpesvirus VP26 core domain assessed by CryoEM density. *PLoS Computational Biology*, **2**, e146.

68 Topf, M., Baker, M.L., John, B., Chiu, W. and Sali, A. (2005) Structural characterization of components of protein assemblies by comparative modeling and electron cryo-microscopy. *Journal of Structural Biology*, **149**, 191.

69 Topf, M., Baker, M.L., Marti-Renom, M.A., Chiu, W. and Sali, A. (2006) Refinement of protein structures by Iterative comparative modeling and cryoEM density fitting. *Journal of Molecular Biology*, **357**, 1655.

70 Baker, M.L., Jiang, W., Bowman, B.R., Zhou, Z.H., Quiocho, F.A., Rixon, F.J.

and Chiu, W. (2003) Architecture of the herpes simplex virus major capsid protein derived from structural bioinformatics. *Journal of Molecular Biology*, **331**, 447.

71 Baker, M.L., Ju, T. and Chiu, W. (2007) Identification of secondary structure elements in intermediate-resolution density maps. *Structure*, **15**, 7.

72 Jiang, W., Baker, M.L., Ludtke, S.J. and Chiu, W. (2001) Bridging the information gap: computational tools for intermediate resolution structure interpretation. *Journal of Molecular Biology*, **308**, 1033.

73 Zhou, Z.H., Baker, M.L., Jiang, W., Dougherty, M., Jakana, J., Dong, G., Lu, G. and Chiu, W. (2001) Electron cryomicroscopy and bioinformatics suggest protein fold models for rice dwarf virus. *Nature Structural Biology*, **8**, 868.

74 Kong, Y. and Ma, J. (2003) A structural-informatics approach for mining beta-sheets: locating sheets in intermediate-resolution density maps. *Journal of Molecular Biology*, **332**, 399.

75 Kong, Y., Zhang, X., Baker, T.S. and Ma, J. (2004) A structural-informatics approach for tracing beta-sheets: building pseudo-C (alpha) traces for beta-strands in intermediate-resolution density maps. *Journal of Molecular Biology*, **339**, 117.

76 Baker, M.L., Jiang, W., Rixon, F.J. and Chiu, W. (2005) Common ancestry of herpesviruses and tailed DNA bacteriophages. *Journal of Virology*, **79**, 14967.

77 Sandberg, K. (2007) *Cellular electron microscopy* (ed. J.R. McIntosh), Elsevier, New York, p. 770.

78 Forster, F. and Hegerl, R. (2007) *Cellular electron microscopy* (ed. J.R. McIntosh), Elsevier, New York, p. 742.

79 Ludtke, S.J., Nason, L., Tu, H., Peng, L. and Chiu, W. (2003) Object oriented database and electronic notebook for transmission electron microscopy. *Microscopy and Microanalysis*, **9**, 556.

80 Marabini, R., Vaquerizo, C., Fernandez, J.J., Carazo, J.M., Engel, A. and Frank, J. (1996) Proposal for a new distributed database of macromolecular and subcellular structures from different areas of microscopy. *Journal of Structural Biology*, **116**, 161.

81 Fellmann, D., Pulokas, J., Milligan, R.A., Carragher, B. and Potter, C.S. (2002) A relational database for cryoEM: experience at one year and 50 000 images. *Journal of Structural Biology*, **137**, 273.

82 Fuller, S.D. (2003) Depositing electron microscopy maps. *Structure (Camb.)*, **11**, 11.

83 Ludtke, S.J., Serysheva, I.I., Hamilton, S.L. and Chiu, W. (2005) The pore structure of the closed RyR1 channel. *Structure (Camb.)*, **13**, 1203.

84 Jiang, Y., Lee, A., Chen, J., Cadene, M., Chait, B.T. and MacKinnon, R. (2002) Crystal structure and mechanism of a calcium-gated potassium channel. *Nature*, **417**, 515.

85 Kuo, A., Gulbis, J.M., Antcliff, J.F., Rahman, T., Lowe, E.D., Zimmer, J., Cuthbertson, J., Ashcroft, F.M., Ezaki, T. and Doyle, D.A. (2003) Crystal structure of the potassium channel KirBac1.1 in the closed state. *Science*, **300**, 1922.

86 Nishida, M., Cadene, M., Chait, B.T. and MacKinnon, R. (2007) Crystal structure of a Kir3.1-prokaryotic Kir channel chimera. *The EMBO Journal*, **26**, 4005.

87 Zagotta, W.N., Olivier, N.B., Black, K.D., Young, E.C., Olson, R. and Gouaux, E. (2003) Structural basis for modulation and agonist specificity of HCN pacemaker channels. *Nature*, **425**, 200.

88 Serysheva, I.I., Ludtke, S.J., Baker, M.L., Cong, Y., Topf, M., Eramian, D., Sali, A., Hamilton, S.L. and Chiu, W. (2008) Subnanometer-resolution electron cryomicroscopy-based domain models for the cytoplasmic region of skeletal muscle RyR channel. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 9610.

89 Wolf, M., Eberhart, A., Glossmann, H., Striessnig, J. and Grigorieff, N. (2003) Visualization of the domain structure of an L-type Ca(2 +) channel using electron

cryo-microscopy. *Journal of Molecular Biology*, **332**, 171.

90 Serysheva, I.I., Ludtke, S.J., Baker, M.R., Chiu, W. and Hamilton, S.L. (2002) Structure of the voltage-gated L-type Ca^{2+} channel by electron cryomicroscopy. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 10370.

91 Block, B.A. and Franzini-Armstrong, C. (1988) The structure of the membrane systems in a novel muscle cell modified for heat production. *The Journal of Cell Biology*, **107**, 1099.

92 Jiang, W., Chang, J., Jakana, J., Weigele, P., King, J. and Chiu, W. (2006) Structure of epsilon15 bacteriophage reveals genome organization and DNA packaging/injection apparatus. *Nature*, **439**, 612.

93 Wikoff, W.R., Liljas, L., Duda, R.L., Tsuruta, H., Hendrix, R.W. and Johnson, J.E. (2000) Topologically linked protein rings in the bacteriophage HK97 capsid. *Science*, **289**, 2129.

94 Jemth, P. and Gianni, S. (2007) PDZ domains: folding and binding. *Biochemistry*, **46**, 8701.

95 Berg, H.C. (2003) The rotary motor of bacterial flagella. *Annual Review of Biochemistry*, **72**, 19.

96 Thomas, D.R., Francis, N.R., Xu, C. and DeRosier, D.J. (2006) The three-dimensional structure of the flagellar rotor from a clockwise-locked mutant of *Salmonella enterica* serovar *Typhimurium*. *Journal of Bacteriology*, **188**, 7039.

97 Suzuki, H., Yonekura, K. and Namba, K. (2004) Structure of the rotor of the bacterial flagellar motor revealed by electron cryomicroscopy and single-particle image analysis. *Journal of Molecular Biology*, **337**, 105.

98 Francis, N.R., Sosinsky, G.E., Thomas, D. and DeRosier, D.J. (1994) Isolation, characterization and structure of bacterial flagellar motors containing the switch complex. *Journal of Molecular Biology*, **235**, 1261.

99 Murphy, G.E., Leadbetter, J.R. and Jensen, G.J. (2006) In situ structure of the complete *Treponema primitia* flagellar motor. *Nature*, **442**, 1062.

100 Murphy, G.E., Matson, E.G., Leadbetter, J.R., Berg, H.C. and Jensen, G.J. (2008) Novel ultrastructures of *Treponema primitia* and their implications for motility. *Molecular Microbiology*, **67**, 1184.

101 Lux, R., Moter, A. and Shi, W. (2000) Chemotaxis in pathogenic spirochetes: directed movement toward targeting tissues? *Journal of Molecular Microbiology and Biotechnology*, **2**, 355.

102 Forster, F., Pruggnaller, S., Seybert, A. and Frangakis, A.S. (2008) Classification of cryo-electron sub-tomograms using constrained correlation. *Journal of Structural Biology*, **161**, 276.

103 Ortiz, J.O., Forster, F., Kurner, J., Linaroudis, A.A. and Baumeister, W. (2006) Mapping 70S ribosomes in intact cells by cryoelectron tomography and pattern recognition. *Journal of Structural Biology*, **156**, 334.

104 Sun, N., Beck, F., Knispel, R.W., Siedler, F., Scheffer, B., Nickell, S., Baumeister, W. and Nagy, I. (2007) Proteomics analysis of *Thermoplasma acidophilum* with a focus on protein complexes. *Molecular and Cellular Proteomics*, **6**, 492.

105 Robinson, C.V., Sali, A. and Baumeister, W. (2007) The molecular sociology of the cell. *Nature*, **450**, 973.

106 Valle, M., Gillet, R., Kaur, S., Henne, A., Ramakrishnan, V. and Frank, J. (2003) Visualizing tmRNA entry into a stalled ribosome. *Science*, **300**, 127.

107 Sewell, B.T., Best, R.B., Chen, S., Roseman, A.M., Farr, G.W., Horwich, A.L. and Saibil, H.R. (2004) A mutant chaperonin with rearranged inter-ring electrostatic contacts and temperature-sensitive dissociation. *Nature Structural & Molecular Biology*, **11**, 1128.

108 Chen, D.H., Song, J.L., Chuang, D.T., Chiu, W. and Ludtke, S.J. (2006) An expanded conformation of single-ring

GroEL-GroES complex encapsulates an 86 kDa substrate. *Structure*, **14**, 1711.

109 Nickell, S., Kofler, C., Leis, A.P. and Baumeister, W. (2006) A visual approach to proteomics. *Nature Reviews. Molecular Cell Biology*, **7**, 225.

110 Sartori, A., Gatz, R., Beck, F., Rigort, A., Baumeister, W. and Plitzko, J.M. (2007) Correlative microscopy: bridging the gap between fluorescence light microscopy and cryo-electron tomography. *Journal of Structural Biology*, **160**, 135.

5
Pushing Optical Microscopy to the Limit: From Single-Molecule Fluorescence Microscopy to Label-Free Detection and Tracking of Biological Nano-Objects

Philipp Kukura, Alois Renn, and Vahid Sandoghdar

5.1
Introduction

The fundamental goal of microscopic imaging is to visualize and identify small objects and to observe their motion. Many techniques based on a wide variety of approaches, including X-ray scattering, scanning probe microscopy, electron diffraction and various optical implementations, have been developed over the past century to improve upon the fundamental limitations of traditional optical microscopy. While each approach has its specific advantages, none by itself can provide a solution to all demands regarding spatial and temporal resolution, sensitivity and *in vivo* imaging capability, as are often desired in biologically motivated studies.

The ultimate resolution would allow one to detect, localize and visualize single molecules, or even atoms. Obtaining structural snapshots at molecular and atomic resolution has become routinely available through X-ray crystallographic techniques. Especially from a biological perspective, such detailed images of molecular and supramolecular structures regularly provide unique insights into their function [1–4]. Comparable resolution is attainable with scanning probe techniques such as scanning tunneling microscopy (STM) or atomic force microscopy (AFM) [5, 6]. While the former is generally limited to nonbiological samples, many biological AFM studies have been reported during the past decades [7, 8]. These experiments have been very informative, both due to their resolution and their ability to apply minute forces on single molecules or nano-objects. Various implementations of electron microscopy have also greatly contributed to biological imaging, due to its ability to provide spatial resolution in the nanometer region, thereby bridging the gap between the ultra high-resolution techniques mentioned above and standard optical imaging approaches [9].

Despite the many successes of these techniques, they all lack an ability to perform real-time, *in vivo* imaging. X-ray scattering experiments require high-quality crystals, which makes studies inside biological media intrinsically impossible. Scanning probe techniques are, by definition, limited to the study of surfaces and are relatively

Nanotechnology, Volume 5: Nanomedicine. Edited by Viola Vogel

ISBN: 978-3-527-31736-3

slow, with acquisition rates rarely exceeding a few Hertz. Electron microscopy requires a vacuum and sometimes metal-coating or cryogenic conditions for high-resolution images, thus making it difficult to perform studies under biologically relevant conditions.

As a consequence, the method of choice for real-time, *in vivo* biological imaging has remained optical microscopy, despite its comparatively low resolution. The optical microscope was invented about 400 years ago, and improvements in resolution to about 200 nm had already taken place by the end of the nineteenth century, through advances in lens design. However, this is a factor of 100 larger than the size of single molecules or proteins that define the ultimately desired resolution, and represents the diffraction limit established by Abbe during the 1880s. Throughout most of the twentieth century, this fundamental limit prevented the optical microscope from opening our eyes to the molecular structure of matter. Nevertheless, numerous recent advances have been made to improve the resolution and, in particular, the contrast of images. In general, these contrast techniques can be divided into two categories, namely linear and nonlinear.

5.1.1 Linear Contrast Mechanisms

The central challenge in visualizing small objects is to distinguish their signals from background scattering, or from reflections that may be caused by other objects or the optics inside the microscope. One possible solution to this problem is provided by dark-field microscopy, which was first reported by Lister in 1830. Here, the design is optimized to reduce all unwanted light to a minimum by preventing the illumination light from reaching the detector [10, 11]. To achieve this, the illumination light is shaped such that it can be removed before detection through the use of an aperture. In general, there are two approaches to achieve dark-field illumination; these are illustrated schematically in Figure 5.1a and b. In Figure 5.1a, the illumination light is modified in such a way as to provide an intense ring of light, for example through the use of a coin-like object blocking all but the outer ring of the incident light. The illumination light can be then removed by an aperture after the sample. In this way, only the light that is scattered by objects of interest, and whose path is thereby deviated, can reach the detector. The other approach, shown in Figure 5.1b, is referred to as total internal reflection microscopy (TIRM) [12]. Here, the illumination is incident on the sample at a steep angle, and is fully reflected at the interface between the glass coverslip and the sample (e.g. water). There, an 'evanescent' region is created, where the light decays exponentially as it enters into the sample. When a particle is placed into this region it scatters energy out of the evanescent part of the beam into the objective. Thus, light will only reach the detector from scatterers that are located within 100 nm of the sample–substrate interface. This technique has been mostly used to detect very weak signals such as those from metallic nanoparticles, or to obtain surface-specific images [13, 14].

Another approach to minimize unwanted light is provided by confocal microscopy, which first appeared during the early 1960s (Figure 5.1c) [15, 16]. In contrast to the

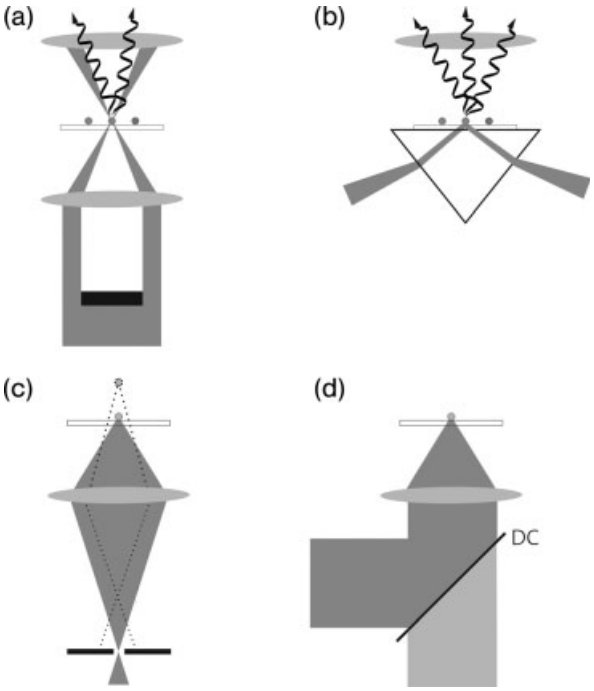


Figure 5.1 Illustration of the major linear contrast mechanisms in optical microscopy. (a) Dark-field illumination; (b) Total internal reflection microscopy (TIRM); (c) Confocal microscopy; (d) Fluorescence microscopy.

previous techniques, where the light from a relatively large sample area is collected ($\sim 20 \times 20\,\mu m$), only light from the focal region of the objective is allowed to reach the detector, thereby considerably reducing the background light. This is achieved through the use of a pinhole that is placed in the confocal plane of the objective. The size of the pinhole is usually chosen so as to match the size of the image of the focal spot. Collecting light in this fashion leads to optical sectioning – that is, the ability to provide three-dimensional (3-D) information by moving the focus in the z-direction through the sample. The major disadvantage of this approach is the need to scan the sample with respect to the illumination, which leads to relatively slow acquisition times.

An alternative for improving the contrast is to exploit the phase of the illumination light. Differential interference contrast (DIC) microscopy, which was introduced during the mid-1950s, takes advantage of slight differences in the refractive index of the specimen to generate additional information about the composition of the sample [17, 18]. For instance, the refractive index of water differs from that of lipids, so areas with a high water content will generate a different signal than those consisting mostly of organic material. The approach splits the illumination light

into two slightly displaced beams (<1 μm) at the focus that are recombined before detection. If the two beams travel through material with different indices of refraction, the phase of one relative to the other will change, and this will lead to a small degree of destructive interference after recombination of the two beams. Such areas will thus appear dark, while areas where the sample is homogeneous will appear bright.

So far, we have been only concerned with improving the contrast using scattered light. A different but powerful method is to use *fluorescence* as a contrast mechanism. The first such reports emerged during the early twentieth century, when ultra-violet light was used for the first time in microscopes. However, the breakthrough occurred during the 1950s, when the application of fluorescence labeling for the detection of antigens was demonstrated [19]. Rather than detecting the scattering of incident light, the illumination light is used to excite molecules, causing them to fluoresce. The advantage of this approach is that the excitation light can be reduced by many orders of magnitude by the use of appropriate filters, as the fluorescence is usually red-shifted in energy (Figure 5.1d). One can then observe the species of interest virtually against zero background because the only photons that can reach the detector must be due to fluorescence emission. Such contrast is virtually unachievable with scattering techniques, even when dark-field illumination is employed, because the background scattering cannot be extinguished to such a high degree. The major sources of contrast are biological autofluorescence, specific labeling of the objects of interest with fluorescent dye molecules, or use of the cellular expression system itself to produce fluorescent proteins [20]. Fluorescence can also be used to introduce specificity by spectrally 'coding' the sample. Examples are fluorescence recovery after photobleaching (FRAP) for studying fluidity [21], fluorescence lifetime imaging [22] or simply simultaneous labeling with different fluorophores.

The spectral resolution of fluorescence is, however, rather poor because the emission is usually very broad in energy. Techniques based on vibrational spectroscopy on the other hand, where the resonances are orders of magnitude sharper, provide a much more unique molecular fingerprint, but at the expense of much-reduced signal intensities. In particular, Raman and infrared microscopy yield information about the composition of the sample, without the need for any label [23]. The experimental set-up is very similar to that used for fluorescence microscopy, but is focused on detecting vibrational resonances rather than fluorescence emission. As a matter of fact, Raman experiments can only be successful when the background fluorescence is reduced to a minimum, because the cross-sections for Raman scattering are orders of magnitudes smaller than that of fluorescence. On the upside, every species has a unique Raman spectrum and can thus be identified without the need for any label. Because these experiments must be performed in a confocal arrangement to achieve a sufficiently high photon flux, they provide highly specific and spatially resolved information, even inside cells. The downside is that the intrinsically low Raman cross-sections require large illumination powers, and this might be problematic for live cell imaging, for reasons of phototoxicity.

5.1.2
Nonlinear Contrast Mechanisms

The recent availability of high-power laser sources producing ultrashort pulses on the order of a hundred femtoseconds (10^{-13} s) or less at high pulse powers (>mJ) has opened up completely new areas in microscopy. When such short and intense pulses are focused to a diffraction-limited spot, peak powers of terawatts per cm^2 and above can be achieved. At such high peak powers, nonlinear effects that involve the simultaneous interaction of multiple photons with the sample become observable. The main microscopy-related application that has emerged from this technological jump is that of *two-photon imaging* [24]. Here, rather than exciting a fluorophore with a single resonant photon, for example at 400 nm, two off-resonant photons at 800 nm are used to produce the same excitation. Although nonresonant two-photon cross-sections are negligible compared to their one-photon counterparts (10^{-50} versus 10^{-16} cm^2), the high peak powers coupled with high pulse repetition rates on the order of 100 MHz can compensate for these dramatically lowered cross-sections. The major advantage of this technique is that optical sectioning comes for free, because the high peak intensities are only produced at the focus of the illuminating beam, making confocal pinholes superfluous. Despite the fact that biological tissue is generally transparent in the near-infrared region, sample heating and the rapid destruction of two-photon fluorophores cannot be avoided due to the necessarily high peak intensities used.

Another prominent example of nonlinear microscopy is based on coherent anti-Stokes Raman scattering (CARS) [25], which is the nonlinear equivalent of Raman microscopy. Here, three incident photons are required to produce a single Raman shifted photon. Additional techniques based on second and third harmonic generation microscopies have also appeared [26, 27]. Finally, a very interesting recent development has been discussed in the context of RESOLFT (reversible saturable optically linear fluorescence transitions). Here, a nonlinear process such as stimulated emission is used to deplete the fluorescence to a subdiffraction-limited spot. As a result, the actual volume from which fluorescent photons are emitted is considerably reduced beyond $\lambda/50$ [28].

In this chapter, we will focus our attention on pushing both the sensitivity and resolution limits of state-of-the-art *linear* microscopic techniques. In particular, we will discuss the capabilities and limitations of single-molecule detection in the light of biological applications. After covering the fundamental aspects of resolution, we will outline recent advances in the detection and tracking of nonfluorescent nano-objects. Scattering based labels show much promise in eliminating many of the limitations of fluorescence microscopy. Yet, by going a step further, we will show how these techniques can be used to detect and follow the motion of *unlabeled* biological nanoparticles.

5.2
Single-Molecule Fluorescence Detection: Techniques and Applications

The fundamental question that arises from the previous discussion of current optical imaging methods is: Why is it so difficult, first to 'see' single molecules, and second to

achieve molecular resolution with optical microscopes? The former is particularly baffling because the fluorescence of single ions trapped in vacuum *can* be observed with the naked eye, as was shown 20 years ago [29].

5.2.1 Single Molecules: Light Sources with Ticks

To understand the intricacies of detecting single molecules in biological environments, it is useful to ask why it is so easy to observe single ions trapped in a vacuum. The answer is simply – the *absence of any background*. In a vacuum, the emitter is alone, with no other objects or molecules nearby that can either scatter the excitation light or fluoresce upon excitation. In this scenario, single-molecule detection simply becomes an issue of having a good enough detector (curiously, the human eye is one of the best light detectors available, being able to detect single photons with almost 70% efficiency [30]). However, the number of photons that any single molecule can emit via *fluorescence* is strictly limited by its intrinsic photophysics, and cannot be increased at will simply by raising the illumination power.

To understand this concept, it is useful to consider the dynamics that follow the absorption of a single photon by a single molecule. Population of the first excited electronic state is followed by an excited state decay which can take place via two major pathways: nonradiative and radiative decay. The former refers in this simple case to a transition from the excited to the ground state, without the emission of a photon. The excess energy is usually deposited in vibrational degrees of freedom, either of the molecule itself or of the surroundings. In the latter case, the energy is lost through the emission of a photon which is typically lower in energy (red-shifted) than the excitation photon due to the Stokes shift. To generate as many detectable photons per unit time as possible for a given excitation power, one requires: (i) a large absorption cross-section; and (ii) a high fluorescence quantum yield – that is, an efficient conversion of absorbed into emitted photons.

The former requirement brings with it a radiative lifetime on the order of nanoseconds, which can be related to the fundamental considerations of absorption and emission. This limits the total number of emitted photons, irrespective of the total incoming photon flux, because a single absorption–emission cycle takes about 10 ns. Thus, even in ideal circumstances no single molecule can emit more than 10^8 photons per second. The restricted collection properties of objectives, imperfect transmission and reflectivity of optics and limited quantum efficiencies of detectors result in typical effective collection efficiencies of $<10\%$. The corresponding count rates on the order of a few million counts per second can indeed be observed in single-molecule experiments at cryogenic temperatures [31]. Under ambient conditions, which are of relevance for biological investigations, photobleaching puts a limit on the applied excitation intensities. The cause of this bleaching is often the generation and further excitation of triplet states that are accessed through intersystem crossing from the first excited singlet state. Despite the low quantum efficiency of the process ($<1\%$), excitation powers must be chosen at the kW cm^{-2} level in order to avoid rapid

photobleaching. At these incident light levels, the observed count rates are below 10^5 photons per second [32].

Photobleaching is also the reason why single molecules usually emit a total of 10^5–10^7 photons before turning dark. The most likely cause of this 'sudden death' is triplet–triplet annihilation with molecular oxygen, which is a particular problem in biological environments. The highly reactive singlet oxygen that is generated attacks the dye and oxidizes it, greatly altering its electronic properties and thereby rendering it dark to the excitation photons [33, 34]. To make matters worse, triplet state formation is thought to be the main cause of phototoxicity [35, 36]. This situation can be improved somewhat by deoxygenating the system, or by using oxygen scavengers for *in vitro* experiments in solution [37].

5.2.2
The Signal-to-Noise Ratio Challenge

In addition to the saturation properties of single molecules discussed above, another difficulty arises from the fact that single molecules cannot be excited very efficiently. Even large single-molecule absorption cross-sections are only on the order of $10^{-16}\,\mathrm{cm}^2$, compared to focal areas that are no smaller than $\sim 10^{-9}\,\mathrm{cm}^2$. Therefore, only one in 10^7 photons that passes through the focus will cause electronic excitation of the molecule. What makes the situation even more problematic, is the fact that a typical focal volume contains on the order of 10^9 molecules. So, even if only one in 10^3 molecules emits a single fluorescence photon per second, the total emission background already matches the maximum fluorescence from the single molecule of about 10^6 photons, even at low temperatures.

As a consequence, initial attempts to detect single molecules used absorption [38], although these were swiftly followed by fluorescence detection [39]. The early studies were performed at cryogenic temperatures, where absorption cross-sections become large so that the saturation regime can be reached at much lower incident powers. At room temperature, however, the task appeared hopeless in the light of the numbers above. One possibility to improve this situation was to develop a technique that is: (i) only surface-sensitive; and (ii) somehow produces a much smaller excitation area. Scanning near-field optical microscopy (SNOM) provides exactly these properties [40, 41]. Here, the light is not focused as in a standard optical microscope, but rather is coupled into a metallized and sharpened tip of an optical fiber equipped with a subwavelength exit hole ($<50\,\mathrm{nm}$). The tip is then brought within tens of nanometers of the surface to be studied and scanned laterally. In this way, only molecules that are on the surface are excited by the evanescent field at the tip's aperture, and only in an area that is comparable to the aperture. Therefore, the number of molecules that can contribute to the background signal becomes considerably smaller compared to the confocal arrangement, making the detection of single molecules much more probable. Indeed, the first room-temperature observation of single molecules was achieved using SNOM [42].

The rather difficult experimental set-up necessary for performing SNOM, along with the limitations to study surfaces, motivated the development of far-field

single-molecule methods for biological studies. It was quickly realized that far-field methods are capable of generating much larger single-molecule signals than SNOM, with much-reduced experimental demands. Therefore, far-field detection has become the method of choice for detecting single molecules in biological environments [43, 44]. This advance was facilitated by the development of low autofluorescence microscope objectives and immersion oils, as well as improved excitation light rejection through the use of dielectric filters and highly efficient single photon detectors such as avalanche photodiodes. Today, single-molecule detection has become an almost standard technique in biology, chemistry and physics [45]. Single-molecule techniques have been used to directly observe the motion and function of single biological nano-objects such as enzymes, viruses or motor proteins in real time [37, 46, 47].

5.2.3
High-Precision Localization and Tracking of Single Emitters

In the previous section, we discussed the difficulties and current solutions to detecting single molecules. However, we are still faced with the problem, that single molecules are much smaller (~1 nm) than the best possible resolution of an optical microscope (~200 nm), which is linked to the wave nature of light [48]. The crucial point is that the image of a point-like emitter is itself not infinitely small but rather appears as an Airy diffraction image (Figure 5.2a), with the ripples originating from the diffraction at the edges of a circular objective, for example. Because these patterns, which are also known as the point spread functions (PSFs), are caused by the lens, the smaller the aperture of the lens the wider the PSF and the lower the resolution, and vice versa. Here, it is useful to define the term numerical aperture, $NA = n\sin(\varphi)$, where φ is defined as the half collection angle of the objective. Therefore, the larger the NA, the higher the resolution of the microscope. The distance from the central maximum to the first minimum is given by $d_{min} = 0.61\lambda/NA$ for a circular aperture, where λ is the wavelength of light and is a common measure of resolution also known as Rayleigh's limit (Figure 5.2b).

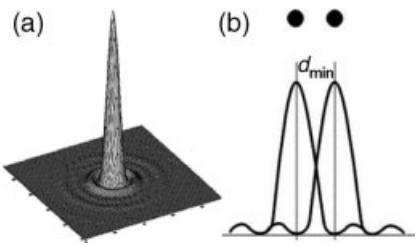


Figure 5.2 Point spread functions and their importance in determining resolution. (a) Surface plot of a typical Airy diffraction pattern; (b) Schematic representation of Rayleigh's criterion. The two graphs represent slices through the 2-D Airy function shown in (a).

Despite the fact that a single molecule smaller than 1 nm yields an image with a diameter of several hundred nanometers, it is possible to determine its location to within a few nanometers. This can be achieved by fitting the data to the theoretical PSF. Thus, the precision of this fit is only limited by the signal-to-noise ratio (SNR) of the acquired Gaussian profile, while the fit tolerance provides the uncertainty in x and y of the emitter's center [49]. To illustrate this, we have acquired confocal images of single rhodamine-labeled GM1 receptors adhered onto an acid-cleaned coverslip (Figure 5.3a). The observed single-step bleaching in Figure 5.3b for the spot highlighted in Figure 5.3a confirms that the image stems from a single molecule. The emission pattern (Figure 5.3c) and the corresponding Gaussian fit (Figure 5.3d) to the highlighted molecule result in a lateral localization accuracy of 10 nm. The high accuracy is due to the fact that all the information in two dimensions can be used for the fit. To illustrate this, three slices and the corresponding fits along the x axis of the spot are shown in Figure 5.3e–g. The center position fluctuates by >20 nm for these three fits due to the limited SNR. A two-dimensional fit, however, provides much higher accuracy, because one effectively fits all slices in every possible direction simultaneously. This approach has been employed in several recent investigations, including the study of lipid diffusion inside supported membrane bilayers [50], and of the mechanism of the molecular motor kinesin stepping along microtubules [51]. While the former study showed a maximum localization accuracy of $\sim$40 nm, the latter state-of-the-art measurements succeeded in realizing molecular resolution ($\sim$1.5 nm) with sub-second time resolution.

As can be seen from the previous discussion, the localization accuracy is critically dependent on the SNR with which the PSF can be measured. The limitation arises from the finite number of detectable photons per molecule. The longer the integration time, the higher the accuracy but also the lower the time resolution. As a rule of thumb, a SNR of 10 is required for a localization accuracy of $\sim$10 nm, which translates into roughly a time resolution on the order of several to tens of milliseconds [49, 52]. This makes tracking beyond video rates difficult if the object is to remain visible against the background, especially for *in vivo* imaging [47]. In addition, the total tracking time is limited to a few seconds because of photobleaching. These issues can only be addressed by using labels with no limitations on the number of emitted photons and on photostability. Nevertheless, single molecules have been used successfully to track individual biological nano-objects in real time. An excellent example is given in Figure 5.4, where single adenoviruses were labeled with single dye molecules and then observed before, during and after cell entry [47].

Inorganic quantum dots have become popular as labels in fluorescence microscopy because of their brightness and extreme photostability [53]. Their inherent toxicity is commonly deactivated through the use of protecting layers, and they have been used successfully in intracellular and *in vivo* studies [54]. A major disadvantage of these labels is that, so far, their emission switches off intermittently at unpredictable times and for unknown durations, a process known as *photoblinking*. In addition, once passivated, they can become as large as 15–20 nm.

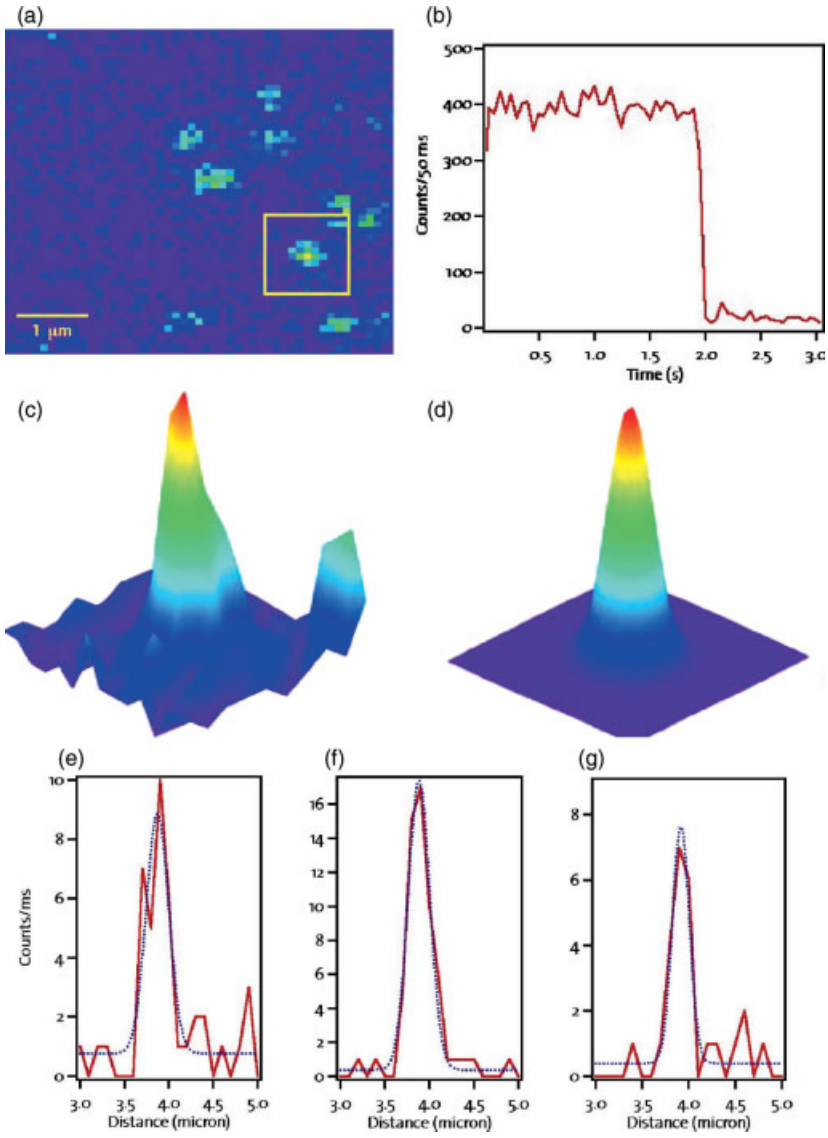


Figure 5.3 High-precision localization of single emitters. (a) Confocal fluorescence scan of a glass coverslip coated with single dye-labeled GM1 receptors. Pixel dwell time: 1 ms; illumination intensity: $\sim$1 kW cm^{-2}; total acquisition time: 10 s; (b) Detector counts as a function of time for the molecule highlighted in (a). The single-step bleaching demonstrates single-molecule sensitivity; (c) Surface plot of the fluorescence collected during the scan from the highlighted molecule; (d) Two-dimensional Gaussian fit to (c); (e–g) Three slices along the *x*-axis of the fluorescence spot (red) accompanied by the corresponding 1-D Gaussian fits (blue, dotted).

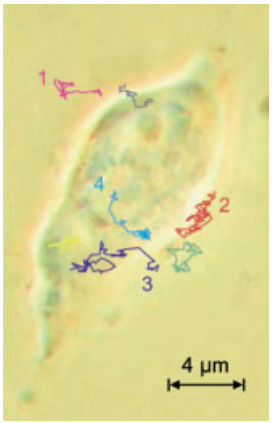


Figure 5.4 Single virus tracking. The various trajectories describe different stages of the infection pathway such as diffusion in solution (1, 2), penetration of the cell membrane (3), diffusion in the cytoplasm (3, 4), penetration of the nuclear envelope (4) and diffusion in the nucleoplasm (5). Adapted from Ref. [47].

5.2.4
Getting Around the Rayleigh Limit: Colocalization of Multiple Emitters

While it may be possible to localize a single emitter with a precision comparable to its size, the fundamental problem defined by Rayleigh's criterion remains if identical objects that are separated by less than half the wavelength of light are to be resolved. This leaves a large gap between the achievable (~200 nm) and the desired molecular resolution (~1 nm). Several approaches have been explored over the past decades, many of which have closed this gap partially and are described in detail [55]. Here, we will focus on concepts that are particularly suited toward the study of single emitters.

One approach is based on the idea that if one emitter could be observed without the other, then the location of each could be pinpointed with high precision by using the methodology outlined above. One way to achieve this task would be to image spectrally orthogonal emitters, as demonstrated by Weiss and coworkers [56], through the use of two inorganic quantum dots fluorescing at different wavelengths. Each of the emitters can be observed independently by separating the emission of the two with a dichroic mirror. Unfortunately, this multicolor colocalization can only be applied to a few particles because fluorescence emission is generally broad, and it is not realistic to use more than two or three emitters simultaneously.

Rather than differentiating between emitters spectrally, another option would be to do this temporally. An early implementation of this approach involved the stepwise photobleaching of individual emitters, as demonstrated by Selvin *et al.* [57]. Here, two emitters (m_1 and m_2) are imaged continuously. The integrated intensity of the emission of both molecules as a function of time shows a two-step behavior due to

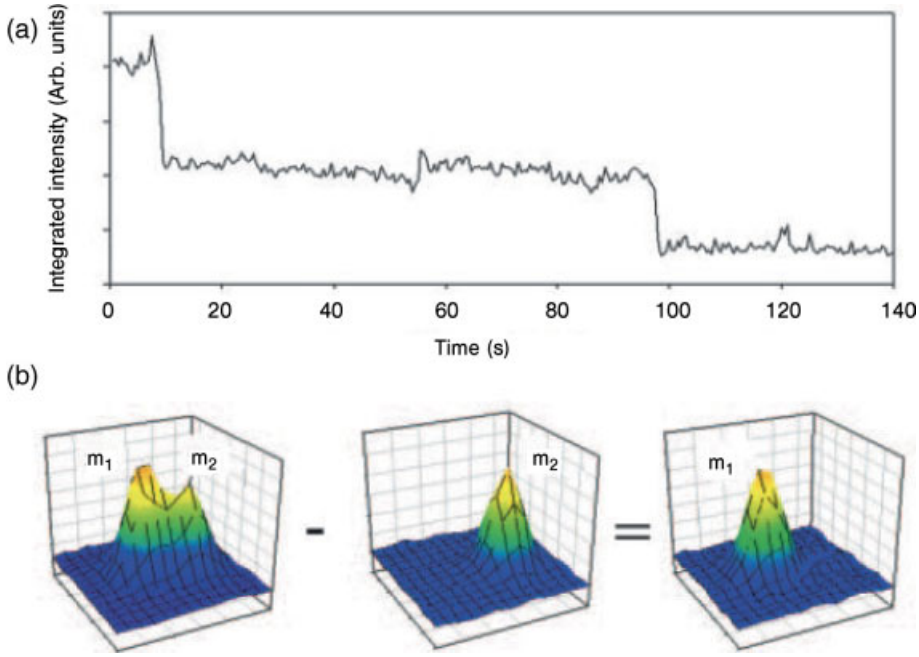


Figure 5.5 Subdiffraction localization through stepwise photobleaching. (a) Summed emission intensity showing clear two-step photobleaching; (b) Schematic representation of the localization procedure. Adapted from Ref. [57].

photobleaching, such as that shown in Figure 5.5a. After the first bleaching event, only m_2 is visible, and this can now be localized with high precision (as described previously). Subtracting the contribution of m_2 from the initial image, where both emitters are present, results in an image representative of m_1 (Figure 5.5b) which can again be localized with high precision. Thus, it is possible to determine the positions of several emitters with near-molecular resolution (down to 1.5 nm). This technique is extremely useful and precise for imaging fairly simple samples containing few fluorophores. However, for general applications such as those required for *in vivo* imaging where many labels are present, it quickly reaches its limit.

This barrier has recently been lifted by a recent approach proposed by Betzig and coworkers [58] and by Zhuang and colleagues [59], based on photoswitchable fluorophores. These methods have been named PALM (photoactivated localization microscopy) and STORM (stochastic optical reconstruction microscopy), respectively. Here, the problem of multiple fluorophores emitting simultaneously is eliminated by initially illuminating the sample in the near-UV (405 nm) at low light levels. This causes a small and stochastically distributed fraction of the total molecules to convert photochemically into an active state. Illumination of the sample in the yellow region (561 nm) then causes the photoactivated molecules to fluoresce.

By observing and fitting the emission pattern from each of these molecules, they can be localized with high precision. Those molecules are subsequently bleached and the process is repeated until the entire sample has been imaged. The resulting images are of spectacular clarity and resolution, especially when compared to standard confocal images, such as in Figure 5.6a–d. The main disadvantage of this approach is currently the low time resolution required by the stochastic activation of a small number of emitters and the reliance on the destruction of the activated signal.

The highest spatial resolution in three dimensions has been achieved at cryogenic temperatures, taking advantage of spectral selectivity [60]. When cooled to a few Kelvin, the absorption and emission from single molecules becomes extremely narrow and highly sensitive to their local environment. Single molecules can then be excited individually and therefore localized with nanometer precision. Whilst, in

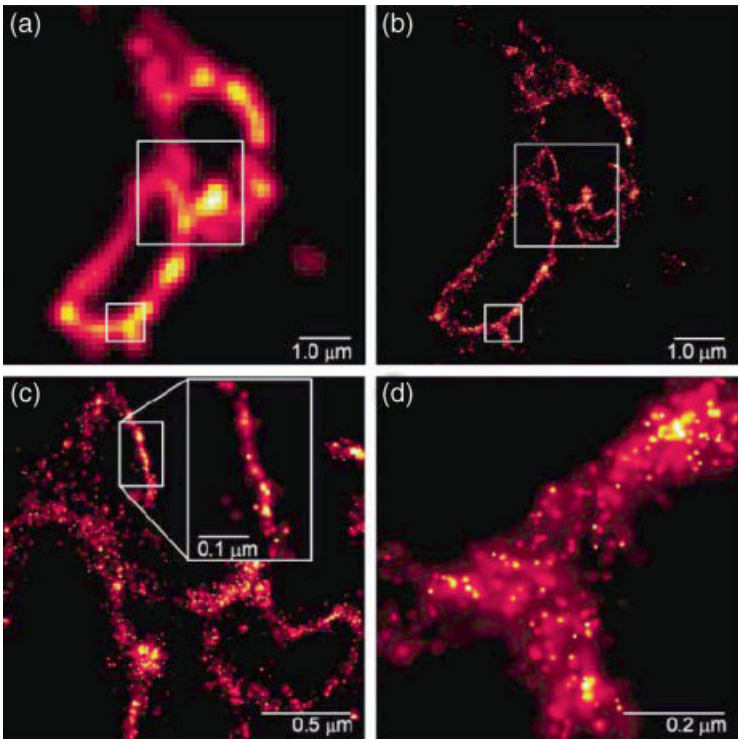


Figure 5.6 Near-molecular resolution using photoactivated localization microscopy (PALM). Comparison between summed molecule TIRF (a) and PALM (b) images from a thin, cryoprepared section of a fixed cell. An enlargement of the large boxed region in (b) reveals smaller associated membranes (c). The inset shows highly localized (10 nm) molecules. An enlargement of the smaller box in (c) shows the distribution of individual molecules within the membrane (d). Adapted from Ref. [58].

principle, this approach should be extendable to biological studies, none has so far been reported due to the high degree of experimental complexity compared to room-temperature studies.

One of the most successful and widespread approaches to achieve spatial information far below the Rayleigh limit involves taking advantage of fluorescence resonance energy transfer (FRET). The principle of this approach is depicted schematically in Figure 5.7. It uses two chromophores, a donor (D) and an acceptor (A), with overlapping absorption and emission bands (Figure 5.7b). When the two emitters are well separated, excitation of the donor will lead to observable emission only from the donor. However, when the two fluorophores are brought into close proximity of each other, the excitation energy originally placed in the donor is efficiently transferred to the acceptor due to the overlapping absorption and emission bands through Förster energy transfer (Figure 5.7c). This indirect excitation of the acceptor leads to fluorescence emission of the acceptor (i.e. far red-shifted compared to that of the donor). Commonly, the emission channels of both the donor and the acceptor are monitored simultaneously by the use of appropriate beam splitters. Thus, a dynamic system where the distance of the two

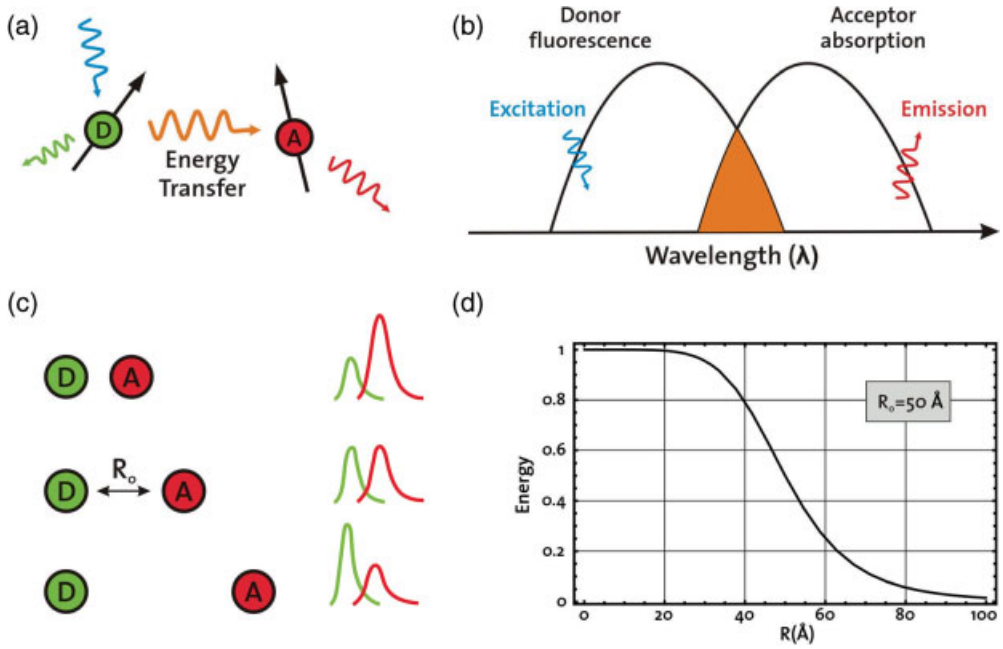


Figure 5.7 Fluorescence resonance energy transfer (FRET). (a) When a donor (D) and an acceptor (A) molecule with overlapping emission and absorption bands (b) are brought into close proximity, energy from the donor is transferred to the acceptor. In this case, red-shifted emission from the acceptor is observed; (c) When the two molecules are separated, donor fluorescence dominates; (d) The distance between the two molecules can be determined with high precision in the 2–8 nm range from the donor to acceptor emission ratio.

labels changes with time, for example due to conformational changes of a protein, will exhibit alternating emission from donor and acceptor [61]. The strong distance-dependence (R^6) of energy transfer efficiency makes FRET an excellent molecular ruler on the sub-10 nm length scale (Figure 5.7d).

5.3
Detection of *Non*-Fluorescent Single Nano-Objects

Despite the amazing capabilities of microscopy using single-molecule labels, two major limitations have become apparent. First, the total number of detectable photons is limited due to photobleaching, thus restricting the total observation times to a few seconds. Second, the saturation properties of single molecules limit the photon count rates to around $10^5\,s^{-1}$ or less. The requirement for a total of $\sim$100 detected photons for reasonable localization [52] results in a maximal time resolution on the order of milliseconds.

Many of the difficulties described above can be eliminated through the use of scattering rather than fluorescence as a contrast mechanism. Here, there is no limit on the number of detected photons because the amount of scattered photons depends only on the incident light level. As a result, an unlimited time resolution is theoretically possible. In addition, because a scattering object acts like a tiny 'mirror', neither bleaching nor blinking is an issue, providing indefinitely long unlimited observation times without dark periods. These advantages have led to the emergence of gold nanoparticles in biological applications with reported time resolutions down to $\sim$25 μs [62]. The downside associated with the use of gold nanoparticles as optical labels is the strong size dependence of the scattering cross-section, which results in a minimum label size on the order of $\sim$30 nm in dark-field detection [14, 63]. The fact that many biological nanoparticles of interest are either comparable (e.g. viruses) or much smaller (proteins) in size than this has restricted the applicability of these labels. In particular, such large labels may strongly perturb the motion of the entity under study.

5.3.1
The Difficulty of Detecting Small Particles Through Light Scattering

The ultimate goal is to detect a single molecule-sized scatterer, as this would combine the advantages of scattering detection and the minimal perturbation of the system. The question becomes: why is it so difficult to detect an object such as a 1 nm gold nanoparticle?

In many ways, the origin of this problem is very similar to that discussed earlier for single molecules. If a single-molecule-sized gold particle could be trapped in a vacuum, the light scattered by the particle could be observed by the naked eye. In a realistic environment, however, background scattering will easily overwhelm the tiny signal generated by the gold particle. For single-molecule fluorescence, one is limited by the maximum photon emission rate, background fluorescence and a small

absorption cross-section. For scattering detection, the situation is even worse because *every* object in the focal volume scatters light and the spectral selectivity available in fluorescence detection is lost, making low background measurements a true challenge.There are two possible sol utions to this problem: (i) the number of background scatterers from the focal volume is minimized; or (ii) the scattering signal from the particles of interest is somehow increased. The former has been the traditional approach to detecting small gold nanoparticles through dark-field or total internal reflection microscopy [14]. In this case, the scattered light intensity depends on the square of the polarizability of the particle which, in the electrostatic approximation, can be written as [64]:

$$\alpha(\lambda) = \frac{\pi d^3}{2} \frac{\varepsilon_p - \varepsilon_m}{\varepsilon_p + 2\varepsilon_m}$$

for a spherical particle of diameter d and dielectric constant ε_p inside a medium of dielectric constant ε_m. It easily follows, that the scattering intensity scales to the sixth power with the particle size; a 5 nm particle will therefore scatter light a million times less efficiently than a 50 nm particle! As a consequence, the signal from scatterers smaller than 30 nm drops below the background, even when a dark-field approach is used [14, 63].

Metallic nanoparticles such as those composed of gold or silver exhibit a so-called 'plasmon resonance'. As can be seen in the equation, such a resonance causes the denominator to approach zero in the specific case of $\varepsilon_p(\lambda) \rightarrow -2\varepsilon_m(\lambda)$, and therefore makes the scattering amplitude large. For spherical gold nanoparticles, this resonance occurs conveniently in the visible region of the spectrum at ~530 nm [65]. For biological nanoparticles, the scattering cross-sections are roughly a factor of three smaller than for gold because of the missing plasmon resonance [66].

To circumvent these difficulties, two approaches have emerged recently that allow the detection of gold nanoparticles down to 5 nm. One method involves taking advantage of the absorption of gold nanoparticles, which scales linearly with particle volume and therefore with the third power of the particle diameter. The resulting drop in the sensitivity of signal on the particle size has been utilized both in direct absorption measurements [67] and in the observation of a change in the refractive index through heating of the surrounding medium caused by the absorption of radiation [68]. We will focus here on an alternative method that measures the electric field directly, thereby achieving d^3 sensitivity without the need to heat the sample [69–71]. As we will show, this approach also brings with it the unique advantage of being able to detect biological nanoparticles *without any labels*.

5.3.2
Interferometric Detection of Gold Nanoparticles

To illustrate our approach, it is useful to consider the simple experimental set-up shown in Figure 5.8. Here, the incident laser beam is focused onto the sample, which in the simplest case consists of gold nanoparticles on a glass coverslip. As detection

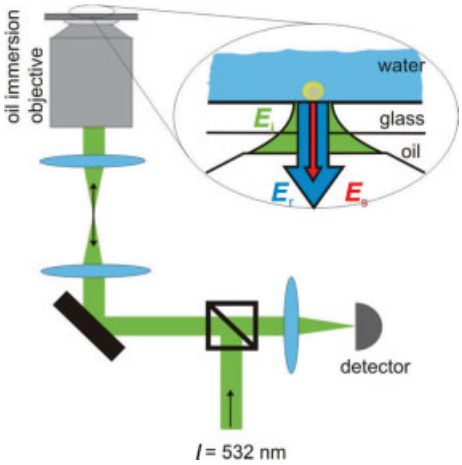


Figure 5.8 Simplified experimental set-up for the interferometric detection of nonfluorescent nanoparticles. The incident light, E_i, is reflected at the glass/water interface (E_r) and collected along with the scattered light from the particle, E_s, by the microscope objective. A portion of this light then passes through the beam splitter and reaches the detector.

occurs in the epi-direction, we are interested in the light returning through the microscope objective. The detector will therefore see the incident light (E_i) reflected at the interface between the glass and the medium. The reflected field reads $E_r = rE_i(\exp(-i\pi/2))$ where r is the field reflectivity and $\pi/2$ denotes the Gouy phase of the reflected focused beam. In addition, light scattered by the particle can be written as $E_s = sE_i = |s|\exp(i\varphi)E_i$ at the detector, where s is proportional to the particle polarizability and therefore to d^3. Here, ϕ signifies the phase change on scattering. The intensity, I_D, measured at the detector is thus

$$I_D = |E_r + E_s|^2 = |E_i|^2(r^2 + s^2 - |r||s|\sin\phi).$$

We can use this equation to illustrate some of the factors discussed above. Dark-field or total internal reflection detection are designed in such a way that $r \rightarrow 0$ and therefore only the scattering term, s^2, is detected. This signal drops very rapidly below the background level ($|rE_i|^2$) for particles <30 nm. In this case, the nature of the observed signal depends on the relative magnitudes of the three terms in the above equation. For large particles, the scattering term, s^2, dominates and the particles appear bright against the background (Figure 5.9). As the particle size decreases, s^2 becomes negligible compared to the other two terms, and only the reflection and interference terms contribute to the detected intensity. The particles appear dark against the background due to the destructive interference between the scattered and the reflected beams caused by a $-\pi/2$ Gouy phase shift of the reflected incident beam (Figure 5.10a and b) [71]. The change-over from bright to dark occurs according to the relative magnitudes of the scattering and interference terms, and therefore occurs

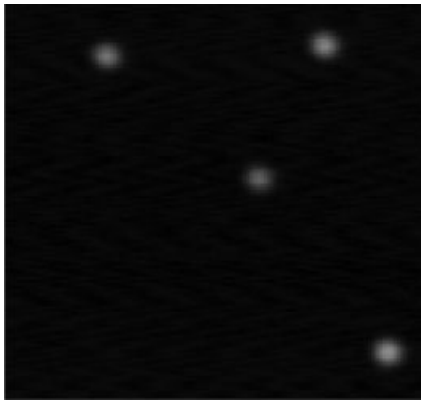


Figure 5.9 Confocal scan of 100 nm gold nanoparticles.

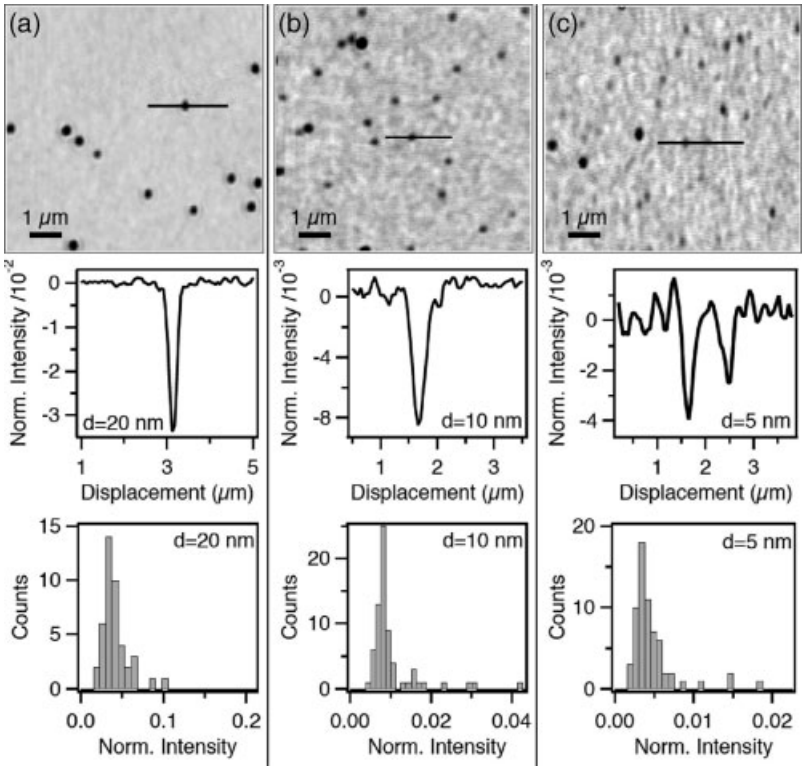


Figure 5.10 Interferometric images of gold nanoparticles spin-coated onto glass coverslips. (a) 20 nm; (b) 10 nm; (c) 5 nm diameter. Representative particle cross-sections, as well as intensity histograms, are provided in each case. Total acquisition time in each case = 10 s; incident power = 2 mW.

earlier for larger r: for example, at 40 nm for air as the surrounding medium, 30 nm for water, and 15 nm for oil that is index-matched fairly well to the glass coverslip.

5.3.2.1 Is it Possible to Detect Molecule-Sized Labels?

To explore the theoretical sensitivity limitations, it is useful to consider the origin of the true noise background that limits the detection sensitivity. This is governed mostly by the noise of the light source itself and the noise of the detector. Both will result in fluctuations in the detected reflected intensity, r, which is the major contributor to the overall detected signal at the detector. Other potential noise sources such as mechanical instabilities of the microscope or beam-pointing instability of the light source are comparatively small and easily corrected for in post-acquisition image processing.

An incident power of 1 mW on the sample will yield $\sim$3 μW of light reaching the detector, taking into account the reflectivity of the glass–water interface and losses due to the limited transmission of optics such as the microscope objective. The ideal detectors for such light intensities are photodiodes, which produce a corresponding photocurrent of 1 μA. The shot noise limit for this photocurrent is on the order of 500 fA $Hz^{-1/2}$, which is about an order of magnitude above the noise of available amplifiers with 10^7 V/A gain, suggesting that shot noise-limited detection is possible. At this amplification and a realistic detection bandwidth for mechanical scanning of 1 kHz, the electronic shot noise amounts to $\sim 1.5 \times 10^{-5}$ rms, which is a factor of $\sim$300 below the magnitude of the signal observed for 5 nm gold particles. A factor of three reduction in size on the other hand, which would lead to molecular sized labels on the order of 1.3 nm, brings with it a factor of 27 reduction in signal intensity. Thus, such molecular-sized labels should be observable with a SNR of 10 at kHz bandwidths with localization accuracies down to 10 nm!

The previous discussion has shown that neither shot noise nor detector noise limit the detectability of such small labels. One other critical noise source remains: the light source itself. Lasers used in confocal microscopes show optical noise on the order of a small percentage over a wide frequency range. Even state-of-the-art, solid-state, diode-pumped lasers rarely perform better than 0.1%. However, external stabilization using optical fibers, acousto-optic modulators and feedback loops has been shown to reduce laser noise from a small percentage to $\sim 5 \times 10^{-5}$ rms with kHz bandwidth [72], which is comparable to the electronic shot noise calculated above. In addition, the use of single-mode fibers in this stabilization scheme significantly reduces the effects of beam-pointing and mode instabilities, further contributing to the overall stability of the system. Given these simple calculations, it becomes evident that the rapid detection of molecular-sized gold scatterers should be possible. We are currently pursuing such experiments.

All of the images presented in Figure 5.10 have been obtained by scanning the sample across the focus using a piezo translation stage that requires 1–10 s per image. By using scanning mirrors rather than a piezo stage, we have shown previously that it is possible to detect 20 nm particles with up to MHz bandwidths – three orders of magnitude above what is possible using single molecules as labels [73]. In addition, rather than scanning the focus across the surface, the use of a feedback loop enables one to lock the focus to a particle and follow its movements rapidly. The feedback loop

is fed by the signal recorded on a four-quadrant detector, where the movement of the particle inside the focus leads to changes in the measured differential voltages. Using the same detector and amplifier combination above, which provide MHz detection bandwidths, the shot noise increases to $\sim 5 \times 10^{-4}$ with mW incident powers. In this way, the tracking of labels as small as 5 nm with MHz bandwidths should be possible.

The sensitivity limitations of the technique are illustrated in Figure 5.10c, which shows a confocal image of 5 nm particles spin-coated on a glass coverslip and covered by water. As can be seen in the image, the particles are visible against the background with a signal contrast on the order of 3×10^{-3}. The distribution width of the signal intensities is in agreement with the manufacturer's specifications with regards to the size of the gold particles, confirming that we are indeed observing single particles.

A close inspection of Figure 5.10b and c reveals the presence of a rather 'noisy' background as the size of the particles and thus their signal magnitude decreases. However, these features are *not* noise, as they are perfectly reproducible in sequential images. Rather, these patterns are due to the surface roughness of the glass coverslips used. Indeed, AFM measurements have shown that the surface roughness amounts to a few nanometers over a few microns. The reproducibility of such nanometer-sized surface roughness demonstrates the excellent sensitivity of this technique to nonmetallic species.

5.3.2.2 The Needle in the Haystack: Finding and Identifying Gold

So far, we have been concerned mostly with detecting and tracking the smallest possible gold nanoparticles. An interesting point to address is how such small labels can be detected in the presence of much larger scatterers, for example in intracellular imaging. Fortunately, gold nanoparticles have a type of 'built-in identification card' in the form of a plasmon resonance in the visible region of the electromagnetic spectrum (Figure 5.11a). As a result, one obtains roughly twice the scattering intensity in the green (532 nm) compared to the blue (488 nm) or red (>560 nm) regions. This wavelength-dependent scattering intensity is in contrast to the constituents of typical biological samples, where the scattering should be roughly identical for both wavelengths.

To demonstrate the possibility of using this interesting feature of gold nanoparticles, we have labeled microtubules with 40 nm gold particles and obtained scattering images simultaneously in the blue and green (Figure 5.11b and c). In the two images, both the nanoparticles and the microtubule are clearly visible. However, when the two images are subtracted from each other (Figure 5.11d), the microtubule disappears while the particles remain. One can thus use this form of differential spectral contrast to ensure that the observed particles are indeed the gold labels of interest and not other scatterers [69].

5.3.3 Combining Scattering and Fluorescence Detection: A Long-Range Nanoscopic Ruler

As yet, we have discussed the advantages and disadvantages of fluorescence and scattering as labels only in biological imaging applications. Now, we present an

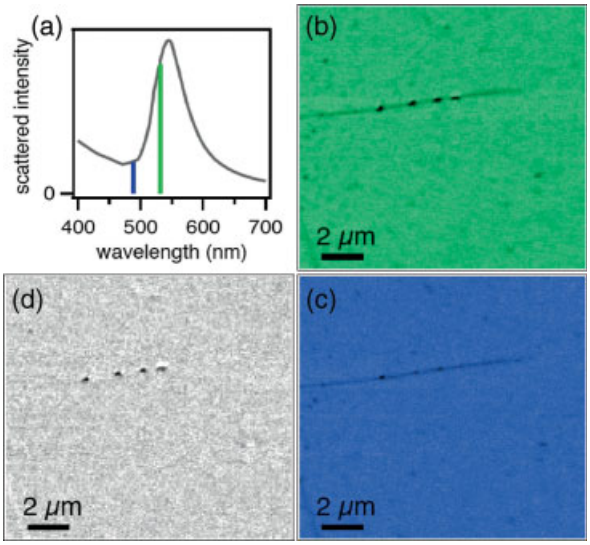


Figure 5.11 Identification of gold scatterers through spectral difference. (a) Plasmon resonance for a 40 nm gold particle; (b) Interferometric image of microtubules labeled with individual 40 nm gold particles acquired at 532 nm; (c) Identical image acquired simultaneously at 488 nm; (d) Subtraction of the blue from the green image. The microtubule with approximately identical scattering cross-sections at the two wavelengths disappears, while the gold particles remain visible.

example where the combination of the two leads to a potentially useful technique. It has been shown previously that nanostructures brought into close vicinity of single emitters can cause enhancement of luminescence and Raman scattering [75]. We have shown recently that a single gold nanoparticle can enhance the fluorescence of a single molecule and the decay rate of its excited state by a factor of 20 [76]. Furthermore, we demonstrated that this strong fluorescence modification is a function of the particle–emitter separation with nanometer sensitivity.

The mechanism of this effect can be intuitively explained as the near-field interaction of the molecular dipole moment, with its image dipole induced in the gold nanoparticle. The dipole–dipole character of this interaction gives rise to a strong distance-dependence much in the same way as in FRET (see also Figure 5.10). However, in this case the interaction range drops much more softly than the $1/r^6$ dependence observed in FRET. Thus, the modification of the fluorescence lifetime close to a nanoparticle can be used as a nanoscopic ruler for distances larger than that of FRET (>10 nm).

To demonstrate this, we have performed studies of systems where a single molecule is linked to a gold nanoparticle with DNA double strands of differing lengths [77]. The techniques of single-molecule detection and microscopy of gold nanoparticles were combined to locate such molecule–particle pairs. The corresponding confocal scans of single functionalized gold nanoparticles of 15 nm

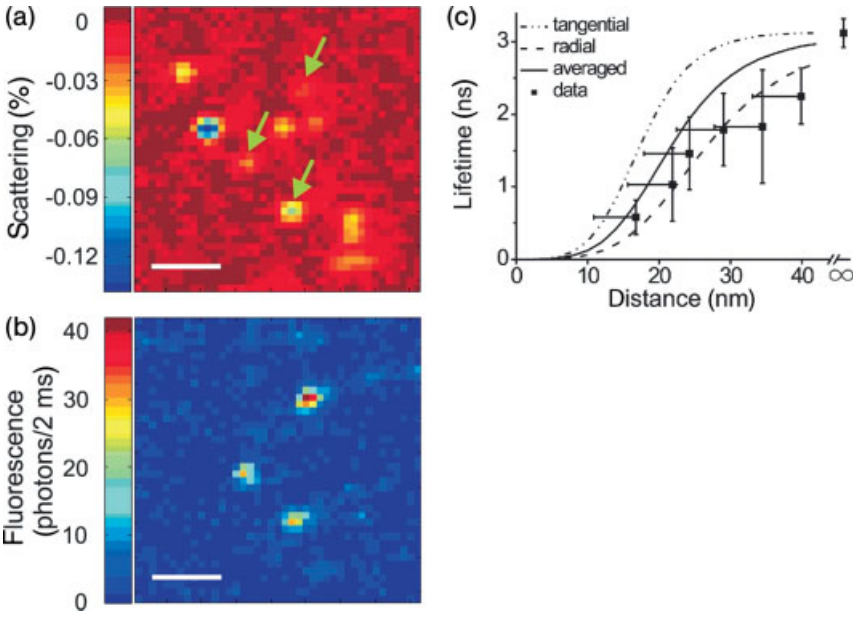


Figure 5.12 Combining scattering and fluorescence detection for a nanoscopic ruler. (a) Scattering image of single 15 nm gold particles functionalized with a single dye molecule via a DNA linker; (b) Simultaneously acquired fluorescence image. The arrows in (a) indicate gold particles that are functionalized with fluorescent markers; (c) Single-molecule fluorescence lifetime dependence on linker length. The dashed and dashed-dotted curves display the calculated fluorescence lifetime for the molecular dipole oriented radially or tangentially with respect to the gold nanoparticle. The solid curve shows the weighted average of the two orientations.

diameter are shown in Figure 5.12a (scattering) and b (fluorescence). As can be seen in the figure, only a fraction of the gold particles contains a fluorescent marker. Figure 5.12c demonstrates the dependence of fluorescence lifetime on linker length. In the absence of a gold nanoparticle, the fluorescence lifetime of the molecule was ~3 ns, but this was reduced to about 0.6 ns for a 15 nm-long DNA linker consisting of 44 base pairs. The interaction length and its slope can be tuned by choosing the particle size and emission wavelength of the dye molecule. The precision of such a nanoscopic ruler is on the order of 1 nm, and is limited by the accuracy with which the fluorescence lifetime can be determined [77].

5.3.4
Label-Free Detection of Biological Nano-Objects

Although we have focused mostly on gold as a scattering label, the previous discussion has also shown that interferometric detection is extremely sensitive to virtually any type of scatterer. This is demonstrated on the one hand by the

observation of nanometer surface roughness on glass coverslips, and on the other hand by the visibility of individual microtubules which are hollow shells of only 24 nm diameter. These results suggest that it may be possible to detect biological nano-objects *without the need for any label*. Such detection brings with it the aforementioned advantages of scattering detection, but more importantly eliminates any outside perturbation on the system which may be introduced by labels; moreover, it eliminates the need for labeling chemistry.

To illustrate the capabilities of the technique in this respect, we have obtained scattering images of unlabeled Simian virus 40 (SV40) virions bound to microscope coverslips. SV40 is a small, 45 nm-diameter, tumor-virus consisting of an outer protein shell of 720 copies of the VP1 protein, with a 5000 base-pair DNA-genome in its core (Figure 5.13a) [78]. As can be seen from the image in Figure 5.13b,

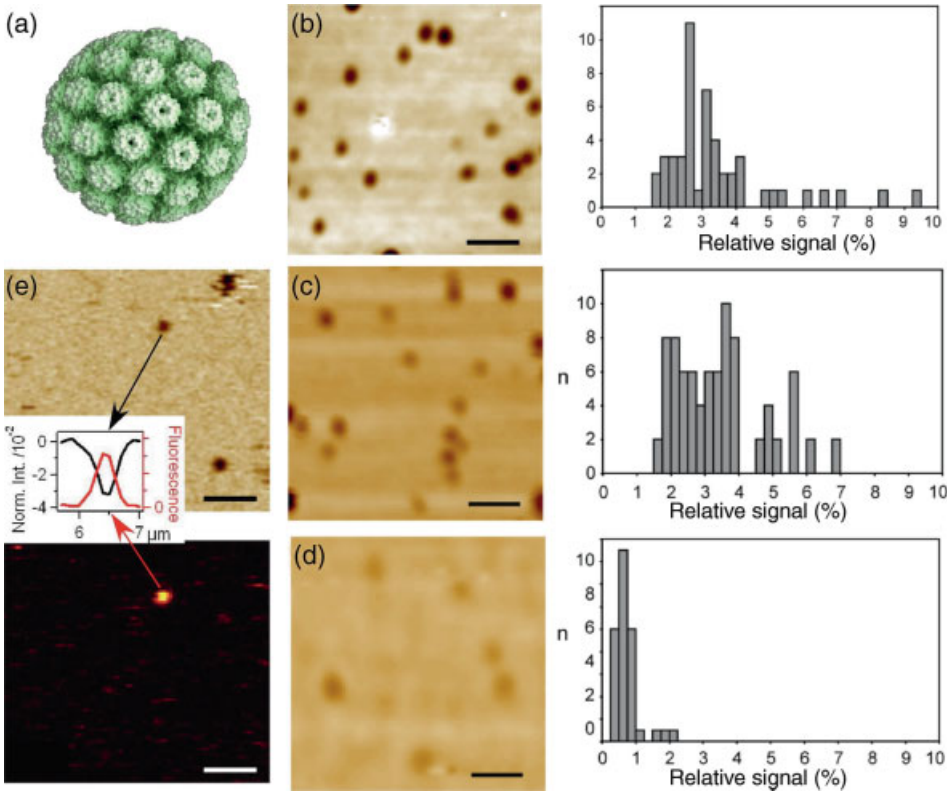


Figure 5.13 Label-free detection of biological nanoparticles. (a) Structure of the SV40 virus as determined from X-ray crystallography; (b) Interferometric images of SV40 adhered to glass and (c) on a supported membrane bilayer; (d) Virus-like particles on a supported membrane bilayer; (e) Simultaneous scattering and fluorescence images of single SV40 viruses labeled with multiple atto-565 fluorophores. The inset shows a cross-section of the particle for both images. Scale bars = 1 µm.

a 45 nm virus shows roughly the same scattering intensity as a 20 nm gold particle ($\sim 3 \times 10^{-2}$). The difference in the observed scattering intensity is due to the lacking plasmon resonance of the virus, and thereby to a reduced polarizability. To test the applicability of this approach in a biologically relevant environment, we have also obtained images of SV40 virions bound to supported lipid bilayer membranes (Figure 5.13c) [66]. Prior to addition of the virus, the membranes showed a homogeneous background signal and few detectable spots caused by unfused vesicles. Within seconds of adding the virus to the solution, Gaussian spots appeared in the image, corresponding to single viruses binding to GM1 pentasaccharide receptors that had been added to the membrane. The signal intensity was comparable to that of viruses bound to the glass coverslips (2.6×10^{-2}). To confirm the validity of the interpretation, we performed two further experiments. First, we acquired images of SV40 virus-like particles, which are identical to the virions, except that the DNA core had been removed. As might be expected, due to the reduced amount of material, the observed signal intensity was lower compared to the viruses at 0.75×10^{-2} (Figure 5.13d). Second, we checked our images by simultaneously acquiring fluorescence and scattering images of fluorescence-labeled SV40 viruses on a supported membrane bilayer (Figure 5.13e). As can be seen in the figure, the two images corresponded perfectly. The inset shows a slice along the virus, demonstrating the complementary nature of the two signals.

In addition, we have performed consecutive confocal scans to investigate diffusion of the virus on the membrane. The trace of the viral motion is superimposed on the image in Figure 5.14a. Computational analysis of the trajectories yielded linear mean square displacement plots, as would be expected for particles undergoing Brownian motion (Figure 5.14b), and exhibited a diffusion constant (D) of 0.0088 ± 0.0004 $\mu m^2\, s^{-1}$. These first results demonstrate the power of the interferometric detection of nano-objects, and its potential for the long-term tracking of unlabelled biological entities. The requirement for this is of course a well-defined sample where unwanted scattering has been eliminated, as in the case of the membrane studies presented here.

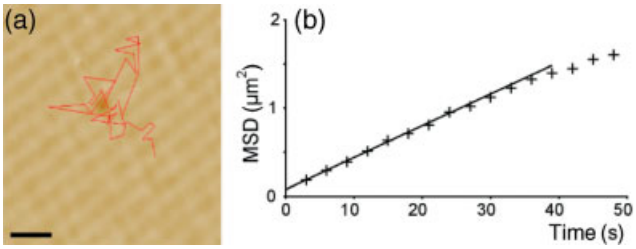


Figure 5.14 Label-free tracking of a single virus diffusing on a supported membrane bilayer. (a) Motion of the virus with 1 s time-resolution acquired over 50 s; (b) Corresponding mean square displacement as deduced from the observed motion in (a), indicating Brownian motion.

5.4 Summary and Outlook

We have discussed the capabilities and limitations of single-molecule fluorescence microscopy, with particular attention being paid to the critical parameters in biological imaging such as the SNR, time resolution and observation time. We have seen that, by detecting and localizing individual fluorescent molecules in a sample, the resolution in optical microscopy can be pushed down to the 1–10 nm regime. The resolution limit in this method is dictated by the noise of the fluorescence signal, and therefore, by the number of photons recorded from each emitter. Currently, problems such as photobleaching and photoblinking prevent an arbitrarily high resolution being achieved in realistic systems. Any efforts to suppress or minimize photobleaching are, therefore, of utmost importance to the future of high-resolution optical microscopy. One interesting possibility is to perform single-molecule detection at cryogenic temperatures where photochemistry is slowed tremendously.

For some imaging and real-time *in vivo* tracking applications, a very promising solution is offered by metallic nanoparticles as alternative labels. The lack of saturation, photobleaching and blinking in light scattering makes such labels ideal candidates to avoid many of the pitfalls of single-molecule spectroscopy. In particular, we have shown how recent advances in the interferometric detection of single gold nanoparticles enable the observation of such labels with sizes down to 5 nm. The ability to illuminate the sample at high power, without saturating the signal, allows a faster integration time and thus a much improved tracking speed, to more than three orders of magnitude above what is possible in single-molecule applications. Finally, we have shown how this interferometric detection technique can be used to observe dielectric objects *without the need for any labels*. Specifically, we have demonstrated label-free detection and the tracking of single SV40 viruses diffusing on artificial lipid bilayer membranes.

The 1990s were witness to a fantastic revival of optical microscopy for high-resolution imaging. Moreover, advances in laser spectroscopy, scanning microscopy, detector technology and photophysics have made it possible to interrogate matter at the single-molecule level, using light *in vitro* and even *in vivo*. In particular, various techniques have shown that the Rayleigh limit can be tackled for specific applications, and have demonstrated optical resolution at the 1–2 nm level. These improvements provide exciting new tools to study a wealth of biological questions at the subcellular level with a spatial resolution more than tenfold higher than can be achieved with conventional confocal microscopy. In other words, the resolution gap between optical microscopy and electron microscopy has been made much smaller. However, what is especially valuable is that this high optical resolution provides a major opportunity to open our eyes to the real-time life processes that occur within a functioning cell.

The road to the optical visualization of every single molecule in the sample, along with its trajectory in the time domain, remains long. However, if the rapid progress made during the past two decades continues then we will have good reason to feel that

this chapter will soon become somewhat of an 'antique'. In fact, since the first concept of this chapter, video rate fluorescence imaging with a focal spot of approximately 60 nm in living cells has been achieved using RESOLFT, and this has resulted in some impressive images of synaptic vesicle movement [79]. Furthermore, the initial 2-D studies with PALM and STORM have now been extended to 3-D imaging with a lateral resolution of 20–30 nm and an axial resolution of 50–60 nm [80]. Finally, single-molecule detection has been successfully extended to the investigation of single labels, such as semiconductor quantum dots, in *absorption* at room temperature. This will surely open the way to optical nanoscopy without a need for efficient fluorescent labels [74].

Acknowledgments

The authors thank U. Hakanson, V. Jacobsen, E. Klotzsch, K. Lindfors, A. Schtalheim, J. Seelig and P. Stoller, who each contributed to the development of the interferometric detection technique, and C. Brunner, H. Ewers, A. Helenius, K. Leslie, A. Smith, V. Vogel and C. Weyman for their collaboration on the biological experiments. These studies were performed within the frame of an Integrated Project of the European Union 'Molecular Imaging'. The authors also acknowledge financial support from the ETH Zurich and the Swiss National Foundation (SNF).

References

1 Preston, G.M. and Agre, P. (1991) Isolation of the CDNA for erythrocyte integral membrane-protein of 28-kilodaltons - member of an ancient channel family. *Proceedings of the National Academy of Sciences of the United States of America*, **88**, 11110.

2 Doyle, D.A., Cabral, J.M., Pfuetzner, R.A., Kuo, L., Gulbis, J.M., Cohen, S.L., Chait, B.T. and MacKinnon, R. (1998) The structure of the potassium channel: molecular basis of K+ conduction and selectivity. *Science*, **280**, 69.

3 Palczewski, K., Kumasaka, T., Hori, T., Behnke, C.A., Motoshima, H., Fox, B.A., Le Trong, I., Teller, D.C., Okada, T., Stenkamp, R.E., Yamamoto, M. and Miyano, M. (2000) Crystal structure of rhodopsin: A G protein-coupled receptor. *Science*, **289**, 739.

4 Cramer, P., Bushnell, D.A. and Kornberg, R.D. (2001) Structural basis of transcription: RNA polymerase II at 2.8 Angstrom resolution. *Science*, **292**, 1863.

5 Binnig, G. and Rohrer, H. (1982) Vacuum tunnel microscope. *Helvetica Physica Acta*, **55**, 726.

6 Binnig, G., Quate, C.F. and Gerber, C. (1986) Atomic force microscope. *Physical Review Letters*, **56**, 930.

7 Shao, Z.F., Yang, J. and Somlyo, A.P. (1995) Biological atomic force microscopy: from microns to nanometers and beyond. *Annual Review of Cell and Developmental Biology*, **11**, 241.

8 Hansma, H.G. (2001) Surface biology of DNA by atomic force microscopy. *Annual Review of Physical Chemistry*, **52**, 71.

9 Davis, L.I. (1995) The nuclear-pore complex. *Annual Review of Biochemistry*, **64**, 865.

10 Pluta, M. (1989) *Advanced Light Microscopy*, Elsevier, Amsterdam.

11 Horio, T. and Hotani, H. (1986) Visualization of the dynamic instability of individual microtubules by dark-field microscopy. *Nature*, **321**, 605.

12 Prieve, D.C., Luo, F. and Lanni, F. (1987) Brownian-motion of a hydrosol particle in a colloidal force-field. *Faraday Discussions*, 297.

13 Joos, U., Biskup, T., Ernst, O., Westphal, I., Gherasim, C., Schmidt, R., Edinger, K., Pilarczyk, G. and Duschl, C. (2006) Investigation of cell adhesion to structured surfaces using total internal reflection fluorescence and confocal laser scanning microscopy. *European Journal of Cell Biology*, **85**, 225.

14 Sonnichsen, C., Geier, S., Hecker, N.E., von Plessen, G., Feldmann, J., Ditlbacher, H., Lamprecht, B., Krenn, J.R., Aussenegg, F.R., Chan, V.Z.H., Spatz, J.P. and Moller, M. (2000) Spectroscopy of single metallic nanoparticles using total internal reflection microscopy. *Applied Physics Letters*, **77**, 2949.

15 Minsky, M. (1988) Memoir on inventing the confocal scanning microscope. *Scanning*, **10**, 128.

16 Wilson, T. (1990) *Confocal Microscopy*, Academic Press, London.

17 Nomarski, G. (1955) Nouveau dispositif pour l'observation en contraste de phase differentiel. *Journal de Physique et le Radium*, **16**, S88.

18 Murphy, D.B. (2001) *Fundamentals of Light Microscopy and Electronic Imaging*, Wiley-Liss, New York.

19 Coons, A.H. and Kaplan, M.H. (1950) Localization of antigen in tissue cells 2. Improvements in a method for the detection of antigen by means of fluorescent antibody. *The Journal of Experimental Medicine*, **91**, 1.

20 Chalfie, M., Tu, Y., Euskirchen, G., Ward, W.W. and Prasher, D.C. (1994) Green fluorescent protein as a marker for gene-expression. *Science*, **263**, 802.

21 Axelrod, D., Koppel, D.E., Schlessinger, J., Elson, E. and Webb, W.W. (1976) Mobility measurement by analysis of fluorescence photobleaching recovery kinetics. *Biophysical Journal*, **16**, 1055.

22 Bastiaens, P.I.H. and Squire, A. (1999) Fluorescence lifetime imaging microscopy: spatial resolution of biochemical processes in the cell. *Trends In Cell Biology*, **9**, 48.

23 Richards-Kortum, R. and Sevick-Muraca, E. (1996) Quantitative optical spectroscopy for tissue diagnosis. *Annual Review of Physical Chemistry*, **47**, 555.

24 Konig, K. (2000) Multiphoton microscopy in life sciences. *Journal of Microscopy*, **200**, 83.

25 Cheng, J.X., Jia, Y.K., Zheng, G.F. and Xie, X.S. (2002) Laser-scanning coherent anti-Stokes Raman scattering microscopy and applications to cell biology. *Biophysical Journal*, **83**, 502.

26 Campagnola, P.J., Wei, M.D., Lewis, A. and Loew, L.M. (1999) High-resolution nonlinear optical imaging of live cells by second harmonic generation. *Biophysical Journal*, **77**, 3341.

27 Barad, Y., Eisenberg, H., Horowitz, M. and Silberberg, Y. (1997) Nonlinear scanning laser microscopy by third harmonic generation. *Applied Physics Letters*, **70**, 922.

28 Hell, S.W. (2007) *Science*, **316**, 1153.

29 Nagourney, W., Janik, G. and Dehmelt, H. (1983) Linewidth of single laser-cooled (Mg-24) + ion in radiofrequency trap. *Proceedings of the National Academy of Sciences of the United States of America*, **80**, 643.

30 Kim, J.E., Tauber, M.J. and Mathies, R.A. (2001) Wavelength dependent cis-trans isomerization in vision. *Biochemistry*, **40**, 13774.

31 Moerner, W.E. and Orrit, M. (1999) Illuminating single molecules in condensed matter. *Science*, **283**, 1670.

32 Xie, X.S. and Trautman, J.K. (1998) Optical studies of single molecules at room temperature. *Annual Review of Physical Chemistry*, **49**, 441.

33 Eggeling, C., Widengren, J., Rigler, R. and Seidel, C.A.M. (1998) Photobleaching of fluorescent dyes under conditions used for

single-molecule detection: evidence of two-step photolysis. *Analytical Chemistry*, **70**, 2651.

34 Renn, A., Seelig, J. and Sandoghdar, V. (2006) Oxygen-dependent photochemistry of fluorescent dyes studied at the single molecule level. *Molecular Physics*, **104**, 409.

35 Foyer, C.H., Lelandais, M. and Kunert, K.J. (1994) Photooxidative stress in plants. *Physiologia Plantarum*, **92**, 696.

36 Hoebe, R.A., Van Oven, C.H., Gadella, T.W.J., Dhonukshe, P.B., Van Noorden, C.J.F. and Manders, E.M.M. (2007) Controlled light-exposure microscopy reduces photobleaching and phototoxicity in fluorescence live-cell imaging. *Nature Biotechnology*, **25**, 249.

37 Yildiz, A., Forkey, J.N., McKinney, S.A., Ha, T., Goldman, Y.E. and Selvin, P.R. (2003) Myosin V walks hand-over-hand: single fluorophore imaging with 1.5-nm localization. *Science*, **300**, 2061.

38 Moerner, W.E. and Kador, L. (1989) Optical-detection and spectroscopy of single molecules in a solid. *Physical Review Letters*, **62**, 2535.

39 Orrit, M. and Bernard, J. (1990) Single pentacene molecules detected by fluorescence excitation in a para-terphenyl crystal. *Physical Review Letters*, **65**, 2716.

40 Pohl, D.W., Denk, W. and Lanz, M. (1984) Optical stethoscopy - image recording with resolution lambda/20. *Applied Physics Letters*, **44**, 651.

41 Lewis, A., Isaacson, M., Harootunian, A. and Muray, A. (1984) Development of a 500-A spatial-resolution light-microscope 1. Light is efficiently transmitted through gamma-16 diameter apertures. *Ultramicroscopy*, **13**, 227.

42 Betzig, E. and Chichester, R.J. (1993) Single molecules observed by near-field scanning optical microscopy. *Science*, **262**, 1422.

43 Nie, S.M., Chiu, D.T. and Zare, R.N. (1994) Probing individual molecules with confocal fluorescence microscopy. *Science*, **266**, 1018.

44 Macklin, J.J., Trautman, J.K., Harris, T.D. and Brus, L.E. (1996) Imaging and time-resolved spectroscopy of single molecules at an interface. *Science*, **272**, 255.

45 Special Issue on Single Molecules (1999) *Science*, **283**, 1667.

46 Vale, R.D., Funatsu, T., Pierce, D.W., Romberg, L., Harada, Y. and Yanagida, T. (1996) Direct observation of single kinesin molecules moving along microtubules. *Nature*, **380**, 451.

47 Seisenberger, G., Ried, M.U., Endress, T., Buning, H., Hallek, M. and Brauchle, C. (2001) Real-time single-molecule imaging of the infection pathway of an adeno-associated virus. *Science*, **294**, 1929.

48 Hecht, E. (2001) *Optics*, Addison Wesley, New York.

49 Thompson, R.E., Larson, D.R. and Webb, W.W. (2002) Precise nanometer localization analysis for individual fluorescent probes. *Biophysical Journal*, **82**, 2775.

50 Schmidt, T., Schutz, G.J., Baumgartner, W., Gruber, H.J. and Schindler, H. (1995) Characterization of photophysics and mobility of single molecules in a fluid lipid-membrane. *The Journal of Physical Chemistry*, **99**, 17662.

51 Yildiz, A., Tomishige, M., Vale, R.D. and Selvin, P.R. (2004) Kinesin walks hand-over-hand. *Science*, **303**, 676.

52 Ober, R.J., Ram, S. and Ward, E.S. (2004) Localization accuracy in single-molecule microscopy. *Biophysical Journal*, **86**, 1185.

53 Alivisatos, A.P., Gu, W.W. and Larabell, C. (2005) Quantum dots as cellular probes. *Annual Review of Biomedical Engineering*, **7**, 55.

54 Michalet, X., Pinaud, F.F., Bentolila, L.A., Tsay, J.M., Doose, S., Li, J.J., Sundaresan, G., Wu, A.M., Gambhir, S.S. and Weiss, S. (2005) Quantum dots for live cells, in vivo imaging, and diagnostics. *Science*, **307**, 538.

55 Pawley, J.B. (2006) *Handbook of Biological Confocal Microscopy*, 3rd edn, Springer Verlag.

56 Lacoste, T.D., Michalet, X., Pinaud, F., Chemla, D.S., Alivisatos, A.P. and Weiss, S. (2000) Ultrahigh-resolution multicolor colocalization of single fluorescent probes. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 9461.

57 Gordon, M.P., Ha, T. and Selvin, P.R. (2004) Single-molecule high-resolution imaging with photobleaching. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 6462.

58 Betzig, E., Patterson, G.H., Sougrat, R., Lindwasser, O.W., Olenych, S., Bonifacino, J.S., Davidson, M.W., Lippincott-Schwartz, J. and Hess, H.F. (2006) Imaging intracellular fluorescent proteins at nanometer resolution. *Science*, **313**, 1642.

59 Rust, M.J., Bates, M. and Zhuang, X.W. (2006) Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). *Nature Methods*, **3**, 793.

60 Hettich, C., Schmitt, C., Zitzmann, J., Kühn, S., Gerhardt, I. and Sandoghdar, V. (2002) Nanometer resolution and coherent optical dipole coupling of two individual molecules. *Science*, **298**, 385.

61 Ha, T., Enderle, T., Ogletree, D.F., Chemla, D.S., Selvin, P.R. and Weiss, S. (1996) Probing the interaction between two single molecules: fluorescence resonance energy transfer between a single donor and a single acceptor. *Proceedings of the National Academy of Sciences of the United States of America*, **93**, 6264.

62 Kusumi, A., Nakada, C., Ritchie, K., Murase, K., Suzuki, K., Murakoshi, H., Kasai, R.S., Kondo, J. and Fujiwara, T. (2005) Paradigm shift of the plasma membrane concept from the two-dimensional continuum fluid to the partitioned fluid: high-speed single-molecule tracking of membrane molecules. *Annual Review of Biophysics and Biomolecular Structure*, **34**, 351.

63 Schultz, S., Smith, D.R., Mock, J.J. and Schultz, D.A. (2000) Single-target molecule detection with nonbleaching multicolor optical immunolabels. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 996.

64 Bohren, C.F. and Huffman, D.R. (1983) *Absorption and Scattering of Light by Small Particles*, John Wiley and Sons, New York.

65 Kelly, K.L., Coronado, E., Zhao, L.L. and Schatz, G.C. (2003) The optical properties of metal nanoparticles: the influence of size, shape, and dielectric environment. *Journal of Physical Chemistry B*, **107**, 668.

66 Ewers, H., Jacobsen, V., Klotzsch, E., Smith, A., Helenius, A. and Sandoghdar, V. (2007) Label-free optical detection and tracking of single virions bound to their receptors in supported membrane bilayers. *Nano Letters*, **7**, 2263.

67 Arbouet, A., Christofilos, D., Del Fatti, N., Vallee, F., Huntzinger, J.R., Arnaud, L., Billaud, P. and Broyer, M. (2004) Direct measurement of the single-metal-cluster optical absorption. *Physical Review Letters*, **93**, 127401.

68 Boyer, D., Tamarat, P., Maali, A., Lounis, B. and Orrit, M. (2002) Photothermal imaging of nanometer-sized metal particles among scatterers. *Science*, **297**, 1160.

69 Jacobsen, V., Stoller, P., Brunner, C., Vogel, V. and Sandoghdar, V. (2006) Interferometric optical detection and tracking of very small gold nanoparticles at a water-glass interface. *Optics Express*, **14**, 405.

70 Ignatovich, F.V. and Novotny, L. (2006) Real-time and background-free detection of nanoscale particles. *Physical Review Letters*, 96.

71 Lindfors, K., Kalkbrenner, T., Stoller, P. and Sandoghdar, V. (2004) Detection and spectroscopy of gold nanoparticles using supercontinuum white light confocal microscopy. *Physical Review Letters*, **93**, 037401.

72 Carter, A.R., King, G.M., Ulrich, T.A., Halsey, W., Alchenberger, D. and Perkins, T.T. (2007) Stabilization of an optical

microscope to 0.1 nm in three dimensions. *Applied Optics*, **46**, 421.

73 Jacobsen, V., Klotzsch, E. and Sandoghdar, V. (2007) *Nano Biophotonics*, Vol. **3** (eds H. Masuhura, S. Kawata and F. Tokunaga), Elsevier, Amsterdam, p. 143.

74 Kukura, P., Celebrano, M., Renn, A. and Sandoghdar, V. (2008) Seeing a single quantum emitter when it is dark. *Nano Letters*, doi 10.1021/nl801735y.

75 Moskovits, M. (1985) Surface-enhanced spectroscopy. *Reviews of Modern Physics*, **57**, 783.

76 Kühn, S., Hakanson, U., Rogobete, L. and Sandoghdar, V. (2006) Enhancement of single-molecule fluorescence using a gold nanoparticle as an optical nanoantenna. *Physical Review Letters*, **97**, 017402.

77 Seelig, J., Leslie, K., Renn, A., Kühn, S., Jacobsen, V., van de Corput, M., Wyman, C. and Sandoghdar, V. (2007) Nanoparticle-induced fluorescence lifetime modification as nanoscopic ruler: demonstration at the single molecule level. *Nano Letters*, **7**, 685.

78 Liddington, R.C., Yan, Y., Moulai, J., Sahli, R., Benjamin, T.L. and Harrison, S.C. (1991) Structure of Simian virus-40 at 3.8-A resolution. *Nature*, **354**, 278.

79 Westphal, V., Rizzoli, S.O., Lauterbach, M.A., Kamin, D., Jahn, R. and Hell, S.W. (2008) Video-rate far-field optical nanoscopy dissects synaptic vesicle movement. *Science*, **320**, 246.

80 Huang, B., Wang, W., Bates, M. and Zhuang, X. (2008) Three-dimensional super-resolution imaging by stochastic optical reconstruction microscopy. *Science*, **319**, 810.

6
Nanostructured Probes for *In Vivo* Gene Detection

Gang Bao, Phillip Santangelo, Nitin Nitin, and Won Jong Rhee

6.1
Introduction

The ability to image specific RNAs in living cells in real time can provide essential information on RNA synthesis, processing, transport and localization, as well as on the dynamics of RNA expression and localization in response to external stimuli. Such an ability will also offer unprecedented opportunities for advancement in molecular biology, disease pathophysiology, drug discovery and medical diagnostics. Over the past decade or so, an increasing amount of evidence has come to light suggesting that RNA molecules have a wide range of functions in living cells, from physically conveying and interpreting genetic information, to essential catalytic roles, to providing structural support for molecular machines, and to gene silencing. These functions are realized through control of the expression level and stability, both temporally and spatially, of specific RNAs in a cell. Therefore, determining the dynamics and localization of RNA molecules in living cells will significantly impact on the molecular biology and medicine.

Many *in vitro* methods have been developed to provide a relative (mostly semi-quantitative) measure of gene expression level within a cell population, by using purified DNA or RNA obtained from cell lysates. These methods include the polymerase chain reaction (PCR) [1], Northern hybridization (or Northern blotting) [2], expressed sequence tag (EST) [3], serial analysis of gene expression (SAGE) [4], differential display [5] and DNA microarrays [6]. These technologies, combined with the rapidly increasing availability of genomic data for numerous biological entities, present exciting possibilities for the understanding of human health and disease. For example, pathogenic and carcinogenic sequences are increasingly being used as clinical markers for diseased states. However, the use of *in vitro* methods to detect and identify foreign or mutated nucleic acids is often difficult in a clinical setting, due to the low abundance of diseased cells in blood, sputum and stool samples. Further, these methods cannot reveal the spatial and temporal variation of RNA within a single cell.

Nanotechnology, Volume 5: Nanomedicine. Edited by Viola Vogel

ISBN: 978-3-527-31736-3

Labeled linear oligonucleotide (ODN) probes have been used to study intracellular mRNA via *in situ* hybridization (ISH) [7], in which cells are fixed and permeabilized to increase the probe delivery efficiency. Unbound probes are removed by washing to reduce the background and achieve specificity [8]. In order to enhance the signal level, multiple probes targeting the same mRNA can be used [7], although fixation agents and other supporting chemicals can have a considerable effect on the signal level [9] and possibly also on the integrity of certain organelles, such as mitochondria. Thus, the fixation of cells (by using either crosslinking or denaturing agents) and the use of proteases in ISH assays may prevent an accurate description of intracellular mRNA localization from being obtained. It is also difficult to obtain a dynamic picture of gene expression in cells using ISH methods.

Of particular interest is the fluorescence imaging of specific messenger RNAs (mRNAs) – in terms of both their expression level and subcellular localization – in living cells. As shown schematically in Figure 6.1, for eukaryotic cells a pre-mRNA molecule is synthesized in the cell nucleus. After processing (including splicing and polyadenylation), the mature mRNAs are transported from the cell nucleus to the cytoplasm, and often are localized at specific sites. The mRNAs are then translated by

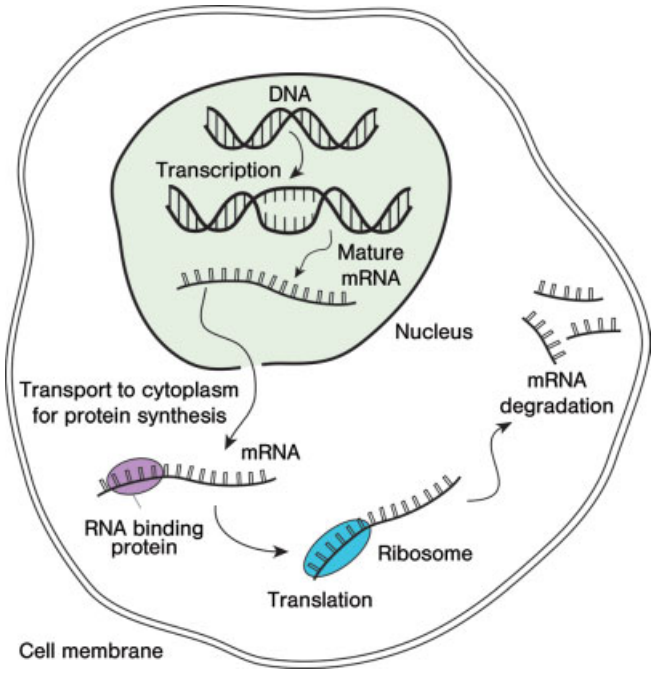


Figure 6.1 The mRNA life cycle. Messenger RNA (mRNA) encoding the chemical 'blueprint' for a protein is synthesized (transcribed) from a DNA template, and the pre-mRNA is processed (spliced) to produce a mature mRNA; this is then transported to specific locations in the cell cytoplasm. The coding information carried by mRNA is used by the ribosomes to produce proteins (translation). After a certain time the message is degraded. mRNAs are almost always complexed with RNA-binding proteins to form ribonucleoprotein (RNP) molecules.

ribosomes to produce specific proteins, and then degraded by RNases after a certain period of time. The limited lifetime of mRNA enables a cell to alter its protein synthesis rapidly, and in response to its changing needs. During the entire life cycle of an mRNA, it is always complexed with RNA-binding proteins to form a ribonucleoprotein (RNP). This has significant implications for the live-cell imaging of mRNAs (as discussed below).

To detect RNA molecules in living cells, with not only high specificity but also high sensitivity and signal-to-background ratio, it is important that the probes recognize RNA targets with high specificity, convert target recognition *directly* into a measurable signal, and differentiate between true and false-positive signals. This is especially important for low-abundance genes and clinical samples containing only a small number of diseased cells. It is also important for the probes to quantify low gene expression levels with great accuracy, and have fast kinetics in tracking alterations in gene expression in real time. For detecting genetic alterations such as mutations, insertions and deletions, the ability to recognize single nucleotide polymorphisms (SNPs) is essential. In order to achieve this optimal performance, it is necessary to have a good understanding of the structure–function relationship of the probes, the probe stability and the RNA target accessibility in living cells. It is also necessary to achieve an efficient cellular delivery of probes, with minimal probe degradation.

In the following sections we will review the fluorescent probes that are most often used for RNA detection, and discuss the critical issues in live-cell RNA detection, including probe design, target accessibility, the cellular delivery of probes, as well as detection sensitivity, specificity and signal-to-background ratio. Emphasis is placed on the design and application of molecular beacons, although some of the issues are common to other oligonucleotide probes.

6.2
Fluorescent Probes for Live-Cell RNA Detection

Several classes of molecular probes have been developed for RNA detection in living cells, including: (i) tagged linear ODN probes; (ii) oligonucleotide hairpin probes; and (iii) probes using fluorescent proteins as reporter. Although probes composed of full-length RNAs (mRNA or nuclear RNA) tagged with a fluorescent or radioactive reporter have been used to study the intracellular localization of RNA [10–12], they are not discussed here as they cannot be used to measure the expression level of specific RNAs in living cells.

6.2.1
Tagged Linear ODN Probes

Single fluorescently labeled linear oligonucleotide probes have been developed for RNA tracking and localization studies in living cells [13–15]. Although these probes may recognize specific endogenous RNA transcripts in living cells via Watson–Crick base pairing, and thus reveal subcellular RNA localization, this approach lacks the

ability to distinguish background from true signal, as both bound probes (i.e. those hybridized to RNA target) and unbound probes give a fluorescence signal. Such an approach might also lack detection specificity, as a partial match between the probe and target sequences could induce probe hybridization to RNA molecules of multiple genes. A novel way to increase the signal-to-noise ratio (SNR) and improve detection specificity is to use two linear probes with a fluorescence resonance energy transfer (FRET) pair of (donor and acceptor) fluorophores [13]. However, the dual-linear probe approach may still have a high background signal due to direct excitation of the acceptor and emission detection of the donor fluorescence. Further, it is difficult for linear probes to distinguish targets that differ by a few bases as the difference in free energy of the two hybrids (with and without mismatch) is typically rather small. This limits the application of linear ODN probes in biological and disease studies.

6.2.2 ODN Hairpin Probes

Hairpin nucleic acid probes have the potential to be highly sensitive and specific in live-cell RNA detection. As shown in Figure 6.2a and b, one class of such probes is that of '*molecular beacons*'; these are dual-labeled oligonucleotide probes with a fluorophore at one end and a quencher at the other end [16]. They are designed to form a stem–loop hairpin structure in the absence of a complementary target, so that the fluorescence of the fluorophore is quenched. Hybridization with the target nucleic acid opens the hairpin and physically separates the fluorophore from quencher, allowing a fluorescence signal to be emitted upon excitation (Figure 6.2b). Under optimal conditions, the fluorescence intensity of molecular beacons can increase more than 200-fold upon binding to their targets [16], and this enables them to function as sensitive probes with a high signal-to-background ratio. The stem–loop hairpin structure provides an adjustable energy penalty for hairpin opening which improves probe specificity [17, 18]. The ability to transduce target recognition *directly* into a fluorescence signal with a high signal-to-background ratio, coupled with an improved specificity, has allowed molecular beacons to enjoy a wide range of biological and biomedical applications. These include multiple analyte detection, real-time enzymatic cleavage assaying, cancer cell detection, real-time monitoring of PCR, genotyping and mutation detection, viral infection studies and mRNA detection in living cells [14, 19–32].

As illustrated in Figure 6.2a, a *conventional molecular beacon* has four essential components: loop, stem, fluorophore (dye) and quencher. The loop usually consists of 15–25 nucleotides and is selected to have a unique target sequence and proper melting temperature. The stem, which is formed by two complementary short-arm sequences, is typically four to six bases long and chosen to be independent of the target sequence (Figure 6.2a).

A novel design of hairpin probes is the *wavelength-shifting molecular beacon*, which can fluoresce in a variety of different colors [33]. As shown in Figure 6.2c, in this design, a molecular beacon contains two fluorophores (dyes): a first fluorophore that absorbs

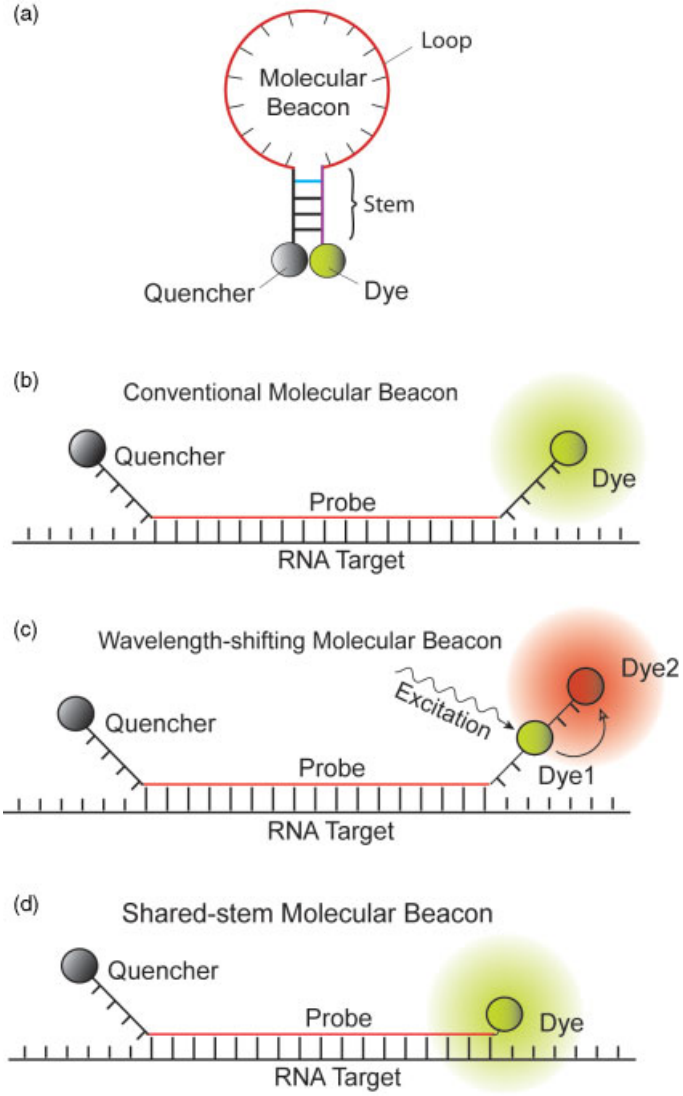


Figure 6.2 Illustrations of molecular beacon designs. (a) Molecular beacons are stem–loop hairpin oligonucleotide probes labeled with a reporter fluorophore at one end and a quencher molecule at the other end; (b) Conventional molecular beacons are designed such that the short complementary arms of the stem are independent of the target sequence; (c) Wavelength-shifting molecular beacons contain two fluorophores: one absorbs in the wavelength range of the monochromatic light source, and the other emits light at the desired emission wavelength due to FRET; (d) Shared-stem molecular beacons are designed such that one arm of the stem participates in both stem formation and target hybridization.

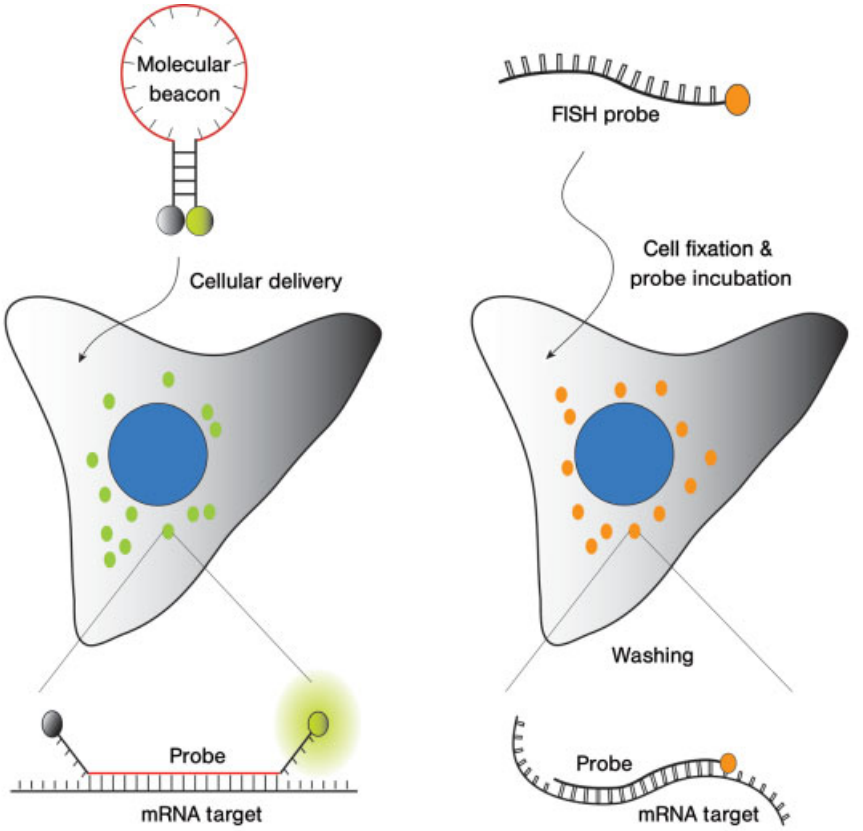
strongly in the wavelength range of the monochromatic light source, and a second fluorophore that emits at the desired emission wavelength due to fluorescence resonance energy transfer from the first fluorophore to the second fluorophore. It has been shown that wavelength-shifting molecular beacons are substantially brighter than conventional molecular beacons, which contain a fluorophore that cannot efficiently absorb energy from the available monochromatic light source.

One major advantage of the stem–loop hairpin probes is that they can recognize their targets with higher specificity than can linear ODN probes. The results of solution studies [17, 18] have suggested that, by using molecular beacons it is possible to discriminate between targets that differ by a single nucleotide. In contrast to current techniques for detecting SNPs – which are often labor-intensive and time-consuming – molecular beacons may provide a simple and promising tool for detecting SNPs in disease diagnosis.

The basic features of molecular beacon versus fluorescence *in situ* hybridization (FISH) are compared in Figure 6.3. Specifically, molecular beacons are dual-labeled hairpin probes of 15–25 nt, while FISH probes are dye-labeled linear oligonucleotides of 40–50 nt. The molecular beacon-based approach has the advantage of detecting RNA in live cells, without the need for cell fixation and washing. However, it does requires the cellular delivery of probes and has a low target accessibility (this is discussed below). The advantage of FISH assays is the ease of probe design due to a better target accessibility. Although FISH assays can be used to image the localization of mRNA in fixed cells, they rely on stringent washing to achieve signal specificity, and do not have the ability to image the dynamics of gene expression in living cells.

In the conventional molecular beacon design, the stem sequence is typically independent of the target sequence (see Figure 6.2b), although sometimes two end bases of the probe sequence, each adjacent to one arm sequence of the stem, could be complementary with each other, thus forming part of the stem (the light blue base of the stem shown in Figure 6.2a). Molecular beacons can also be designed such that all the bases of one arm of the stem (to which a fluorophore is conjugated) are complementary to the target sequence, thus participating in both stem formation and target hybridization (shared-stem molecular beacons) [34] (Figure 6.2d). The advantage of this shared-stem design is to help fix the position of the fluorophore that is attached to the stem arm, limiting its degree-of-freedom of motion, and increasing the FRET in the dual-FRET molecular beacon design (as discussed below).

A dual-FRET molecular beacon approach was developed [26–28] to overcome the difficult that, in live-cell RNA detection, molecular beacons are often degraded by nucleases or open due to nonspecific interaction with hairpin-binding proteins, causing a significant amount of false-positive signal. In this dual-probe design, a pair of molecular beacons labeled with a donor and an acceptor fluorophore, respectively are employed (Figure 6.4). The probe sequences are chosen such that this pair of molecular beacons hybridizes to adjacent regions on a single RNA target (Figure 6.4). As FRET is very sensitive to the distance between donor and acceptor fluorophores, and typically occurs when the donor and acceptor fluorophores are



Molecular beacon	FISH
Hairpin probe of 15–25 nt	Linear probe of 40–50 nt
Live cell assay	With fixed cells
Low target accessibility	Good target accessibility
Requires cellular delivery	Requires cell fixation & probe incubation

Figure 6.3 Comparison of molecular beacon and FISH approaches.

within ~10 nm, the FRET signal is generated by the donor and acceptor beacons only if both probes are bound to the same RNA target. Thus, the sensitized emission of the acceptor fluorophore upon donor excitation serves as a positive signal in the FRET-based detection assay; this can be differentiated from non-FRET false-positive signals due to probe degradation and nonspecific probe opening. This approach combines the low background signal and high specificity of molecular

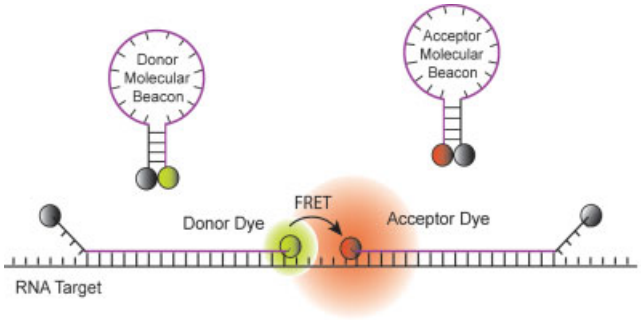


Figure 6.4 A schematic showing the concept of dual-FRET molecular beacons. Hybridization of the donor and acceptor molecular beacons to adjacent regions on the same mRNA target results in FRET between donor and acceptor fluorophores upon donor excitation. By detecting the FRET signal, fluorescence signals due to probe/target binding can be readily distinguished from that due to molecular beacon degradation and nonspecific interactions.

beacons with the ability of FRET assays in differentiating between true target recognition and false-positive signals, leading to an enhanced ability to quantify RNA expression in living cells [28].

6.2.3 Fluorescent Protein-Based Probes

In addition to oligonucleotide probes, tagged RNA-binding proteins such as those with green fluorescent protein (GFP) tags have been used to detect mRNA in live cells [35]. One limitation here is that it requires the identification of a unique protein, which only binds to the specific mRNA of interest. To address this issue, a coat protein of the RNA bacteriophage MS2 was tagged with GFP, after which a RNA sequence corresponding to several MS2 binding sites was introduced to the mRNA of interest. This allowed for the specific targeting of the *nanos* mRNA in live *Drosophila* eggs [36]. The GFP-MS2 approach has been used to track the localization and dynamics of RNA in living cells with single-molecule sensitivity [37, 38]. However, as unbound GFP-tagged MS2 proteins also produce a fluorescence signal, the background signal in the GFP-MS2 approach could be high, leading to a low signal-to-background ratio in live-cell imaging of RNA.

An interesting fluorescent protein-based approach that overcomes this problem is to utilize the fluorescent protein complementation [39, 40]. In this method (split-GFP), a RNA-binding protein is dissected into two fragments, which are respectively fused to the split fragments of a fluorescent protein. Binding of the two tagged fragments of the RNA-binding protein to adjacent sites on the same mRNA molecule (or two parts of an aptamer sequence inserted to the mRNA sequence) brings the two halves of the fluorescent protein together, thus reconstituting the fluorescent protein and restoring fluorescence [40]. Alternatively, two RNA-binding proteins that bind specifically to adjacent sites on the same mRNA molecule can be

tagged with the split fragments of a fluorescent protein, such that their binding to the target mRNA results in the restoration of fluorescence [39]. The advantage of this novel approach is that the background signal is low; there is no fluorescence signal unless the RNA-binding proteins (or protein fragments) are bound to the target mRNA. The split-GFP method, however, may have difficulties in tracking the dynamics of RNA expression in real time, as reconstitution of the fluorescent protein from the split fragments typically takes 2–4 h, during which time the RNA expression level may change. Transfection efficiency may also be a major concern in the GFP-based approaches, in that usually only a small percentage of the cells express the fluorescent proteins following transfection. This limits the application of the split-GFP methods in detecting diseased cells using mRNA as a biomarker for the disease.

6.3 Probe Design and Structure–Function Relationships

6.3.1 Target Specificity

There are three major design issues of molecular beacons: probe sequence; hairpin structure; and fluorophore/quencher selection. In general, the probe sequence is selected to ensure specificity, and to have good target accessibility. The hairpin structure, as well as the probe and stem sequences, are determined to have the proper melting temperature, while the fluorophore–quencher pair should produce a high signal-to-background ratio. To ensure specificity, for each gene to target, it is possible to use the NCBI BLAST [41] or similar software to select multiple target sequences of 15–25 bases that are unique to the target RNA. As the melting temperature of the molecular beacons affects both the signal-to-background ratio and detection specificity (especially for mutation detection), it is often necessary to select the target sequence with a balanced G-C content, and to adjust the loop and stem lengths and the stem sequence of the molecular beacon to realize the optimal melting temperature. In particular, it is necessary to understand the effect of molecular beacon design on melting temperature so that, at 37 °C, single-base mismatches in target mRNAs can be differentiated. This is also a general issue for detection specificity in that, for any specific probe sequence selected, there might be multiple genes in the mammalian genome that have sequences which differ from the probe sequence by only a few bases. Therefore, it is important to design the molecular beacons so that only the specific target RNA would produce a strong signal.

Several approaches can be taken to validate the signal specificity. For example, one could either upregulate or downregulate the expression level of a specific RNA, quantify the level using RT-PCR, and then compare the PCR result with that of molecular beacon-based imaging of the same RNA in living cells. However, complications may arise when the approach used to change the RNA expression level in living cells has an effect on multiple genes, as this would lead to some ambiguity, even

when the PCR and beacon results match. It is possible that the best way to down-regulate the level of a specific mRNA in live cells is to use small interfering RNA (siRNA) treatment, which typically leads to a >80% reduction of the specific mRNA level. As the effect of siRNA treatment varies depending on the specific probe used, the siRNA delivery method, cell type and optimization of the protocol (i.e. probe design and delivery method/conditions) is often needed.

6.3.2
Molecular Beacon Structure–Function Relationships

The loop, stem lengths and sequences are critical design parameters for molecular beacons, since at any given temperature they largely control the fraction of molecular beacons that are bound to the target [17, 18]. In many applications, the choices of the probe sequence are limited by target-specific considerations, such as the sequence surrounding a single nucleotide polymorphism (SNP) of interest. However, the probe and stem lengths, and stem sequence, can be adjusted to optimize the performance (i.e. specificity, hybridization rate and signal-to-background ratio) of a molecular beacon for a specific application [17, 34].

In order to demonstrate the effect of molecular beacon structure on its melting behavior, the melting temperature for molecular beacons with various stem–loop structures is displayed in Figure 6.5a. In general, the melting temperature was found to increase with probe length, but appeared to plateau at a length of ~20 nucleotides. The stem length of the molecular beacon was also found to have a major influence on the melting temperature of the molecular beacon–target duplexes.

While both the stability of the hairpin probe and its ability to discriminate targets over a wider range of temperatures increase with increasing stem length, it is accompanied by a decrease in the hybridization on-rate constant (see Figure 6.5b).

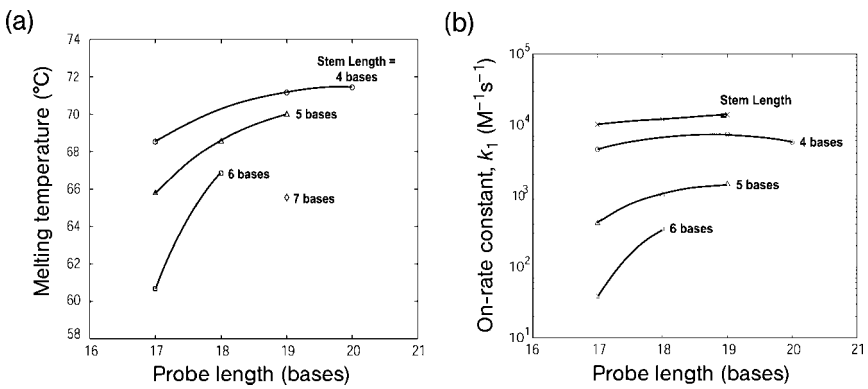


Figure 6.5 Structure–function relationships of molecular beacons. (a) Melting temperatures for molecular beacons with different structures in the presence of target; (b) The rate constant of hybridization k_1 (on-rate constant) for molecular beacons with various probe and stem lengths hybridized to their complementary targets.

For example, molecular beacons with a four-base stem had an on-rate constant up to 100-fold higher than did molecular beacons with a six-base stem. Changing the probe length of a molecular beacon may also influence the rate of hybridization, as shown in Figure 6.5b.

The results of thermodynamic and kinetic studies showed that, if the stem length was too large then it would be difficult for the beacon to open on hybridization. But, if the stem length was too small, then a large fraction of beacons might open due to the thermal force. Likewise, and relative to the stem length, whilst a longer probe might lead to a lower dissociation constant, it might also reduce the specificity, as the relative free energy change due to a one base mismatch would be smaller. A long probe length may also lead to coiled conformations of the beacons, resulting in reduced kinetic rates. Consequently, the stem and probe lengths must be carefully chosen in order to optimize both hybridization kinetics and molecular beacon specificity [17, 34]. In general, molecular beacons with longer stem lengths have an improved ability to discriminate between wild-type and mutant targets in solution, over a broader range of temperatures. This effect can be attributed to the enhanced stability of the molecular beacon stem–loop structure and the resulting smaller free energy difference between closed (unbound) molecular beacons and molecular beacon–target duplexes, which generates a condition where a single-base mismatch reduces the energetic preference of probe–target binding. Longer stem lengths, however, are accompanied by a reduced probe–target hybridization kinetic rate. On a similar note, molecular beacons with short stems have faster hybridization kinetics but suffer from lower signal-to-background ratios compared to molecular beacons with longer stems.

6.3.3
Target Accessibility

One critical issue in molecular beacon design is target accessibility, as is the case for most oligonucleotide probes for live-cell RNA detection. It is well known that a functional mRNA molecule in a living cell is always associated with RNA-binding proteins, thus forming a RNP. An mRNA molecule also often has double-stranded portions and forms secondary (folded) structures (Figure 6.6). Therefore, when designing a molecular beacon it is necessary to avoid targeting mRNA sequences that are double-stranded, or occupied by RNA-binding proteins, for otherwise the probe will have to penetrate into the RNA double strand or compete with the RNA-binding protein in order to hybridize to the target. In fact, molecular beacons designed to target a specific mRNA often show no signal when delivered to living cells. One difficulty in molecular beacon design is that, although predictions of mRNA secondary structure can be made using software such as *Beacon Designer* (www.premierbiosoft.com) and *mfold* (http://www.bioinfo.rpi.edu/applications/mfold/old/dna/), they may be inaccurate due to limitations of the biophysical models used, and the limited understanding of protein–RNA interaction. Therefore, for each gene to be targeted it may be necessary to select multiple unique sequences along the target RNA, and then to design, synthesize and test the corresponding molecular beacons in living cells in order to select the best target sequence.

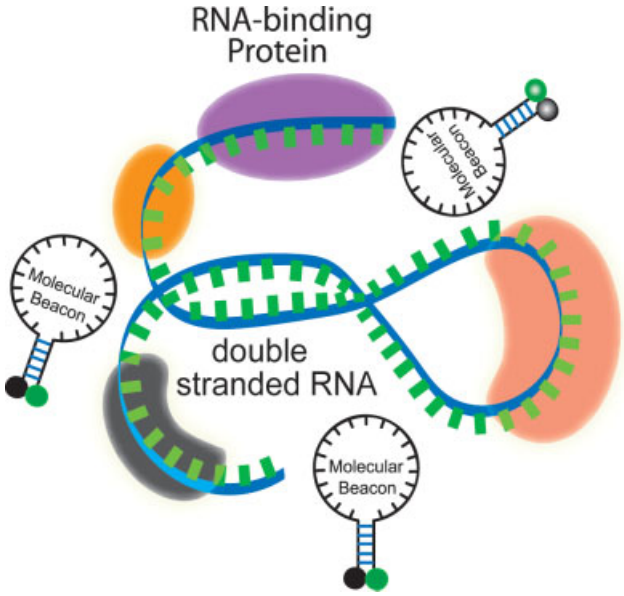


Figure 6.6 A schematic illustration of a segment of the target mRNA with a double-stranded portion and RNA-binding proteins. A molecular beacon must penetrate into the mRNA double strand or compete with the RNA-binding protein(s) in order to hybridize to the target.

In aiming to reveal the possible molecular beacon design rules, the accessibility of BMP-4 mRNA was studied using different beacon designs [42]. Specifically, molecular beacons were designed to target the start codon and termination codon regions, the siRNA and anti-sense oligonucleotide probe sites (which were identified previously) and also the sites that were chosen at random. All of the target sequences are unique to BMP-4 mRNA. Of the eight molecular beacons designed to target BMP-4 mRNA, only two were found to produce a strong signal: one which targeted the start codon region, and one which targeted the termination codon region. It was also found that, even for a molecular beacon which functioned well, shifting its targeting sequence by only a few bases towards the 3′ or 5′ ends caused a significant reduction in the fluorescence signal from beacons in a live-cell assay. This indicated that the target accessibility was quite sensitive to the location of the targeting sequence. These results, together with molecular beacons validated previously, suggest that the start and termination codon regions and the exon–exon junctions are more accessible than other locations in an mRNA.

6.3.4
Fluorophores and Quenchers

With a correct backbone synthesis and fluorophore/quencher conjugation, a molecular beacon can – in theory – be labeled with any desired reporter–quencher pair.

However, the correct selection of the reporter and quencher could also improve the signal-to-background ratio and multiplexing capabilities. The selection of a fluorophore label for a molecular beacon as reporter is normally less critical than for the hairpin probe design, as many conventional dyes can yield satisfactory results. However, the correct selection may yield additional benefits such as an improved signal-to-background ratio and multiplexing capabilities. As each molecular beacon utilizes only one fluorophore, it is possible to use multiple molecular beacons in the same assay, assuming that the fluorophores are chosen with minimal emission overlap [19]. Molecular beacons can even be labeled simultaneously with two fluorophores – that is with 'wavelength shifting' reporter dyes (see Figure 6.2c), allowing multiple reporter dye sets to be excited by the same monochromatic light source but to fluoresce in a variety of colors [33]. Clearly, multicolor fluorescence detection of different beacon/target duplexes may in time become a powerful tool for the simultaneous detection of multiple genes.

For dual-FRET molecular beacons (see Figure 6.4), the donor fluorophores typically emit at shorter wavelengths compared with the acceptor. Energy transfer then occurs as a result of long-range dipole–dipole interactions between the donor and acceptor. The efficiency of such energy transfer depends on the extent of the spectral overlap of the emission spectrum of the donor with the absorption spectrum of the acceptor, the quantum yield of the donor, the relative orientation of the donor and acceptor transition dipoles [43], and the distance between the donor and acceptor molecules (usually four to five bases). In selecting the donor and acceptor fluorophores so as to create a high signal-to-background ratio, it is important to optimize the above parameters, and to avoid direct excitation of the acceptor fluorophore at the donor excitation wavelength. It is also important to minimize donor emission detection at the acceptor emission detection wavelength. Examples of FRET dye pairs include Cy3 (donor) with Cy5 (acceptor), TMR (donor) with Texas Red (acceptor), and fluorescein (FAM) (donor) with Cy3 (acceptor).

By contrast, it is relatively straightforward to select the quencher molecules. Organic quencher molecules such as dabcyl, BHQ2 (blackhole quencher II) (Biosearch Tech), BHQ3 (Biosearch Tech) and Iowa Black (IDT) can all effectively quench a wide range of fluorophores by both FRET and the formation of an exciton complex between the fluorophore and the quencher [44].

6.4
Cellular Delivery of Nanoprobes

One of the most critical aspects of measuring the intracellular level of RNA molecules using synthetic probes is the ability to deliver the probes into cells via the plasma membrane, which itself is quite lipophilic and restricts the transport of large, charged molecules. Thus, the plasma membrane serves as a very robust barrier to polyanionic molecules such as hairpin oligonucleotides. Further, even if the probes enter the cells successfully, the efficiency of delivery in an imaging assay should be defined not only by how many probes enter the cell, or how many cells have probes internalized,

but also by how many probes remain functioning inside the cells. This is a different situation from both antisense and gene delivery applications, where the reduction in level of protein expression is the final metric used to define efficiency or success. For measuring RNA molecules (including mRNA and rRNA) in the cytoplasm, a large amount of the probe should remain in the cytoplasm.

Existing cellular delivery techniques can be divided into two categories, namely *endocytic* and *nonendocytic*. Endocytic delivery typically employs cationic and polycationic molecules such as liposomes and dendrimers, whereas nonendocytic methods include microinjection and the use of cell-penetrating peptides (CPPs) or streptolysin *O* (SLO). Probe delivery via the endocytic pathway typically takes 2–4 h. It has been reported that ODN probes internalized via endocytosis are predominantly trapped inside endosomes and lysosomes, where they are degraded by the action of cytoplasmic nucleases [45]. Consequently, only 0.01% to 10% of the probes remain functioning after having escaped from endosomes and lysosomes [46].

Oligonucleotide probes (including molecular beacons) have been delivered into cells via microinjection [47]. In most cases, the ODNs were accumulated rapidly in the cell nucleus and prevented the probes from targeting mRNAs in the cell cytoplasm. The depletion of intracellular ATP or lowering the temperature from 37 to 4 °C did not have any significant effect on ODN nuclear accumulation, thus ruling out any active, motor protein-driven transport [47]. It is unclear if the rapid transport of ODN probes to the nucleus is due to electrostatic interaction, or is driven by a microinjection-induced flow, or the triggering of a signaling pathway. There is no fundamental biological reason why ODN probes should accumulate in the cell nucleus, but to prevent such accumulation streptavidin (60 kDa) molecules were conjugated to linear ODN probes via biotin [13]. After being microinjected into the cells, the dual-FRET linear probes could hybridize to the same mRNA target in the cytoplasm, resulting in a FRET signal. More recently, it was shown that when transfer RNA (tRNA) transcripts were attached to molecular beacons with a 2′-*O*-methyl backbone and injected into the nucleus of HeLa cells, the probes were exported into the cytoplasm. Yet, when these constructs were introduced into the cytoplasm, they remained cytoplasmic [48]. However, even without the problem of unwanted nuclear accumulation, microinjection is an inefficient process for delivering probes into a large number of cells.

Another nonendocytic delivery method is that of *toxin-based cell membrane permeabilization*. For example, SLO is a pore-forming bacterial toxin that has been used as a simple and rapid means of introducing oligonucleotides into eukaryotic cells [49–51]. SLO binds as a monomer to cholesterol and oligomerizes into a ring-shaped structure to form pores of approximately 25–30 nm in diameter, allowing the influx of both ions and macromolecules. It was found that SLO-based permeabilization could achieve an intracellular concentration of ODNs which was approximately 10-fold that achieved with electroporation or liposomal-based delivery. As cholesterol composition varies with cell type, however, the permeabilization protocol must be optimized for each cell type by varying the temperature, incubation time, cell number and SLO concentration. One essential feature of toxin-based permeabilization is that it is reversible. This can be achieved by introducing oligonucleotides with SLO under

serum-free conditions and then removing the mixture and adding normal media with the serum [50, 52].

Cell-penetrating peptides have also been used to introduce proteins, nucleic acids and other biomolecules into living cells [53–55]. Included among the family of peptides with membrane-translocating activity are antennapedia, HSV-1 VP22 and the HIV-1 Tat peptide. To date, the most widely used peptides are HIV-1 Tat peptide and its derivatives, due to their small sizes and high delivery efficiencies. The Tat peptide is rich in cationic amino acids (especially arginine, which is very common in many CPPs); however, the exact mechanism of CPP-induced membrane translocation remains elusive.

A wide variety of cargos have been delivered to living cells, both in cell culture and in tissues, using CPPs [56, 57]. For example, Allinquant *et al.* [58] linked the antennapedia peptide to the 5′ end of DNA oligonucleotides (with biotin on the 3′ end) and incubated both peptide-linked ODNs and ODNs alone (as control) with cells. By detecting biotin via a streptavidin–alkaline phosphatase amplification, the peptide-linked ODNs were shown to be internalized very efficiently into all cell compartments compared to control ODNs. Moreover, no indication of endocytosis was found. Similar results were obtained by Troy *et al.* [59], with a 100-fold increase in antisense delivery efficiency when the ODNs were linked to antennapedia peptides. Recently, Tat peptides were conjugated to molecular beacons using different linkages (Figure 6.7); the resultant peptide-linked molecular beacons were delivered into living cells to target glyceraldehyde phosphate dehydrogenase (GAPDH) and survivin mRNAs [29]. It was shown that, at relatively low concentrations, peptide-linked molecular beacons were internalized into living cells within 30 min, with near-100% efficiency. Further, peptide-based delivery did not interfere with either specific targeting by, or hybridization-induced florescence of, the probes. In addition, the peptide-linked molecular beacons were seen to possess self-delivery, targeting and reporting functions. In contrast, the liposome-based (Oligofectamine) or dendrimer-based (Superfect) delivery of molecular beacons required 3–4 h and resulted in a punctate fluorescence signal in the cytoplasmic vesicles and a high background in

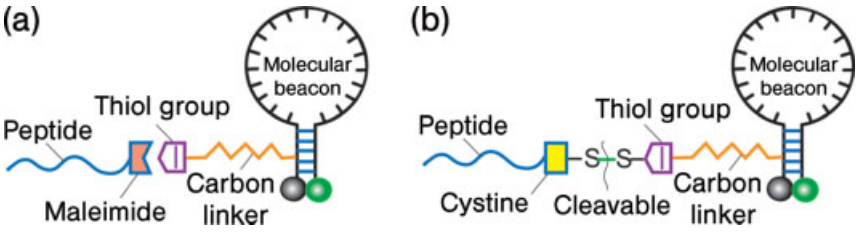


Figure 6.7 A schematic of peptide-linked molecular beacons. (a) A peptide-linked molecular beacon using the thiol–maleimide linkage in which the quencher arm of the molecular beacon stem is modified by adding a thiol group which can react with a maleimide group placed to the C terminus of the peptide to form a direct, stable linkage; (b) A peptide-linked molecular beacon with a cleavable disulfide bridge in which the peptide is modified by adding a cysteine residue at the C terminus; the cysteine then forms a disulfide bridge with the thiol-modified molecular beacon. This disulfide bridge design allows the peptide to be cleaved from the molecular beacon by the reducing environment of the cytoplasm.

both the cytoplasm and nucleus of cells [29]. These results showed clearly that the cellular delivery of molecular beacons using a peptide-based approach is far more effective than conventional transfection methods.

6.5 Living Cell RNA Detection Using Nanostructured Probes

Sensitive gene detection in living cells presents a significant challenge. In addition to issues of target accessibility, detection specificity and probe delivery (as discussed above), the achievement of a high detection sensitivity and a high signal-to-background ratio requires not only careful design of the probes and advanced fluorescence microscopy imaging, but also a better understanding of RNA biology and probe–target interactions. It is likely that different applications have different requirements on the properties of probes. For example, the rapid determination of RNA expression level and localization requires fast probe/target hybridization kinetics, whereas the long-term monitoring of gene expression dynamics requires probes with a high intracellular stability.

To demonstrate the capability of molecular beacons in the sensitive detection of specific endogenous mRNAs in living cells, dual-FRET molecular beacons were designed to detect K-ras and survivin mRNAs in HDF and MIAPaCa-2 cells, respectively [28]. K-*ras* is one of the most frequently mutated genes in human cancers [60]. A member of the G-protein family, K-ras is involved in transducing growth-promoting signals from the cell surface. Survivin, one of the inhibitor of apoptosis proteins (IAPs), is normally expressed during fetal development but not in most normal adult tissues [61], and thus can be used as a tumor biomarker for several types of cancer. Each FRET probe pair consisted of two molecular beacons – one labeled with a donor fluorophore (Cy3, donor beacon) and a second labeled with an acceptor fluorophore (Cy5, acceptor beacon). These molecular beacons were designed to hybridize to adjacent regions on an mRNA target so that the two fluorophores lay within the FRET range (~6 nm) when probe/target hybridization occurred for both beacons. BHQ-2 and BHQ-3 were used as quenchers for the donor and acceptor molecular beacons, respectively. One pair of molecular beacons targets a segment of the wild-type K-*ras* gene, the codon 12 mutations of which are involved in the pathogenesis of many cancers. A negative control dual-FRET molecular beacon pair was also designed ('random beacon pair'), the specific 16-base target sequence of which was selected using random walking, and thus had no exact match in the mammalian genome. It was found that detection of the FRET signal significantly reduced false-positives, leading to sensitive imaging of K-*ras* and survivin mRNAs in live HDF and MIAPaCa-2 cells. For example, FRET detection gave a ratio of 2.25 of K-*ras* mRNA expression in stimulated versus unstimulated HDF cells, which was comparable to a ratio of 1.95 using RT-PCR but contrasted to the single-beacon result of 1.2. The detection of survivin mRNA also indicated that, compared to the single-beacon approach, dual-FRET molecular beacons gave a lower background signal, which in turn led to a higher signal-to-background ratio [28].

6.5.1 Biological Significance

An intriguing discovery in detecting K-*ras* and survivin mRNAs using dual-FRET molecular beacons is the clear and detailed mRNA localization in living cells [28]. To demonstrate this point, a fluorescence image of K-*ras* mRNA in stimulated HDF cells is shown in Figure 6.8a, indicating an intriguing filamentous localization pattern. The localization pattern of K-*ras* mRNA was further studied and found to be colocalized with mitochondria inside live HDF cells [62]. As K-ras proteins interact with proteins such as Bcl-2 in the mitochondria to mediate both anti-apoptotic and pro-apoptotic pathways, it seems that cells localize certain mRNAs where the corresponding proteins can easily bind to their partners.

The survivin mRNA, however, is localized in MIAPaCa-2 cell very differently. As shown in Figure 6.8b, in which the fluorescence image was superimposed with a white-light image of the cells, the survivin mRNAs seemed to localize in a nonsymmetrical pattern within MIAPaCa-2 cells, often to one side of the nucleus of the cell. These mRNA localization patterns raise many interesting biological questions. For example, how are mRNAs transported to their destination, and how is the destination recognized? Also, to which subcellular organelle might the mRNAs be colocalized? And what is the biological implication of mRNA localization? Although mRNA localization in living cells is believed to be closely related to the post-transcriptional regulation of gene expression, much remains to be seen if such localization indeed targets a protein to its site of function by producing the protein 'right on the spot'.

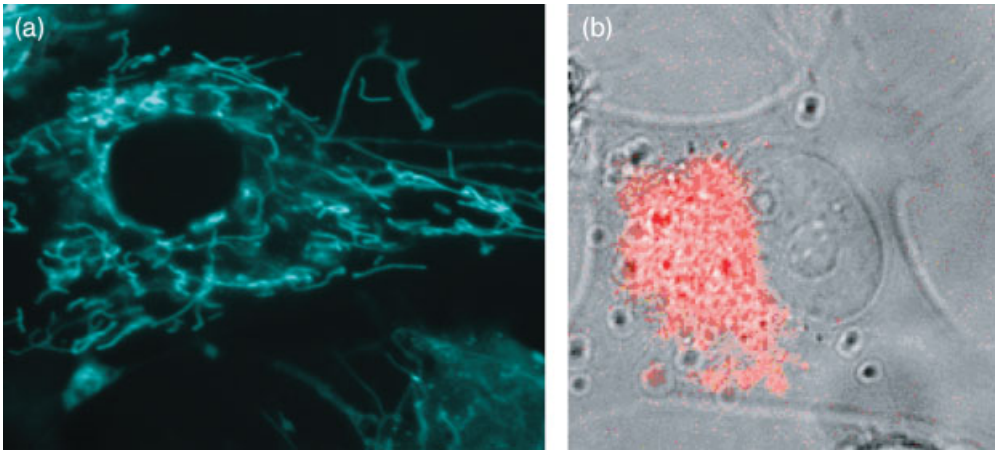


Figure 6.8 mRNA localization in HDF and MIAPaCa-2 cells. (a) Fluorescence images of K-*ras* mRNA in stimulated HDF cells. Note the filamentous K-*ras* mRNA localization pattern; (b) A fluorescence image of survivin mRNA localization in MIAPaCa-2 cells. Note that survivin mRNAs are often localized to one side of the nucleus of the MIAPaCa-2 cells.

The transport and localization of oskar mRNA in *Drosophila melanogaster* oocytes has also been visualized [26]. In these studies, molecular beacons with a 2′-*O*-methyl backbone were delivered into cells using microinjection, and the migration of oskar mRNA was tracked in real time, from the nurse cells where it is produced to the posterior cortex of the oocyte where it is localized. Clearly, the direct visualization of specific mRNAs in living cells with molecular beacons will provide important insights into the intracellular trafficking and localization of RNA molecules.

As another example of targeting specific genes in living cells, molecular beacons were used to detect the viral genome and characterize the spreading of bovine respiratory syncytial virus (bRSV) in living cells [63]. It was found that a molecular beacon signal could be detected in single living cells infected by bRSV with high detection sensitivity, and the signal revealed a connected, highly three-dimensional, amorphous inclusion-body structure not seen in fixed cells. Figure 6.9 shows the molecular beacon signal indicating the spreading of viral infection at days 1, 3, 5 and

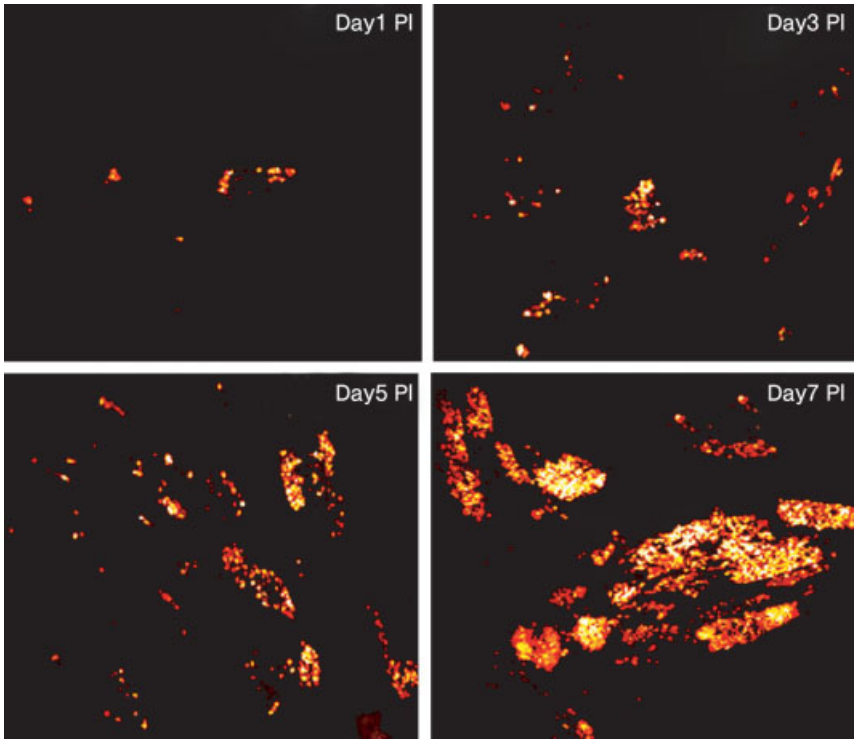


Figure 6.9 Live-cell fluorescence imaging of the genome of bovine respiratory syncytial virus (bRSV) using molecular beacons, showing the spreading of infection in host cells at days 1, 3, 5 and 7 post-infection (PI). Primary bovine turbinate cells were infected by a clinical isolate of bRSV, CA-1, with a viral titer of $2 \times 10^{3.6}\,TCID_{50}\,ml^{-1}$. Molecular beacons were designed to target several repeated sequences of the gene-end-intergenic-gene-start signal within the bRSV genome, with a SNR of 50–200.

7 post-infection, and demonstrates the ability of molecular beacons to monitor and quantify – in real time – the viral infection process. Molecular beacons were also used to image the viral genomic RNA (vRNA) of human RSV (hRSV) in live Vero cells, revealing the dynamics of filamentous virion egress, and providing an insight as to how viral filaments bud from the plasma membrane of the host cell [64].

6.6 Engineering Challenges in New Probe Development

Nanostructured molecular probes such as molecular beacons have the potential to enjoy a wide range of applications that require the sensitive detection of genomic sequences. For example, molecular beacons can be used as a tool for the detection of single-stranded nucleic acids in homogeneous *in vitro* assays [65, 66]. Surface-immobilized molecular beacons used in microarray assays allow for the high-throughput parallel detection of nucleic acid targets, while avoiding the difficulties associated with PCR-based labeling [65, 67]. Another novel application of molecular beacons is the detection of double-stranded DNA targets using PNA 'openers' that form triplexes with the DNA strands [68]. Further, proteins can be detected by synthesizing an 'aptamer molecular beacon' [69, 70] which, upon binding to a protein, undergoes a conformational change that results in the restoration of fluorescence.

The most exciting application of nanostructured oligonucleotide probes, however, is that of living cell *gene detection*. As demonstrated, molecular beacons can detect endogenous mRNA in living cells with high specificity, sensitivity and signal-to-background ratio, and thus have the potential to provide a powerful tool for both laboratory and clinical studies of gene expression *in vivo*. For example, molecular beacons can be used in high-throughput cell-based assays to quantify and monitor the dose-dependent changes of specific mRNA expression in response to different drug leads. The ability of molecular beacons to detect and quantify the expression of specific genes in living cells will also facilitate disease studies, such as viral infection detection and cancer diagnosis.

A number of challenges exist in the detection and quantification of RNA expression in living cells. In addition to the issues of probe design and target accessibility, quantifying gene expression in living cells in terms of mRNA copy-number per cell poses a significant challenge. For example, it is necessary to distinguish between true and background signals, to determine the fraction of mRNA molecules hybridized with probes, and to quantify the possible self-quenching effect of the reporter, especially when mRNA is highly localized. As the fluorescence intensity of the reporter may be altered by the intracellular environment, it is also necessary to create an internal control by, for example, injecting fluorescently labeled oligonucleotides with known quantity into the same cells and obtaining the corresponding fluorescence intensity. Furthermore, unlike RT-PCR studies – where the mRNA expression is averaged over a large number of cells (usually $>10^6$) – in the optical imaging of mRNA expression in living cells only a relatively small number of cells (typically

<1000) are observed. Therefore, the average copy number per cell may change with the total number of cells observed due to the (often large) cell-to-cell variation of mRNA expression.

Another issue in living cell gene detection using hairpin ODN probes is the possible effect of probes on normal cell function, including protein expression. As has been revealed in antisense therapy research, the complementary pairing of a short segment of an exogenous oligonucleotide to mRNA can have a profound impact on protein expression levels, and even cell fate. For example, tight binding of the probe to the translation start site may block mRNA translation. Binding of a DNA probe to mRNA can also trigger RNase H-mediated mRNA degradation. However, the probability of eliciting antisense effects with hairpin probes may be very low when low concentrations of probes (<200 n*M*) are used for mRNA detection, in contrast to the high concentrations (typically 20 μ*M*; [51]) employed in antisense experiments. Further, it generally takes 4 h before any noticeable antisense effect occurs, whereas the visualization of mRNA with hairpin probes requires less than 2 h after delivery. However, it is important to carry out a systematic study of the possible antisense effects, especially for molecular beacons with a 2′-*O*-methyl backbone, which may also trigger unwanted RNA interference.

As a new approach for *in vivo* gene detection, nanostructured probes can be further developed to have an enhanced sensitivity and a wider range of applications. For example, it is likely that hairpin ODN probes with quantum dot as the fluorophore will have a better ability to track the transport of individual mRNAs from the cell nucleus to the cytoplasm. Hairpin ODN probes with a near-infrared (NIR) dye as the reporter, combined with peptide-based delivery, have the potential to detect specific RNAs in tissue samples, animals or even humans. It is also possible to use lanthanide chelate as the donor in a dual-FRET probe assay and to perform time-resolved measurements to dramatically increase the SNR, thus achieving high sensitivity while detecting low-abundance genes. Although very challenging, the development of these and other nanostructured ODN probes will significantly enhance our ability to image, track and quantify gene expression *in vivo*, and provide a powerful tool for basic and clinical studies of human health and disease.

There are many possibilities for nanostructured probes to become clinical tools for disease detection and diagnosis. For example, molecular beacons could be used to perform cell-based early cancer detection using clinical samples such as blood, saliva and other body fluids. In this case, cells in the clinical sample are separated, while the molecular beacons designed to target specific cancer genes are delivered to the cell cytoplasm for detecting mRNAs of the cancer biomarker genes. Cancer cells having a high level of the target mRNAs (e.g. survivin) or mRNAs with specific mutations that cause cancer (e.g. K-*ras* codon 12 mutations) would show high levels of fluorescence signal, whereas normal cells would show just a low background signal. This would allow cancer cells to be distinguished from normal cells. When using this approach, the target mRNAs would not be diluted compared to approaches using a cell lysate, such as PCR. Thus, molecular beacon-based assays have the potential for the positive identification of cancer cells in a clinical sample, with high specificity and sensitivity. It might also be possible to detect cancer cells *in vivo*

by using NIR-dye-labeled molecular beacons in combination with endoscopy. Nanostructured probes could also be used for the cell-based detection of other diseases. As illustrated above, well-designed molecular beacons can rapidly detect viral infection in living cells, with high specificity and sensitivity. Another possibility might be to analyze the vulnerability of atherosclerotic plaques by designing nanostructured probes to image biomarkers (mRNAs or proteins) of vulnerable plaques in blood samples. Although there remain significant challenges, imaging methods using nanostructured probes possess a truly great potential to become a powerful clinical tool for disease detection and diagnosis.

Acknowledgments

These studies were supported by the National Heart Lung and Blood Institute of the NIH as a Program of Excellence in Nanotechnology (HL80711), by the National Cancer Institute of the NIH as a Center of Cancer Nanotechnology Excellence (CA119338), and by the NIH Roadmap Initiative in Nanomedicine through a Nanomedicine Development Center award (PN2EY018244).

References

1 Saiki, R.K., Scharf, S., Faloona, F., Mullis, K.B., Horn, G.T., Erlich, H.A. and Arnheim, N. (1985) *Science*, **230**, 1350.

2 Alwine, J.C., Kemp, D.J., Parker, B.A., Reiser, J., Renart, J., Stark, G.R. and Wahl, G.M. (1979) *Methods in Enzymology*, **68**, 220.

3 Adams, M.D., Dubnick, M., Kerlavage, A.R., Moreno, R., Kelley, J.M., Utterback, T.R., Nagle, J.W., Fields, C. and Venter, J.C. (1992) *Nature*, **355**, 632.

4 Velculescu, V.E., Zhang, L., Vogelstein, B. and Kinzler, K.W. (1995) *Science*, **270**, 484.

5 Liang, P. and Pardee, A.B. (1992) *Science*, **257**, 967.

6 Schena, M., Shalon, D., Davis, R.W. and Brown, P.O. (1995) *Science*, **270**, 467.

7 Bassell, G.J., Powers, C.M., Taneja, K.L. and Singer, R.H. (1994) *The Journal of Cell Biology*, **126**, 863.

8 Buongiorno-Nardelli, M. and Amaldi, F. (1970) *Nature*, **225**, 946.

9 Behrens, S., Fuchs, B.M., Mueller, F. and Amann, R. (2003) *Applied and Environmental Microbiology*, **69**, 4935.

10 Huang, Q. and Pederson, T. (1999) *Nucleic Acids Research*, **27**, 1025.

11 Glotzer, J.B., Saffrich, R., Glotzer, M. and Ephrussi, A. (1997) *Current Biology*, **7**, 326.

12 Jacobson, M.R. and Pederson, T. (1998) *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 7981.

13 Tsuji, A., Koshimoto, H., Sato, Y., Hirano, M., Sei-Iida, Y., Kondo, S. and Ishibashi, K. (2000) *Biophysical Journal*, **78**, 3260.

14 Dirks, R.W., Molenaar, C. and Tanke, H.J. (2001) *Histochemistry and Cell Biology*, **115**, 3.

15 Molenaar, C., Abdulle, A., Gena, A., Tanke, H.J. and Dirks, R.W. (2004) *The Journal of Cell Biology*, **165**, 191.

16 Tyagi, S. and Kramer, F.R. (1996) *Nature Biotechnology*, **14**, 303.

17 Tsourkas, A., Behlke, M.A., Rose, S.D. and Bao, G. (2003) *Nucleic Acids Research*, **31**, 1319.

18 Bonnet, G., Tyagi, S., Libchaber, A. and Kramer, F.R. (1999) *Proceedings of the*

National Academy of Sciences of the United States of America, **96**, 6171.
19 Tyagi, S., Bratu, D.P. and Kramer, F.R. (1998) *Nature Biotechnology*, **16**, 49.
20 Li, J.J., Geyer, R. and Tan, W. (2000) *Nucleic Acids Research*, **28**, E52.
21 Molenaar, C., Marras, S.A., Slats, J.C., Truffert, J.C., Lemaitre, M., Raap, A.K., Dirks, R.W. and Tanke, H.J. (2001) *Nucleic Acids Research*, **29**, E89.
22 Sokol, D.L., Zhang, X., Lu, P. and Gewirtz, A.M. (1998) *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 11538.
23 Vet, J.A., Majithia, A.R., Marras, S.A., Tyagi, S., Dube, S., Poiesz, B.J. and Kramer, F.R. (1999) *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 6394.
24 Kostrikis, L.G., Tyagi, S., Mhlanga, M.M., Ho, D.D. and Kramer, F.R. (1998) *Science*, **279**, 1228.
25 Piatek, A.S., Tyagi, S., Pol, A.C., Telenti, A., Miller, L.P., Kramer, F.R. and Alland, D. (1998) *Nature Biotechnology*, **16**, 359.
26 Bratu, D.P., Cha, B.J., Mhlanga, M.M., Kramer, F.R. and Tyagi, S. (2003) *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 13308.
27 Tsourkas, A., Behlke, M.A., Xu, Y. and Bao, G. (2003) *Analytical Chemistry*, **75**, 3697.
28 Santangelo, P.J., Nix, B., Tsourkas, A. and Bao, G. (2004) *Nucleic Acids Research*, **32**, e57.
29 Nitin, N., Santangelo, P.J., Kim, G., Nie, S. and Bao, G. (2004) *Nucleic Acids Research*, **32**, e58.
30 Tyagi, S. and Alsmadi, O. (2004) *Biophysical Journal*, **87**, 4153.
31 Peng, X.H., Cao, Z.H., Xia, J.T., Carlson, G.W., Lewis, M.M., Wood, W.C. and Yang, L. (2005) *Cancer Research*, **65**, 1909.
32 Medley, C.D., Drake, T.J., Tomasini, J.M., Rogers, R.J. and Tan, W. (2005) *Analytical Chemistry*, **77**, 4713.
33 Tyagi, S., Marras, S.A. and Kramer, F.R. (2000) *Nature Biotechnology*, **18**, 1191.
34 Tsourkas, A., Behlke, M.A. and Bao, G. (2002) *Nucleic Acids Research*, **30**, 4208.
35 Brodsky, A.S. and Silver, P.A. (2002) *Methods (San Diego, Calif.)*, **26**, 151.
36 Forrest, K.M. and Gavis, E.R. (2003) *Current Biology*, **13**, 1159.
37 Shav-Tal, Y., Darzacq, X., Shenoy, S.M., Fusco, D., Janicki, S.M., Spector, D.L. and Singer, R.H. (2004) *Science*, **304**, 1797.
38 Haim, L., Zipor, G., Aronov, S. and Gerst, J.E. (2007) *Nature Methods*, **4**, 409.
39 Ozawa, T., Natori, Y., Sato, M. and Umezawa, Y. (2007) *Nature Methods*, **4**, 413.
40 Valencia-Burton, M., McCullough, R.M., Cantor, C.R. and Broude, N.E. (2007) *Nature Methods*, **4**, 421.
41 States, D.J., Gish, W. and Altschul, S.F. (1991) *Methods (San Diego, Calif.)*, **3**, 66.
42 Rhee, W.J., Santangelo, P.J., Jo, H. and Bao, G. (2007) *Nucleic Acids Research*, **36**, e30.
43 Lakowicz, J.R. (1999) *Principles of Fluorescence Spectroscopy*, 2nd edn, Plenum Press, New York.
44 Marras, S.A., Kramer, F.R. and Tyagi, S. (2002) *Nucleic Acids Research*, **30**, e122.
45 Price, N.C. and Stevens, L. (1999) *Fundamentals of Enzymology: The Cell and Molecular Biology of Catalytic Proteins*, 3rd edn, Oxford University Press, New York.
46 Dokka, S. and Rojanasakul, Y. (2000) *Advanced Drug Delivery Reviews*, **44**, 35.
47 Leonetti, J.P., Mechti, N., Degols, G., Gagnor, C. and Lebleu, B. (1991) *Proceedings of the National Academy of Sciences of the United States of America*, **88**, 2702.
48 Mhlanga, M.M., Vargas, D.Y., Fung, C.W., Kramer, F.R. and Tyagi, S. (2005) *Nucleic Acids Research*, **33**, 1902.
49 Giles, R.V., Ruddell, C.J., Spiller, D.G., Green, J.A. and Tidd, D.M. (1995) *Nucleic Acids Research*, **23**, 954.

50 Barry, M.A. and Eastman, A. (1993) *Archives of Biochemistry and Biophysics*, **300**, 440.

51 Giles, R.V., Spiller, D.G., Grzybowski, J., Clark, R.E., Nicklin, P., Tidd, D.M. (1998) *Nucleic Acids Research*, **26**, 1567.

52 Walev, I., Bhakdi, S.C., Hofmann, F., Djonder, N., Valeva, A., Aktories, K. and Bhakdi, S. (2001) *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 3185.

53 Snyder, E.L. and Dowdy, S.F. (2001) *Current Opinion in Molecular Therapeutics*, **3**, 147.

54 Wadia, J.S. and Dowdy, S.F. (2002) *Current Opinion in Biotechnology*, **13**, 52.

55 Becker-Hapak, M., McAllister, S.S. and Dowdy, S.F. (2001) *Methods (San Diego, Calif.)*, **24**, 247.

56 Wadia, J.S. and Dowdy, S.F. (2005) *Advanced Drug Delivery Reviews*, **57**, 579.

57 Brooks, H., Lebleu, B. and Vives, E. (2005) *Advanced Drug Delivery Reviews*, **57**, 559.

58 Allinquant, B., Hantraye, P., Mailleux, P., Moya, K., Bouillot, C. and Prochiantz, A. (1995) *The Journal of Cell Biology*, **128**, 919.

59 Troy, C.M., Derossi, D., Prochiantz, A., Greene, L.A. and Shelanski, M.L. (1996) *The Journal of Neuroscience*, **16**, 253.

60 Minamoto, T., Mai, M. and Ronai, Z. (2000) *Cancer Detection and Prevention*, **24**, 1.

61 Altieri, D.C. and Marchisio, P.C. (1999) *Laboratory Investigation; A Journal of Technical Methods and Pathology*, **79**, 1327.

62 Santangelo, P.J., Nitin, N. and Bao, G. (2005) *Journal of Biomedical Optics*, **10**, 44025.

63 Santangelo, P., Nitin, N., LaConte, L., Woolums, A. and Bao, G. (2006) *Journal of Virology*, **80**, 682.

64 Santangelo, P.J. and Bao, G. (2007) *Nucleic Acids Research*, **35**, 3602.

65 Liu, X. and Tan, W. (1999) *Analytical Chemistry*, **71**, 5054.

66 Kambhampati, D., Nielsen, P.E. and Knoll, W. (2001) *Biosensors and Bioelectronics*, **16**, 1109.

67 Steemers, F.J., Ferguson, J.A. and Walt, D.R. (2000) *Nature Biotechnology*, **18**, 91.

68 Kuhn, H., Demidov, V.V., Coull, J.M., Fiandaca, M.J., Gildea, B.D. and Frank-Kamenetskii, M.D. (2002) *Journal of the American Chemical Society*, **124**, 1097.

69 Hamaguchi, N., Ellington, A. and Stanton, M. (2001) *Analytical Biochemistry*, **294**, 126.

70 Yamamoto, R., Baba, T. and Kumar, P.K. (2000) *Genes to Cells: Devoted to Molecular & Cellular Mechanisms*, **5**, 389.

7
High-Content Analysis of Cytoskeleton Functions by Fluorescent Speckle Microscopy

Kathryn T. Applegate, Ge Yang, and Gaudenz Danuser

7.1
Introduction

In 1949, Linus Pauling observed that hemoglobin in patients with sickle cell anemia is structurally different from that in healthy individuals [1]. This seminal discovery of a 'molecular disease' overturned a century-old notion that all diseases were caused by structural problems at the cellular level. Today, we know that disease can arise from aberrations in the expression, regulation or structure of a single molecule. Frequently, such aberrations interfere with one or more of the cell's basic morphological activities, including cell division, morphogenesis and maintenance in different tissue environments, or cell migration.

The advent of molecular pathology precipitated the rise of molecular biology and genomics, which in turn jump-started other large-scale '-omics' fields. In parallel, sophisticated imaging, quantitative image analysis and bioinformatics approaches were developed. These methods have enabled a quantum leap in our knowledge base about the molecular underpinnings of life, and what goes wrong during disease. Much has already been translated to the clinic. For example, mutation and gene expression profiles can be used to prescribe targeted drugs to breast cancer patients [2], and in the US many states have adopted metabolic screening programs to test newborns for a growing number of disorders [3].

Yet on the whole, the genomic era has failed to yield the 'goldmine of personalized interventions' that it first promised. Drug development pipelines rely heavily on high-throughput screens to identify compounds that have a desired effect on the biochemical activity of a particular drug target. These screens, however, cannot resolve whether a 'hit' will be active in living cells and specific to the pathway of interest. To avoid this limitation, high-content screens – also called 'phenotypic' or 'imaging' screens – use automated image analysis methods to detect desired changes in cells photographed under the light microscope [4]. Changes in gross cell morphology or the spatial activity of a protein of interest become apparent when analyzing a large population of cells. Although the identification of a molecular target causing a

Nanotechnology, Volume 5: Nanomedicine. Edited by Viola Vogel

ISBN: 978-3-527-31736-3

phenotypic change can be rate-limiting, assays can often be designed with the drug target already in mind. The larger challenge is to derive meaningful, quantitative phenotypic information from images [4].

The problem of extracting phenotypic information is compounded by the fact that many phenotypic differences can only be resolved in *time*; the dynamics of molecules, not just their concentrations and localization, are important in disease development. In other words, a cell may look healthy in a still image, but an analysis of the underlying dynamics of cell-adhesion proteins, for example, may reveal that the cell has metastatic potential. Current phenotypic screens can only distinguish between coarse, spatially oriented phenotypes [5], while many diseases exhibit extremely subtle, yet significant, phenotypes. New methods are needed to extract and correlate dynamic descriptors if we are to design drugs and other nanomedical intervention strategies with minimal side effects.

In this chapter, we review quantitative fluorescent speckle microscopy (qFSM), a relatively young imaging technology that has been used to characterize the dynamic infrastructure of the cell. qFSM has the potential to become a unique assay for live-cell phenotypic screening that will guide the development of drugs and other nanomedical strategies based on the dynamics of subcellular structures. We begin by summarizing how regulation of the cytoskeleton contributes to important cell morphological processes that go awry in disease. We then describe how fluorescent speckles form to mark the dynamics of subcellular structures. Next, we illustrate critical biological insights that have been gleaned from qFSM experiments. We conclude with new applications and an outlook on the future of qFSM.

7.2 Cell Morphological Activities and Disease

The filamentous actin (F-actin), intermediate filament (IF) and microtubule (MT) cytoskeleton systems are key mediators of cell morphology (Figure 7.1). Each filament system is unique in its physical properties and extensive subset of associated proteins [6]. Endogenous and exogenous chemical and mechanical signals control the precise arrangement of these dynamic polymers, and defects in their regulation are seen in a wide variety of diseases [7].

7.2.1 Cell Migration

One of the most fundamental cell morphological functions is migration. Many cell types in the body are motile, including fibroblasts, epithelial cells, neurons, leukocytes and stem cells. Failure to migrate, or migration to the wrong location in the body, can lead to congenital heart or brain defects, atherosclerosis, chronic inflammation, neurodegenerative disease, compromised immune response, defects in wound healing and tumor metastasis [8].

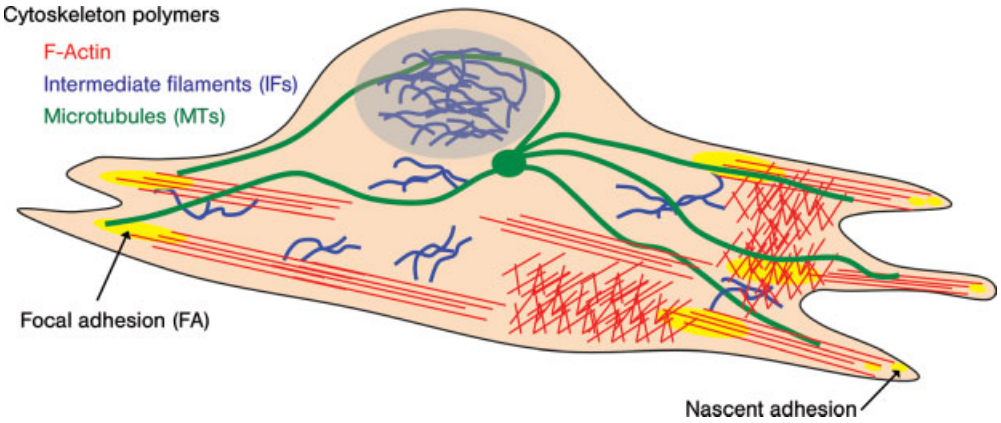


Figure 7.1 The F-actin, microtubule and intermediate filament cytoskeletons and adhesions in a migrating cell. A mesh-like F-actin network (red) at the leading edge drives the plasma membrane forward. Contractile F-actin bundles (red) linked to strong adhesions (yellow) in the front and weak adhesions in the back promote tail retraction. MTs (green) are implicated in cell polarization and adhesion regulation. The IFs (blue) are a heterogeneous group of filamentous proteins which help maintain cell structural integrity.

Cell migration is a remarkably complex behavior at the molecular level. Cell crawling involves three basic steps [9]: (i) leading edge protrusion; (ii) contraction of the F-actin polymer network; and (iii) tail retraction. Net movement of the cell cannot occur unless the F-actin network is anchored to the substrate. Otherwise, the protruding forces generated by F-actin polymerization at the leading edge plasma membrane and the contraction of the network by myosin molecular motors would simply deform the cell, without creating traction. Cells accomplish this via adhesion organelles located on the cell's ventral surface (Figure 7.1). In the case of uniform cell adhesion, actin network contraction would pull equally on the front and rear of the cell, resulting in zero movement. Directional movement thus requires an adhesion *gradient*, which is established by adhesions assembling at the protruding edge and weakening as they mature towards the cell rear [10]. The control of such a spatial gradient is extremely complex and involves many molecular components [11], the interactions of which are still poorly understood.

Like F-actin and adhesions, MTs are also important for migration in many cell types. They are implicated in polarizing the cell by delivering signaling molecules and regulating the turnover of adhesions [12]. In addition, mechanical interactions between MTs and F-actin may also contribute to the control of cell movement and morphology [13]. Tight spatial and temporal coordination between the F-actin, MT and adhesion systems is critical for cell migration.

7.2.2 Cell Division

Division is another cellular function that depends on a complex and highly regulated series of molecular events. Division is essential during embryogenesis and

development, and also occurs constantly in tissues of the adult body. In the intestines alone, approximately 10^{10} old cells are shed and replaced every day [14]. When DNA replication is complete, the pairs of chromosomes must be pulled apart symmetrically and segregated to opposite ends of the cell. Segregation is accomplished by a dynamic structure called the spindle, which is composed of MTs and motor proteins. In the final step of cell division – cytokinesis – the spindle elongates and a contractile actin structure develops to pinch off the membrane, partitioning organelles and cytoplasm into two daughter cells. While these processes progress with remarkable fidelity in healthy individuals, unchecked and faulty cell division are hallmarks of oncogenesis [15] and age-related disorders [16]. Cell cycling of adult neurons may be implicated in Alzheimer's disease [17]. Analysis of the architectural dynamics of the F-actin and MT structures involved in these steps is critical if we are to intervene with abnormal events in cell division associated with disease development.

7.2.3
Response to Environmental Changes

Cells also must be able to respond to physical changes in the environment. Almost 15 years ago, Wang *et al*. reported that applying a mechanical stimulus to integrin transmembrane receptors in adhesions caused the cytoskeleton to stiffen in proportion to the load [18]. The increased stiffness required the presence of intact F-actin, MTs and IFs. Such adaptation is central to the formation of multicellular tissues and functional organelles [19]. Recent studies have also provided evidence that mechanical cues relayed through the cytoskeleton systems dictate stem cell fate. Naïve mesenchymal stem cells cultured on a soft, brain-like matrix differentiate into neurons, while those cultured on a more rigid, bone-like matrix develop into osteoblasts [20]. When F-actin contraction by myosin motors is inhibited, lineage specification by elasticity is blocked. Mechanistic analyses of these differentiation-defining processes will require a quantitative analysis of the underlying cytoskeleton dynamics.

7.2.4
Cell–Cell Communication

Many diseases are directly linked to abnormalities in cell–cell or intracellular communication. For example, normal epithelial cells grow into organized, confluent monolayers because their growth is constantly monitored and controlled by cell–cell contacts. Carcinoma cells, on the other hand, lose this cue and grow into a mass. In addition to cancer, the loss of control over cell–cell contacts can result in embryonic death, severe developmental defects, neuropathy, skin-blistering diseases, diabetes, autoimmune disorders and atherosclerosis [21]. Besides the communication between cells in direct attachment, cells also communicate via a number of pathways, from the endo- or exocytosis of a few receptor-bound molecules to the uptake of large particles via phagocytosis [6]. Pathogens have managed to hijack these pathways to enter and exit cells [22]. In all of these communication pathways, the three cytoskeleton systems are on center stage. Analogously, the import of drugs or nanomedical

devices into defective cells requires the specific activation of one of these pathways in the right place, at the right time. Thus, our ability to understand disease and to precisely manipulate cells depends on our ability to analyze the cytoskeleton structure and dynamics *in situ* in living cells. We will now introduce qFSM as one of the emerging tools to achieve this goal.

7.3 Principles of Fluorescent Speckle Microscopy (FSM)

As reviewed by Danuser and Waterman-Storer [23], FSM enables quantitative analysis of the dynamics of subcellular structures *in vitro* and *in vivo*. Like many innovations in science, FSM was discovered by accident when Clare Waterman-Storer noticed that the microinjection of a small amount of X-rhodamine-labeled tubulin into cells gave rise to MTs with a speckled appearance [24]. Over the past ten years, sample preparation, speckle imaging and computational image analysis have been developed and improved to yield robust measurements of intracellular flow and assembly/disassembly dynamics for both MT and F-actin structures. Very recently, FSM has also been applied to measure interaction dynamics between molecular assemblies, such as transient F-actin and adhesion coupling in migrating epithelial cells [25] and F-actin and MT comovement in epithelial cells [26] and neurons [27]. Moreover, computational post-processing of FSM data has yielded indirect information, such as intracellular forces [28]. Thus, FSM is a prime imaging mode for interrogating the cytoskeleton's many roles in cell physiology and disease.

FSM is a twist on conventional live-cell fluorescence microscopy, where structures are visualized either by the expression of a fluorescent protein fused to the protein of interest, or by the microinjection of a covalently labeled protein into the cell. Typically in fluorescence microscopy, high expression levels or large amounts of injected fluorescent protein are necessary to achieve a high signal-to-noise ratio (SNR). This approach reveals protein localization and, to an extent, the movement of molecular structures in cells. However, the ability to report protein dynamics is limited due to the inherently high background fluorescence from out-of-focus incorporated and diffusing, unincorporated fluorescent protein. In addition, it is impossible to detect the movement or turnover of protein subunits within a larger, uniformly labeled structure. Laser photobleaching and photoactivation can help by marking structures in defined regions of the cell and allowing the measurement of recovery or movement in the local region at steady state [29–34]. Similar to these techniques is the ratiometric method of fluorescence localization after photobleaching (FLAP), which shows the diffusive and convective transport of unincorporated protein [35, 36].

In contrast to conventional fluorescence microscopy, FSM probes the dynamics of protein assemblies which contain very few (<1%) fluorescent subunits among a vast majority of unlabeled subunits. When imaged by high-resolution, diffraction-limited optics, the scattered distribution of fluorescence yields a punctate (speckled) pattern that reveals the motion of the entire structure, the reorganization of subunits within

the structure, and the association and dissociation of subunits. Thus, FSM provides the same information as the aforementioned photomarking techniques, but does so across much or all of the cell simultaneously. As FSM does not require active marking, it allows the continuous detection of nonsteady-state dynamics within protein assemblies, and reveals spatial and temporal relationships between these dynamic events at submicron and second resolution. FSM also reduces out-of-focus fluorescence and improves the visibility of fluorescently labeled structures and their dynamics in three-dimensional (3-D) polymer arrays, such as the mitotic spindle [37–39].

In the early years of its development, FSM used wide-field epifluorescence light microscopy and digital imaging with a sensitive, low-noise, cooled charge-coupled-device (CCD) camera [24]. Since then, FSM has been transferred to confocal and total internal reflection fluorescence (TIRF) microscopes [37, 40–42]. The development of fully automated, computer-based tracking and the statistical analysis of speckle behavior proved to be critical steps in establishing FSM as a routine method for measuring cytoskeleton architectural dynamics. Thus, FSM is an integrated technology in which sample preparation, imaging and image analysis are optimized to achieve detailed information about polymer dynamics.

7.4
Speckle Image Formation

7.4.1
Speckle Formation in Microtubules (MTs): Stochastic Clustering of Labeled Tubulin Dimers in the MT Lattice

MTs exhibit a variation in fluorescence intensity along their lattices when cells are injected with a small amount of labeled tubulin dimers, leading to a speckled appearance (Figure 7.2a and b). Several possibilities exist for how speckles arise in this situation:

- The fluorescent tubulin could form oligomers or aggregates on the MT.
- Cellular organelles or MT-associated proteins (MAPs) could be bound to the MT and conceal or quench the fluorescence in some regions.
- Random variation of the number of fluorescent tubulin subunits in each diffraction-limited image region along the MT could occur as the MT assembles from a pool of labeled and unlabeled dimers [43].

The first hypothesis was discounted by showing that labeled tubulin dimers sediment similarly to unlabeled dimers in an analytical ultracentrifugation assay. Next, it was shown that MTs assembled from purified tubulin *in vitro* exhibited similar speckle patterns to MTs in cells, where MAPs and organelles are present [43]. Thus, the most plausible explanation for MT speckle formation is that variations exist in the number of fluorescent tubulin subunits in each resolution-limited image region along the MT.

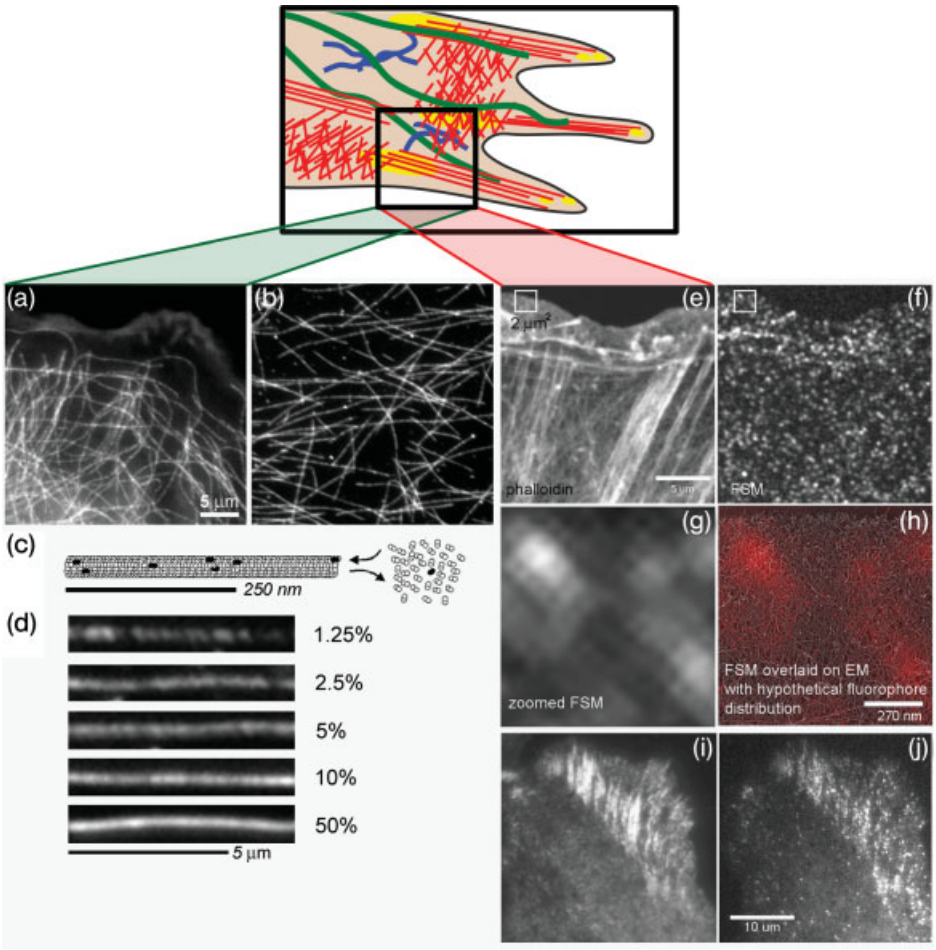


Figure 7.2 Speckle formation in microtubules, F-actin and FAs. The cartoon is an inset of Figure 7.1, showing that MTs (a) and F-actin (e, f) in cells can be imaged in separate channels. (a, b) Comparison of random speckle pattern of fluorescence along MTs for (a) a living epithelial cell microinjected with X-rhodamine-labeled tubulin and for (b) MTs assembled *in vitro* from 5% X-rhodamine-labeled tubulin; (c) Model for fluorescent speckle pattern formation in a MT grown from a tubulin pool containing a small fraction of labeled dimers; (d) Dependence of speckle contrast on the fraction of labeled tubulin dimers. (e, f) Speckle formation in actin filament networks. An epithelial cell was microinjected with a low level of X-rhodamine-labeled actin, fixed, and stained to show structure with Alexa-488 phalloidin. (e) Phalloidin image showing the organization of actin filaments in amorphous filament networks and bundles; (f) In the single FSM image, much of the structural information is lost, but time-lapse FSM series contain dynamic information of filament transport and turnover not accessible with higher-level labeling of the cytoskeleton; (g) Close-up of 2×2 µm window in panels e and f; (h) Colorized speckle signal overlaid onto a quick-freeze deep etch image of the same-sized region of the actin cytoskeleton in the leading edge of a fibroblast (kindly provided by Tatyana Svitkina). The hypothetical fluorophore distribution (red) could give rise to such a speckle pattern, indicating the scale of speckles in comparison with the polymer network ultrastructure. It also illustrates that a small proportion of the total actin fluoresces and that fluorophores from different filaments contribute to the same speckle; (i, j) Low-level expression of the GFP-tagged FA protein vinculin results in speckled FAs in TIRF images. A cell expressing GFP-vinculin was fixed and immunofluorescently stained with antibodies to (i) vinculin to reveal the position of FAs, which in the (j) GFP channel appear speckled because of the low level of incorporation of GFP-vinculin. Figure reproduced with permission from Ref. [23].

To understand how speckles originate, consider the incorporation of fluorescent subunits into the helical MT lattice, which consists of 1625 α/β tubulin dimers per micron (Figure 7.2c) [44]. The image of an individual MT under the light microscope results from a convolution of the fluorophore distribution along the MT with the point-spread function (PSF) of the microscope. Ignoring the vertical dimension, the in-focus slice of the PSF is given by the Airy disk. Given the emission wavelength λ of the fluorophore and the numerical aperture (NA) of the microscope objective, the first ring of the Airy disk with zero intensity has a radius $r = 0.61\,\lambda/\mathrm{NA}$ [45]. Objects separated by less than r cannot be resolved. For X-rhodamine-labeled MTs ($\lambda = 620$ nm) at a high NA (1.4), the Airy disk radius $r = 270$ nm, which corresponds to 440 tubulin subunits (270 nm × 1.625 subunits nm^{-1}). A given fraction of fluorescent dimers, f, produces a mean number of fluorescent dimers $n = 440 \cdot f$ per PSF. Variations in the number of fluorescent dimers per PSF relative to this mean produce the speckle pattern along the MT. The speckle pattern contrast in this example can be approximated by the ratio of the standard deviation and the mean of a binomial distribution with 440 elements: $c = \sqrt{440 \cdot f \cdot (1-f)}/(440 \cdot f)$. Accordingly, the contrast c can be increased by decreasing f (Figure 7.2d) or by making the Airy disk smaller – that is, effectively lowering the number of tubulin subunits per Airy disk. The latter is accomplished by using optics with the highest NA possible. Experiments have shown that fractions in the range of 0.5 to 2%, where speckles consisting of three to eight fluorophores are optimal for the speckle imaging of individual microtubules [46].

7.4.2
Speckle Formation in Other Systems: The Platform Model

Conventional fluorescence microscopy of the F-actin network reveals a variety of actin-based structures, such as stress fibers and filopodia, but this approach is not conducive to the observation of structural dynamics (see Section 7.3). When fluorescently labeled actin is injected into a cell at a low level relative to the amount of endogenous, unlabeled actin, actin-rich structures appear relatively evenly speckled (Figure 7.2e and f) [26, 47–50]. Importantly, while the speckle pattern reveals the local dynamics of actin structures the overall architectural organization of the cytoskeleton is no longer visible. Speckle formation can also be found when expressing green fluorescent protein (GFP)-fused actin at a very low level [48, 51] or by injecting trace amounts of the labeled actin-binding molecule phalloidin [52, 53]. In contrast to speckle formation in isolated MTs, labeled actin subunits bind within a highly cross-linked 3-D network of F-actin filaments [54–56]. The mesh size of an F-actin network in living cells is nearly always below the resolution limit of the light microscope (Figure 7.2g and h). Consequently, unless f is kept extremely low so that only one fluorophore falls into the PSF volume [48], most fluorescent speckles arise from subunits on multiple actin filaments.

The same concept applies to speckle formation within adhesion sites [40]. GFP-fusions to adhesion proteins, including vinculin, talin, paxilin, α-actinin, zyxin and $\alpha_v$integrin [57], have been expressed in epithelial cells from crippled promoters to achieve very low expression levels. Labeled proteins assembling with endogenous,

unlabeled proteins give the adhesions a speckled appearance (Figure 7.2i and j). As with F-actin networks, the speckles represent randomly distributed fluorescent adhesion proteins that are temporarily clustered in the adhesion complex within the volume of one PSF.

In summary, a speckle is defined as a diffraction-limited region that is significantly higher in fluorophore concentration (i.e. higher in fluorescence intensity) than its local neighborhood. For a speckle signal to be detected in an image, the contributing fluorescent molecules must be associated with a molecular scaffold, or 'speckle platform', during the 0.1–2.0 s exposure time required by most digital cameras to acquire the dim FSM image. Conversely, unbound, diffusible fluorescent molecules visit many pixels during the exposure and yield an evenly distributed background signal, instead of speckles [48]. The same idea was illustrated for the MT MAP ensconsin [58] and for the MT kinesin motor Eg5 [59]. Thus, association with the platform can occur when labeled subunits either *become* part of the platform, as with tubulin or actin, or simply *bind* to it, as in the case of cytoskeleton-associated proteins or adhesion molecules.

7.5
Interpretation of Speckle Appearance and Disappearance

7.5.1
Naïve Interpretation of Speckle Dynamics

Following the platform model, one would expect the appearance of a speckle to correspond to the local association of subunits with the platform. Conversely, the disappearance of a speckle would mark the local dissociation of subunits. In other words, FSM allows – in principle – the direct kinetic measurement of subunit turnover in space and time via speckle lifetime analysis. In addition, once a speckle is formed, it may undergo motion that indicates the coordinated movement of labeled subunits on the platform and/or the movement of the platform itself.

7.5.2
Computational Models of Speckle Dynamics

The interpretation of speckle dynamics becomes significantly more complicated when individual speckles arise from fluorophores distributed over multiple polymers. To examine how speckle appearance and disappearance relate to the rates of assembly and disassembly of F-actin, we performed Monte Carlo simulations of fluorophore incorporation into growing and shrinking filaments in dense, branched networks, and generated synthetic FSM time-lapse sequences [60]. The first lesson learned from this modeling was that the speckle density is independent of whether the network assembles or disassembles; it depends only on how many Airy disks can be resolved per square micron. With NA = 1.4/100× optics, this amounts to ∼4 (approximately 2 × 2), as confirmed by Figure 7.3a. The graph displays the mean

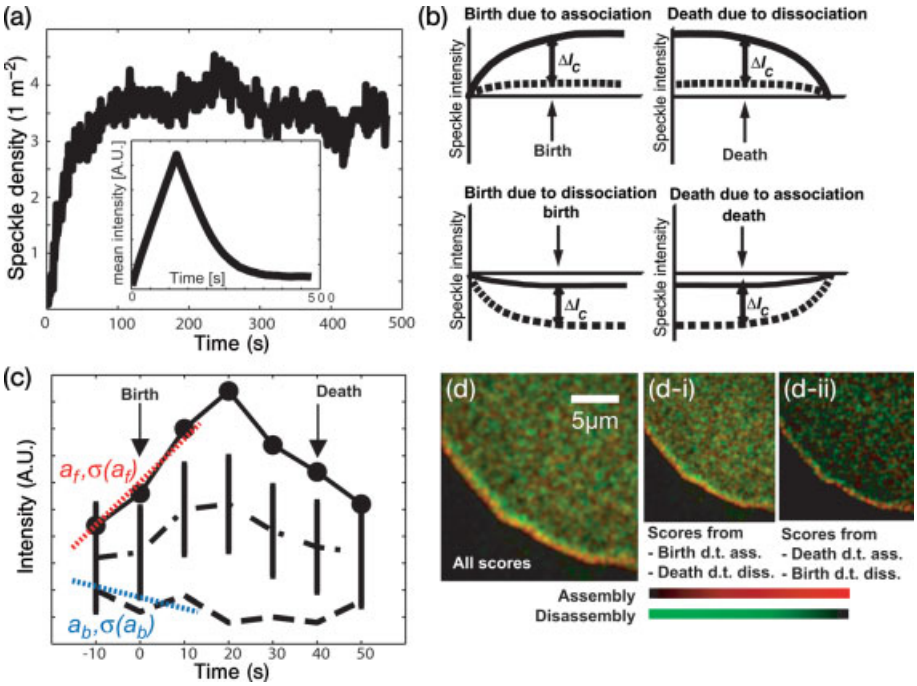


Figure 7.3 Relationship between speckle appearance (birth) and disappearance (death) and the turnover in the underlying macromolecular assembly. (a) Simulated speckle density in an actin filament network assembling for 120 s and disassembling for 360 s. Inset: Mean intensity indicating the overall change in bound fluorophore over time; (b) Classification of speckle birth and death due to monomer association and dissociation with the network. A speckle appears when the difference between foreground (solid line) and background (dashed line) is greater than a threshold ΔI_C, which is a function of the camera noise and the shot noise of the signal [60]; (c) Measurement of intensity changes in foreground (solid line) and background (dashed line) during a speckle birth. The entire lifetime of the speckle is shown (40 s). Dash-dotted line: Mean between foreground and background; error bars: ΔI_C computed in every time point. Birth and death are defined as the time points at which the intensity difference exceeds ΔI_C for the first and the last time. Red dotted line: Regression line to the foreground intensity values before, at and after birth. Blue dotted line: Regression line to the background intensity values. The cause of speckle birth is inferred by statistical classification of the two slopes and their standard deviations (see text). The statistically more dominant of the two slopes, if also significant relative to image noise, defines the score of the event (foreground slope in the example given); (d) Spatial averaging of scores accumulated over a defined time window of an FSM time-lapse sequence yields maps of net polymerization (red) and depolymerization (green). Scores from birth due to association and death due to dissociation (d-i) or from birth due to dissociation and death due to association (d-ii) reveal the same spatial distribution of polymerization and depolymerization. Figure in parts reproduced with permission from Refs [60, 83].

speckle density from five simulations of a network that starts with no fluorophores, assembles for 120 s (inset: mean fluorescence intensity increases), and disassembles for 360 s (inset: mean fluorescence intensity decreases) at equal rates. The density does not change after saturation at 100 s and remains constant, despite the further

addition of fluorophores for another 20 s. This suggests that monomer association can cause an equal number of speckle appearances and disappearances. The same holds true in the opposite sense during network disassembly.

Whereas, the NA of the optics defines the maximum number of resolvable speckles per unit area, the labeling ratio influences the speckle density indirectly. For multi-fluorophore speckles the ratio f is the main determinant of speckle contrast. When increasing this ratio, the speckle density drops because the difference between the peak intensity of a speckle and its surroundings is no longer distinguishable from intensity fluctuations due to noise. Similarly, at labeling ratios where speckles represent the image of single fluorophores ($f < 0.1\%$), a further decrease in f reduces the speckle density proportionally. Across the optimal range of ratios for multi-fluorophore speckles ($0.5\% < f < 3\%$) the density is almost constant. These model predictions were largely confirmed experimentally by Adams *et al.* [40].

The origin of constant speckle density in the range of $0.5\% < f < 3\%$ is illustrated in Figure 7.3b. A speckle represents a local image intensity maximum significantly above the surrounding background. The critical intensity difference ΔI_c depends on both the camera noise and the shot noise. The shot noise is by itself a function of the speckle intensity [60]. Speckles may appear (speckle birth) for two reasons: (i) the intensity of a local maximum becomes brighter because of the association of fluorescent subunits; or (ii) the intensity of the surrounding background becomes dimmer because of subunit dissociation in the neighborhood. In both cases, a speckle birth is detected when the peak-to-background intensity difference exceeds ΔI_c. Analogously, speckles may disappear (speckle death) either because of subunit dissociation in the location of a speckle or because of subunit association in the neighborhood.

7.5.3
Statistical Analysis of Speckle Dynamics

With the classification scheme in Figure 7.3b, speckles become time-specific, diffraction-limited probes of turnover of subcellular structures. The change in foreground or background intensity that causes the birth or death of a speckle is, on average, proportional to the net number of subunits Δm added to or removed from the PSF volume between two frames. This defines an algorithm for the local measurement of network assembly or disassembly kinetics [60]:

- Calculation of changes in foreground and background intensities. After detection of a speckle birth/death event, regression lines are fitted to the foreground and background intensities for one time point before, during and after the event (Figure 7.3c). Intensity values before birth and after death are extrapolated [60]. The line fits provide two estimates a_f and a_b of the slopes of foreground and background intensity variation. They also yield the standard deviations σa_f and σa_b of the slopes, which are derived from the residuals of the intensity values to the regression line. Noisy data, poorly represented by the regression model, generate large values of σ; intensity values in perfect match with the model result in small values of σ.

- Each of the two slopes is tested for statistical significance. Insignificant intensity changes are discarded.
- If both foreground and background slopes are significant, the one with the higher significance (lower *p*-value) is selected as the cause of the event. In the example in Figure 7.3c the foreground slope has the higher significance. The magnitude of the more significant slope is recorded as the score of the birth/death event. If neither foreground nor background slope is statistically significant, no score is generated.

Score values represent instantiations of a random variable with an expectation value $\mu = \alpha\Delta m{\cdot}f$ and variance $\sigma^2 = \alpha^2\Delta m^2{\cdot}f(1-f)$, where α denotes the unknown intensity of one fluorophore. In addition, the scores are perturbed by noise. However, assuming that the net rate Δm remains constant for a small probing window, the intrinsic score variation and noise are approximately eliminated by averaging all scores falling into the window. The choice of the window size depends on the density of significant scores and the demand for spatial or temporal resolution. The more scores averaged by time integration, the less spatial averaging is required, and *vice versa*.

Figure 7.3d displays rates of actin assembly (red) and disassembly (green) of the F-actin network at the edge of an epithelial cell. Score values were averaged over 10 min, reflecting the steady-state turnover. The two smaller panels indicate the rate distributions calculated from scores extracted from speckle births due to monomer association and from speckle deaths due to monomer dissociation only (Figure 7.3d-i), and from births due to monomer dissociation and from deaths due to monomer association only (Figure 7.3d-ii). Both panels display the same distribution of loci of strong assembly (for example, the cell edge) and disassembly but at different event densities. Figure 7.3d-i thus corresponds to the naïve interpretation of speckle appearance and disappearance. These events contribute ~70% of all scores; the other ~30% of significant scores is related to the counterintuitive cases of speckle birth and death, and neglecting these would significantly reduce the sample size. How many intuitive versus counterintuitive cases occur depends on the fraction of labeled monomers. The lower the fraction, the fewer counterintuitive cases are observed, with a lower boundary defined by the single-fluorophore speckle regime, where all speckle appearances are due to monomer association and all disappearances are due to monomer dissociation.

The processing of only short time intervals around speckle birth and death events focuses the analysis on image events that are more likely to have originated from monomer exchange rather than from intensity fluctuations due to image noise, bleaching and in- and out-of-focus speckle motion. In addition, the algorithm rejects ~60% of all speckle birth and death events as insignificant [61]. That is, these events are not classifiable as induced by monomer exchange with the certainty the user chooses as the confidence level for the analysis. Bleaching affects all speckle scores, and thus can be corrected based on global drifts in the image signal [60]. It has also been shown that, with a NA = 1.4 objective lens, focus drifts smaller than 100 nm over three frames (e.g. 30 nm per 1–5 s) have no effect on the mapping of network

turnover. Thus, the statistical model described in this section provides a robust method for calculating spatiotemporal maps of assembly and disassembly of subcellular structures such as F-actin networks.

7.5.4
Single- and Multi-Fluorophore Speckles Reveal Different Aspects of the Architectural Dynamics of Cytoskeleton Structures

Intuitively, it seems that FSM would be most powerful if implemented as a single-molecule imaging method, where speckle appearances and disappearances unambiguously signal association and dissociation of fluorescent subunits to the platform [48]. However, the much simpler signal analysis of single-molecule images is counterweighed by several disadvantages not encountered when using multifluorophore speckles. First, establishing that an image contains only single-fluorophore speckles can be challenging, especially when the signal of one fluorophore is close to the noise floor of the imaging system. Especially in 3-D structures a large number of speckles have residual contributions from at least one other fluorophore, and those mixtures must be eliminated from the statistics. Second, the imaging of single-fluorophore speckles is practically more demanding than multifluorophore FSM and requires longer camera exposures to capture the very dim signals, thus reducing the temporal resolution. Third, in addition to the lower temporal resolution, single-fluorophore FSM offers lower spatial resolution because the density of speckles drops significantly with the extremely low labeling ratio required for single-molecule-imaging conditions. In dense, crosslinked structures such as the F-actin network the intensity variation during appearance and disappearance events of multifluorophore speckles can distinguish between fast and slow turnover. In contrast, single-fluorophore speckles deliver on/off information only. Thus, in order to measure rates of the turnover of molecular structures on a continuous scale, single-fluorophore speckle analysis must rely on spatial and temporal averaging, which further decreases the resolution, while multifluorophore speckles provide this information at a finer spatial and shorter temporal scale.

On the other hand, multifluorophore speckles cannot resolve the dynamics of closely apposed individual units with subcellular structures. If the dynamics of the individual building blocks of a structure is of interest, and the lower density of spatial and temporal samples can be afforded, single-fluorophore speckles are adequate probes. If information about individual units *and* the ensemble of units is needed, single-fluorophore and multifluorophore speckle imaging can be combined in two spectrally distinct channels. This has been demonstrated in an analysis of the *Xenopus* spindle [62], where combined single- and multifluorophore qFSM revealed the overall dynamics of MTs in the spindle, as well as the dynamics and length distribution of individual MTs within densely packed bundles inside the spindle (see Section 7.9.2).

In summary, FSM can probe different aspects of the architectural dynamics of subcellular structures at different spatial and temporal scales via modulation of the

ratio of labeled to unlabeled subunits. Currently, the exact labeling ratio is difficult to control in a given experiment. Statistical clustering analysis of the resulting speckle intensity is required to identify the distribution of the numbers of fluorophores within speckles. In the near future, these mathematical methods will be complemented with sophisticated molecular biology that will allow relatively precise titration of the labeled subunits. Together, these approaches will be invaluable to a systematic mapping of the heterogeneous dynamics of complex subcellular structures such as the cytoskeleton.

7.6 Imaging Requirements for FSM

Time-lapse FSM requires the imaging of high-resolution, diffraction-limited regions containing one to ten fluorophores and the inhibiting of fluorescence photobleaching. This requires a sensitive imaging system with little extraneous background fluorescence, efficient light collection, a camera with low noise, high quantum efficiency, high dynamic range, high resolution, and the suppression of fluorescence photobleaching with illumination shutters and/or oxygen scavengers [49, 63, 64]. In addition, all fluorescently labeled molecules must be functionally competent to bind their platform; otherwise, they will contribute to diffusible background and reduce the speckle contrast [65]. The reader is referred to a review by Gupton and Waterman-Storer [66] for an in-depth discussion of the hardware requirements for obtaining FSM images.

Because FSM is achieved by the level of fluorescent protein in the sample, it is adaptable to various modes of high-resolution fluorescence microscopy, such that the specific advantages of each mode can be exploited in combination with the quantitative capabilities of FSM. For example, we have performed FSM on both spinning-disk confocal microscopy [41] and total internal reflection fluorescence microscopy (TIRFM) [40] systems to gain speckle data in two spectral channels with the specific image advantages of confocal and TIRFM. A comparison of FSM images of the actin cytoskeleton in migrating epithelial cells acquired by wide-field epifluorescence, spinning-disk confocal microscopy and TIRFM is shown in Figure 7.4a–c. Clearly, speckle contrast is improved by reducing out-of-focus fluorescence with either of the latter techniques. Contrast in TIRFM images is further improved over the spinning-disk confocal image because the evanescent field excitation depth is reduced to 100–200 nm into the specimen. When quantified, the effect of the reduced effective imaging volume on modulation and detectability of actin and focal adhesion (FA) speckles showed that TIRF-FSM indeed affords major improvements in these parameters over wide-field epifluorescence for imaging macromolecular assemblies at the ventral surface of living cells, both in thin peripheral and thick central cell regions [40]. Importantly, to date FSM has proved to be incompatible with all commercial laser-scanning confocal microscope systems. This is because these instruments use photomultipliers as detectors that are noisy and have a limited dynamic range compared to the

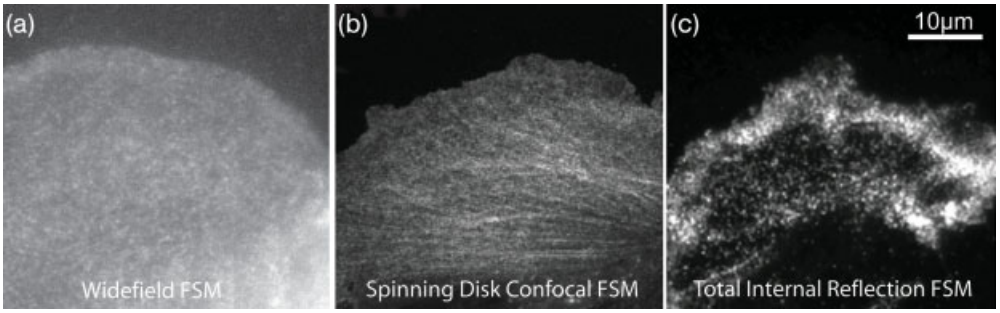


Figure 7.4 Comparison of X-rhodamine–actin FSM images of the edge of migrating Ptk1 epithelial cells using (a) wide-field epifluorescence, (b) spinning-disk confocal microscopy and (c) TIRF microscopy. Panels (a) and (b) were acquired using a Nikon 100 × 1.4 NA Plan Apo phase contrast objective lens and a 14 bit Hamamatsu Orca II camera with 6.7 micron pixels. Panel (c) was acquired with a Nikon 100 × 1.45 NA Plan Apo TIRF objective lens and a 14 bit Hamamatsu Orca II ER with 6.4 micron pixels. Note that speckle contrast and the ability to detect speckles in more central cell regions increases from panels (a) to (c). Note, however, in the TIRF image that speckles are very bright a few microns back from the edge, most likely where the cell is in closer contact with the substrate. Figure reproduced with permission from Ref. [23].

low-noise, high dynamic range CCDs used with spinning-disk confocal microscope systems.

7.7
Analysis of Speckle Motion

7.7.1
Tracking Speckle Flow: Early and Recent Developments

In addition to revealing the kinetics of association and dissociation of subunits to and from the molecular platform, speckles also show the movement of subunits within the platform and of the platform itself. In early applications of FSM, speckle motion was quantified by hand-tracking a few speckles – a tedious, error-prone and incomplete way of analyzing the wealth of information contained by these images [26, 27, 48]. Alternatively, kymographs provided average estimates of speckle velocities [37, 38, 49, 50, 59, 67–70].

Initial attempts to automate the extraction of more complete speckle flow maps from FSM time-lapse sequences of F-actin networks relied on correlation-based tracking. The speckled area of a source frame in the movie was divided into small probing windows, with each window being displaced until the normalized cross-correlation of the window with the signal of the next frame in the movie was maximized. This approach reported the average motion of all the speckles falling into the window. The window size pitted robustness in correlation against spatial resolution: the larger the window, the more unique was the speckle pattern to be

recognized in the target frame. On the other hand, larger windows increased the averaging of distinct speckle motions within the window.

Underlying the method of cross-correlation tracking is the assumption that the signal of a probing window, although translocated in space, does not change between source and target frame. In practice, this assumption is always violated because of noise. However, the cross-correlation of two image signals appears to be tolerant toward spatially uncorrelated noise, making it a prime objective function in computer vision tracking [71–73]. The many speckle appearances and disappearances in F-actin networks, however, introduce signal perturbations that cannot be tackled by the cross-correlation function [74]. Instead, we have developed a particle flow method, in which the movement of each speckle was tracked individually [74]. Speckles were linked between consecutive frames by nearest-neighbor assignment in a distance graph, in which conflicts between multiple speckles in the source frame competing for the same speckle in the target frame were resolved by global optimization [75]. An extension of the graph to linking speckles in three consecutive frames allowed enforcement of smooth and unidirectional trajectories, so that speckles moving in antiparallel flow fields could be tracked [74].

Surprisingly, cross-correlation-based tracking was successful in measuring average tubulin flux in meiotic spindles [76]. Simulated time-lapse sequences showed that if a significant subpopulation of speckles in the probing window moves jointly, then the coherent component of the flow can be estimated even when the rest of the speckles move randomly or, as in the case of the spindle apparatus, a smaller population moves coherently in opposite directions. However, the tracking result will be ambiguous if the window contains multiple, coherently moving speckle subpopulations of equal size. Miyamoto *et al*. [76] carefully chose windows in the central region of a half-spindle, where the motion of speckles towards the nearer of the two poles dominated speckle motion in the opposite direction and random components. The approach was aided further by several features of the spindle system: tubulin flux in a spindle is quasi-stationary; speckle appearances and disappearances are concentrated at the spindle midzone and in the pole regions, which were both excluded from the probing window; and the flow fields were approximately parallel inside the probing window.

Encouraged by these results, we returned to cross-correlation tracking of speckle flow in F-actin networks [77]. The advantage of cross-correlation tracking over particle flow tracking is that there is no requirement to detect the same speckle in at least two consecutive frames. Hence, speckle flows can be tracked in movies with high noise levels and weak speckle contrast [77]. In order to avoid trading correlation stability for spatial resolution, we capitalized on the fact that cytoskeleton transport is often stationary on the timescale of minutes. Thus, although the correlation of a single pair of probing windows in source and target frames is ambiguous (Figure 7.5b-i), rendering the tracking of speckle flow impossible (Figure 7.5c-i), time-integration of the correlation function over *multiple* frame pairs yields robust displacement estimates for probing windows as small as the Airy disk area (Figure 7.5b-ii, c-ii). Figure 7.5d presents a complete high-resolution speckle flow map extracted by integration over 20 frames ($\sim$3 min).

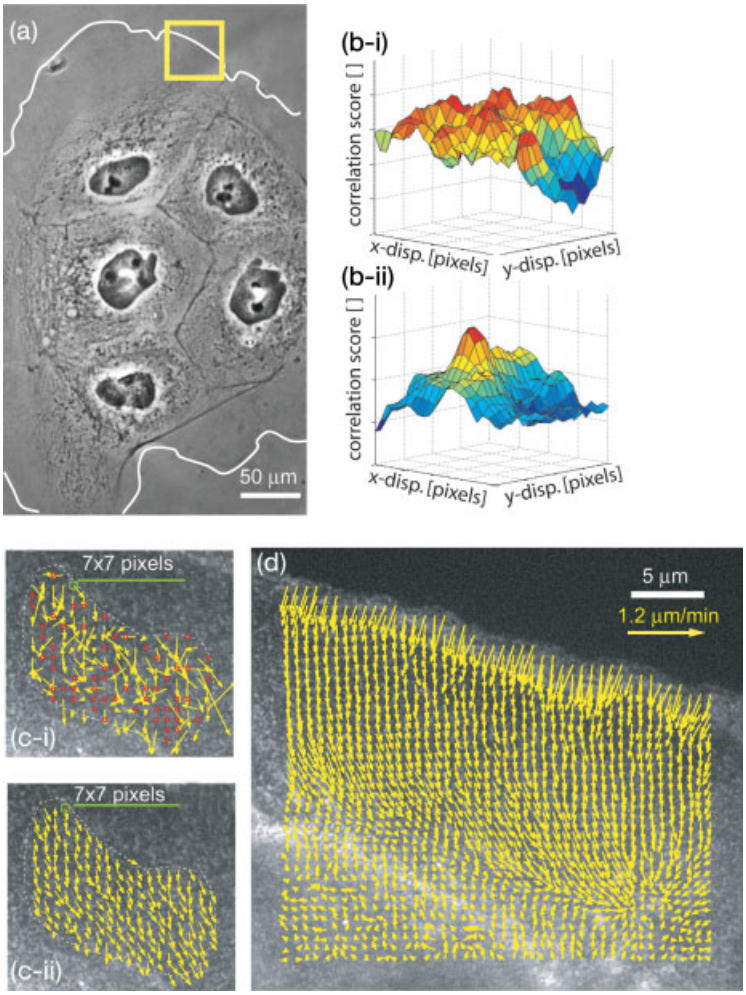


Figure 7.5 Tracking quasi-stationary speckle flow using multiframe correlation. (a) Island of epithelial cells; (b) Cross-correlation for a single frame pair (b-i) and integrated for 20 frame pairs (b-ii); (c) Region of a speckled actin network tracked with a probing window of 7×7 pixels (400×400 nm) using a single frame pair (c-i) and 20 frame pairs (c-ii); (d) Speckle flow map corresponding to the inset in (a) extracted by integration of the correlation score over 20 frame pairs. Speckle flow in this movie is almost stationary, justifying the time integration. Figure reproduced with permission from Ref. [77].

7.7.2
Tracking Single-Speckle Trajectories

The extraction of kinetic data according to Figure 7.3 requires the accurate localization of speckle birth and death events. For this, it was necessary to devise methods capable of tracking full trajectories at the single-speckle level. The large number

(>100 000) of dense speckles poses a significant challenge. Details of the current implementation of single-particle tracking of speckles are described by Ponti *et al.* [78]. Our approach follows the framework of most particle-tracking methods – that is, the detection of speckles as particles on a frame-by-frame basis, and the subsequent assignment of corresponding particles in consecutive frames. Assignment is iterated to close gaps in the trajectories created by the short-term instability of the speckle signal. Our implementation of this framework included two algorithms that address particularities of the speckle signal:

- Speckles are detected in an iterative statistical framework, which accounts for signal overlap between proximal speckles.
- Speckle assignments between consecutive frames are executed in a hybrid approach combining speckle flow and single-speckle tracking.

Speckle flow fields are extracted iteratively from previous solutions of single-speckle trajectories [78], or by initial correlation-based tracking [77]. The fields are then employed to propagate speckle motion from the source to the target frame, prior to establishing the correspondence between the projected speckle position and the effective speckle position in the target frame by global nearest-neighbor assignment [79, 80].

Motion propagation allows us to cope with two problems of FSM data. First, in many cases the magnitude of speckle displacements between two frames significantly exceeds half the distance between speckles. Hence, no solution to the correspondence problem exists without prediction of future speckle locations. Second, speckles undergo sharp spatial gradients in speed and direction of motion. A global propagation scheme discarding regional variations will thus fail, whereas an iterative extraction of the flow field permits a gradually refined trajectory reconstruction in these areas.

Figure 7.6a displays the single-speckle trajectories for speckles initiated in the first 20 frames of the same movie for which speckle flow computation is demon-

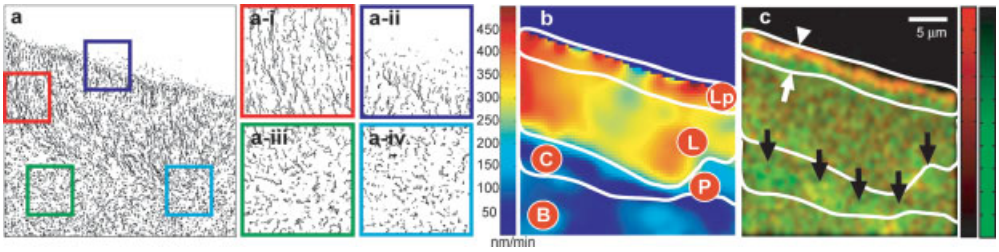


Figure 7.6 Tracking single-speckle trajectories. (a) Trajectories of speckles initiated in the first 20 frames of an actin FSM movie. (a-i to a-iv) Close-ups in different areas indicating regional variation in directional persistence, velocity and lifetime of the trajectories; (b) Speed distribution averaged over all 220 frames of the movie; (c) Distribution of polymerization (red channel) and depolymerization (green channel) calculated from scores averaged over 220 frames. Four regions of the actin network with distinct kinematic (motion) and kinetic (turnover) properties can be segmented (see text). Figure reproduced with permission from Ref. [23].

strated in Figure 7.5. The color-framed close-ups indicate regional differences between trajectories. Window (*a-i*) contains mostly straight trajectories with an average lifetime of 88 s; the trajectories in window (*a-ii*) are also straight, with an average lifetime of 60 s. In contrast, trajectories in windows (*a-iii*) and (*a-iv*) exhibit less directional persistence and have average lifetimes of 65 s and 59 s, respectively. It was concluded from such data that the F-actin cytoskeleton is regulated in a regionally variable fashion.

Figure 7.6b and c present the steady-state speed of actin network transport and turnover extracted from ∼100 000 trajectories. Three different patterns of turnover are recognized that correspond to regions with different average speeds. At the cell edge, a ∼1 μm-wide band of network assembly (red color; white arrowhead) abuts a ∼1 μm-wide band of disassembly (green color; white arrow). The yellow shade in the assembly band indicates that filament polymerization and depolymerization significantly overlap. This 2 μm-wide cell border, which is referred to as the lamellipodium (Lp), exhibits on average the fastest F-actin retrograde flow. Predominant disassembly is found ∼10 μm from the cell edge (black arrows), where the speed of F-actin flow is minimal. Here, the retrograde flow of the cell front encounters the anterograde flow of the cell body (B); this region is thus called the 'convergence zone' (C). Between the lamellipodium and the convergence zone is a region called the lamella (L), where assembly and disassembly alternate in a random pattern, accompanied by relatively coherent retrograde flow of moderate speed. The same pattern of network turnover is observed in the cell body.

7.7.3
Mapping Polymer Turnover Without Speckle Trajectories

It frequently occurs that a lower speckle contrast or a high image noise does not allow the precise identification of single-speckle trajectory endpoints. However, the trackable subsections of the trajectories are usually sufficient to extract the overall structure of speckle flow. In this case, an alternative scheme relying on the continuity of the optical density of the speckle field permits the mapping of turnover at lower resolution [81], as indicated in Figure 7.7 for the example of a crawling fish keratocyte, a cell system where the generation of clear fluorescent speckle patterns has proven difficult [52].

7.8
Applications of FSM for Studying Protein Dynamics *In Vitro* and *In Vivo*

Applications of FSM have thus far focused mostly on the study of F-actin and MT cytoskeleton systems, although other systems have also been analyzed in this way. A

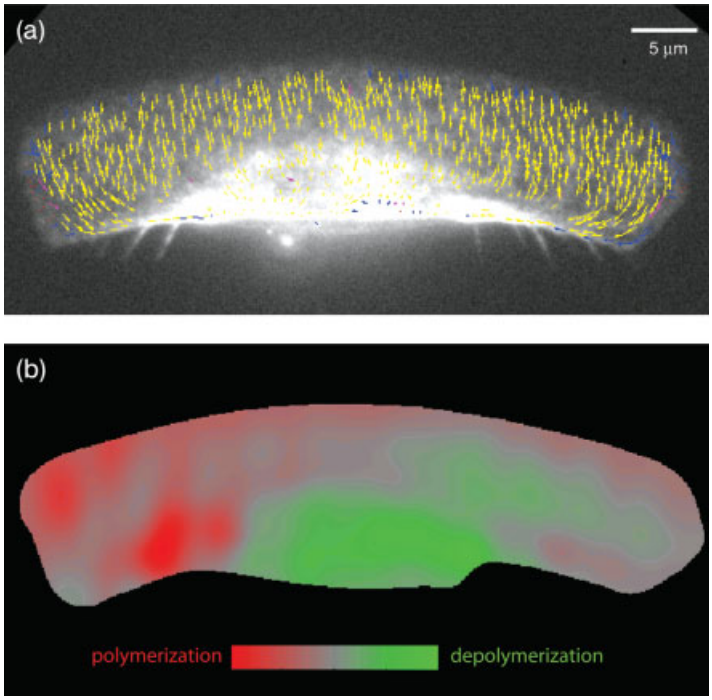


Figure 7.7 Reconstruction of F-actin network turnover from speckle flow in a fish keratocyte with poor speckle contrast. (a) Flow is calculated using multiframe correlation (as in Figure 7.5); (b) The calculated turnover map is based on conservation of image intensity. Without explicit identification of speckle births and deaths, essential details in the fine structure of the network turnover are lost.

summary of the FSM literature can be found in Table 7.1, where the major biological findings and technical advances in FSM made to date are listed. Most of the FSM data analysis has been limited to kymograph measurements of average speckle flow (see Section 7.7) and to the manual tracking of a few hundred speckles to extract lifetime information [48] and selected trajectories of cytoskeleton structures [26, 27, 69]. A systematic analysis of the full spatiotemporal information offered by FSM regarding transport and turnover in molecular assemblies is far from complete, although significant progress has already been made. In the following section, we describe comprehensive qFSM analyses both of F-actin cytoskeleton dynamics in migrating epithelial cells, and of the architectural dynamics of the spindle. Some of the most interesting results of these studies, showcasing the technical possibilities of qFSM, are also summarized.

Table 7.1 Selection of fluorescent speckle microscopy (FSM) general references and biological findings by subject.

Previous FSM Reviews and Methods Chapters

Previous reviews:

- Speckle image formation and the first FSM applications studying microtubule and F-actin dynamics *in vitro* and *in vivo*[a,b]
- Introduction to the 'platform concept' of FSM, where labeled subunit association with and dissociation from a macromolecular structure produces a stochastic signal indicative of turnover and movement of the structures[c]
- Algorithms development for quantitative FSM (through 2003)[d]
- FSM imaging of cytoskeleton dynamics in neurons[e]
- Comparison of the FSM signal in wide-field epifluorescence and total internal reflection fluorescence (TIRF) microscopy[f]
- FSM imaging and signal analysis, history of biological questions, and corresponding methodological advances[g]

Methods:

- Methods for microscope set-up and cell preparations for FSM imaging in living cells[h,i,j,k,j]
- Method of intracellular force reconstruction by FSM analysis[m]

FSM Analysis of Microtubule (MT) Motion and Assembly Dynamics

MTs assembled from pure tubulin (in vitro):

- MTs assembled from labeled and unlabeled pure tubulin *in vitro* exhibit fluorescent speckles, showing that cellular factors or organelles do not contribute to the speckle pattern[n]
- Speckles containing one fluorophore can be detected using conventional wide-field epifluorescence[o]
- For MTs *in vitro*, treadmilling is not unidirectional, suggesting that it is powered by differences in dynamic instability between plus and minus ends[p]
- The visualization of individual speckled MTs in a pool of tubulin, even when labeled at ratios <2%, requires a reduction of out-of-focus fluorescence by spinning-disk confocal microscopy[p]
- The KinI subfamily of kinesin-related proteins mediates depolymerization of MTs at both ends *in vitro*[q]

MT flux in meiotic spindles assembled from Xenopus laevis *egg extracts (in vitro):*

- FSM allows detailed analysis of dense polymers like the spindle[a]
- Spinning disk FSM reveals MT bundles, whereas in wide-field FSM those bundles are not detectable[r]
- MTs both polymerize and depolymerize at the kinetochores[r]
- Flux rates are different for kinetochore and nonkinetochore MT bundles[r,s]
- Monopoles do not exhibit MT flux during spontaneous bipolarization, and the onset of flux is correlated with the onset of bipolarity. This suggests that arrays of antiparallel MTs are required for flux generation[t]
- Disruption of depolymerization factors at the poles yields kinesin Eg5-dependent elongation of the metaphase spindle[u,v]
- MT flux is predominantly driven by ensembles of processive kinesin Eg5 motors[w]
- FSM and cross-correlation analysis reveal MT minus ends throughout the spindle, with more at the poles than at the spindle equator[x]
- MTs within bundles move at heterogeneous speeds, and FSM can be used to determine the length distribution of individual MTs in the spindle[y]

(*Continued*)

Table 7.1 (*Continued*)

MT flux in mitotic spindles in tissue culture cells (in vivo):
- The majority of poleward flux of kinetochore MTs in mammalian PtK1 epithelial cells is driven by a polar pulling-in mechanism, whereas Eg5, which plays a dominant role in *Xenopus* egg extracts, makes a minor contribution[z]

MT flux in spindles of other cell systems:
- MT flux makes a significant contribution to poleward chromosome movement during anaphase A in *Drosophila melanogaster* embryos[aa,bb]
- Three mitotic motors exhibit different roles in anaphase B in *Drosophila* embryos[cc]
- Two functionally distinct MT-destabilizing *Kin*I enzymes are responsible for normal chromatid-to-pole motion in *Drosophila*[dd]
- MT flux in crane-fly spermatocytes increases from metaphase to anaphase and is faster than chromosome poleward motion, suggesting that MT plus ends are still polymerizing[ee]

MTs in interphase tissue cells:
- In living cells, optimal speckle contrast occurs at fractions of labeled tubulin in the 0.1–0.5% range, where the fluorescence of each speckle corresponds to one to seven fluorophores per resolvable unit[n,o]
- Cytoplasmic dynein, a MT-associated motor, promotes the formation and growth of immobile MTs in organized astral arrays, as opposed to organizing the array by powering the motion of pre-existing polymers[ff]
- Overexpression of Ncd, a kinesin-14, in mammalian fibroblasts results in generation of sliding forces between adjacent MTs in bundles[gg]

MTs in neurons:
- In the axon shaft proximal to the cell body, individual MTs are stationary, suggesting that tubulin dimers are transported down the axon to promote axonal growth and branching[hh]
- In regions of axon growth (i.e. growth cones and interstitial branches), short segments of MTs move, suggesting that exploratory behavior of neurons is promoted by MT transport[ii]
- Measurement of MT growth and transport in growth cones reveals they grow towards the periphery while being transported towards the axon[jj,kk]

MT dynamics in S. cerevisiae *and* S. pombe*:*
- During mating of S. *cerevisiae*, the nucleus and spindle pole body are oriented and tethered to the shmoo tip by a MT-dependent search and capture mechanism, where MT growth and shrinkage are localized mostly to the shmoo tip[ll]
- Kinetochore MTs grow and shrink only at the plus ends and do not exhibit poleward flux[mm]
- Astral MT plus end growth and shortening at the cell cortex plays an important role in positioning the nucleus during interphase and the spindle during mitosis[nn]

MT dynamics in pathogens:
- Dynamics of MTs in the fungal pathogen *Ustilago maydis* determines cell polarity[oo]
- MT dynamics through the cell cycle alter morphogenesis in the fungal pathogen *Candida albicans*[pp]

FSM Analysis of Microtubule-Associated Protein (MAP) Dynamics

MT plus-end binding proteins:
- CLIP-170 binds at the growing MT end, stays stationary relative to the MT, and dissociates after some time[qq]
- MT assembly in meiotic *Xenopus* egg extract spindles is visualized by localization of speckle-like EB1 comets[rr]

MT motors:
- The motor protein Eg5 stays stationary in spindles despite the flux of MTs, suggesting that Eg5 may be bound to a putative non-MT spindle matrix[ss]

(*Continued*)

Table 7.1 (*Continued*)

Other MAPs:
- Co-imaging of full-length ensconsin or its MT-binding domain (EMTB) conjugated to GFP with fluorescent MTs suggests that dynamics of MAP:MT interactions is at least as rapid as tubulin: MT dynamics in the polymerization reaction[tt]
- Binding and unbinding of ensconsin generates a speckle pattern along the MT, the dynamics of which can be evaluated to study the phosphorylation-dependent regulation of the turnover[uu]
- Multi-GFP tandems on MAPs significantly increase the speckle contrast and stability[uu]

FSM Analysis of F-actin Dynamics in Migrating and Non-migrating Cells

Actin network dynamics in migrating tissue cells:
- F-actin in polarized cells is organized in four distinct zones: a lamellipodium with rapid retrograde flow and constant polymerization; a lamella with slower retrograde flow, a contraction zone with no flow, and a zone of anterograde flow. The spatial transition from retrograde to anterograde flow suggests the presence of a contractile belt powered by myosin II which may drive cell migration[vv]
- Single-fluorophore speckles can reveal F-actin turnover, as known from speckle analysis in MTs, despite the complex filamentous structure of the F-actin meshwork[ww]
- Statistical clustering analysis of single speckle dynamics reveals two kinetically and kinematically distinct, yet spatially overlapping, actin networks that mediate cell protrusion[xx]
- Spatiotemporal correlation of F-actin assembly maps and GFP-Arp2/3 clustering indicate that, in the lamellipodium, actin assembly is mediated by Arp2/3, while lamellar assembly is independent of Arp2/3 activity[yy]
- Arp2/3- and cofilin-regulated assembly of the lamellipodia is not required for epithelial cell protrusion[zz]
- Molecular kinetics of Arp2/3 and capping protein can be measured by single-molecule FSM[aaa]
- The dynamics of actin-binding proteins (capping protein, Arp2/3, tropomyosin) exhibit spatial differentiation in the lamellipodium and lamella of *Drosophila* S2 cells[bbb]

Actin network dynamics in neurons:
- Steady-state retrograde flow in neuronal growth cones depends on both myosin II contractility and actin-network treadmilling[ccc]

Actin networks in keratocytes and keratocyte fragments:
- Mechanical stimulation of keratocyte fragments activates acto-myosin contraction and causes directional motility[ddd]
- Keratocytes exhibit F-actin retrograde flow relative to the substrate[eee] in a biphasic relationship between flow magnitude and adhesiveness[fff]
- Actin and myosin undergo polarized assembly, suggesting force generation occurs at the lamellipodium/cell body transition zone[ggg]
- Directed motility is initiated by symmetry breaking actin-myosin network reorganization and contractility at the cell rear[hhh]

Actin in contact-inhibited epithelial cells:
- Unlike migrating cells, cortical actin in contact-inhibited cells is spatially stationary but undergoes rapid turnover[iii, jjj]
- The spatiotemporal mapping of F-actin network turnover from speckle signal analysis during appearance and disappearance events can be carried out at high resolution in contact-inhibited cells[jjj]

Actin dynamics in S. cerevisiae, *using GFP-tubulin:*
- FSM can be used to visualize bud-associated assembly and motion of F-actin cables in budding yeast[kkk]

Multi-Spectral FSM Analysis of F-actin and Other Macromolecular Structures

(*Continued*)

Table 7.1 (*Continued*)

Co-motion of the F-actin cytoskeleton and other structures, using spectrally distinct fluorescent analogues:

- FSM of MTs and F-actin in cytoplasmic extracts of *Xenopus* eggs confirms two basic types of interaction between the polymers: a cross-linking activity and a motor-mediated interaction[lll]
- Dynamic interactions between MTs and actin filaments are required for axon branching and directed axon outgrowth[mmm]
- Direction of MT growth is guided by the tight association of MTs with F-actin bundles[nnn]
- F-actin contraction may be involved in the breaking of MTs[vv,ooo]
- Rho and Rho effectors have differential effects on F-actin and MT dynamics during growth cone motility[ppp]
- In migrating epithelial cells, the dynamics of MTs and F-actin is coordinated by signaling pathways downstream of Rac1[qqq]
- FSM of F-actin and several focal adhesion (FA) proteins reveals a differential transmission of F-actin network motion through the adhesion structure to the extracellular matrix[rrr]

FSM Analysis of Protein Turnover in Focal Adhesions (FAs)

- FSM of low-level GFP-fusion protein expression, in combination with TIRF microscopy, allows quantification of molecular dynamics within FA protein assemblies at the ventral surface of living cells[f,rrr]

[a]Waterman-Storer, C.M., Desai, A., Bulinski, J.C. and Salmon E.D. (1998) *Curr. Biol.*, **8**, 1227.
[b]Keating, T.J. and Borisy, G.G. (2000) *Curr Biol.*, **10**, R22.
[c]Waterman-Storer, C.M. and Danuser, G. (2002) *Curr. Biol.*, **12**, R633.
[d]Danuser, G. and Waterman-Storer, C.M. (2003) *J. Microsc.*, **211**, 191.
[e]Dent, E.W. and Kalil, K. (2003) in *Methods in Enzymology*, Vol. 361, Academic Press, San Diego, pp. 390.
[f]Adams, M., Matov, A., Yarar, D., Gupton, S., Danuser, G. and Waterman-Storer, C.M. (2004) *J. Microsc.*, **216**, 138.
[g]Danuser, G. and Waterman-Storer, C.M. (2006) *Annu. Rev. Biophys. Biomol. Struct.*, **35**, 361.
[h]Adams, M.C., Salmon, W.C., Gupton, S.L., Cohan, C.S., Wittmann, T., Prigozhina, N. and Waterman-Storer, C.M. (2003) *Methods*, **29**, 29.
[i]Waterman-Storer, C.M. (2002) in *Current Protocols in Cell Biology* (eds J.S. Bonifacino, M. Dasso, J.B. Harford, J. Lippincott-Schwartz and K.M. Yamada), Wiley, New York.
[j]Maddox, P.S., Moree, B., Canman, J.C. and Salmon, E.D. (2003) *Methods in Enzymology*, **360**, 597.
[k]Gupton, S.L. and Waterman-Storer, C.M. (2006) in *Cell Biology: A Laboratory Handbook*, Vol. 3, 3 edn. (eds J. Celis, N. Carter, K. Simons, J.V. Small, T. Hunter and D. Shotton), Academic Press, San Diego, pp. 137.
[l]Waterman-Storer, C., Desai, A. and Salmon, E.D. (1999) in *Methods in Cell Biology*, Vol. 61, p. 155.
[m]Ji, L., Loerke, D., Gardel, M. and Danuser, G. (2007) in *Methods in Cell Biology*, Vol. 83.
[n]Waterman-Storer, C.M. and Salmon, E.D. (1998) *Biophys. J.*, **75**, 2059.
[o]Waterman-Storer, C.M. and Salmon, E.D. (1999) *FASEB J*, **13**, 225.
[p]Grego, S., Cantillana, V. and Salmon, E.D. (2001) *Biophys. J.*, **81**, 66.
[q]Hunter, A.W., Caplow, M., Coy, D.L., Hancock, W.O., Diez, S., Wordeman, L. and Howard, J. (2003) *Molecular Cell*, **11**, 445.
[r]Maddox, P., Straight, A., Coughlin, P., Mitchison, T.J. and Salmon, E.D. (2003) *J. Cell Biol.*, **162**, 377.
[s]Vallotton, P., Ponti, A., Waterman-Storer, C.M. Salmon, E.D. and Danuser, G. (2003) *Biophys. J.*, **85**, 1289.
[t]Mitchison, T.J., Maddox, P., Groen, A., Cameron, L., Perlman, Z., Ohi, R., Desai, A., Salmon, E.D. and Kapoor, T.M. (2004) *Mol. Biol. Cell*, **15**, 5603.
[u]Shirasu-Hiza, M., Perlman, Z.E., Wittmann, T., Karsenti, E. and Mitchison, T.J. (2004) *Curr. Biol.*, **14**, 1941.
[v]Gaetz, J. and Kapoor, T.M. (2004) *J. Cell Biol.*, **166**, 465.
[w]Miyamoto, D.T., Perlman, Z.E., Burbank, K.S., Groen, A.C. and Mitchison, T.J. (2004) *J. Cell Biol.*, **167**, 813.
[x]Burbank, K.S., Groen, A.C., Perlman, Z.E., Fisher, D.D. and Mitchison, T.J. (2006) *J. Cell Biol.*, **175**, 369.

[y]Yang, G., Houghtaling, B.R., Gaetz, J., Liu, J.Z., Danuser, G. and Kapoor, T.M. (2007) *Nat. Cell Biol.*, **9**, 1233.
[z]Cameron, L.A., Yang, G., Cimini, D., Canman, J.C., Evgenieva, O.K., Khodjakov, A., Danuser, G. and Salmon, E.D. (2006) *J. Cell Biol.*, **173**, 173.
[aa]Maddox, P., Desai, A., Oegema, K., Mitchison, T.J. and Salmon, E.D. (2002) *Curr. Biol.*, **12**, 1670.
[bb]Brust-Mascher, I. and Scholey, J.M. (2002) *Mol. Biol. Cell*, **13**, 3967.
[cc]Brust-Mascher, I., Civelekoglu-Scholey, G., Kwon, M., Mogilner, A. and Scholey, J.M. (2004) *Proc. Natl Acad. Sci. USA*, **101**, 15938.
[dd]Rogers, G.C., Rogers, S.L., Schwimmer, T.A., Ems-McClung, S.C., Walczak, C., Vale, R.D., Scholey, J.M. and Sharp, D.J. (2004) *Nature*, **427**, 364.
[ee]LaFountain, J.R. Jr., Cohan, C.S., Siegel, A.J. and LaFountain, D.J. (2004) *Mol. Biol. Cell*, **15**, 5724.
[ff]Vorobiev, I., Malikov, V. and Rodionov, V. (2001) *Proc. Natl Acad. Sci. USA*, **98**, 10160.
[gg]Oladipo, A., Cowan, A. and Rodionov, V. (2007) *Mol. Biol. Cell*, **18**, 3601.
[hh]Chang, S., Svitkina, T.M., Borisy, G.G. and Popov, S.V. (1999) *Nat. Cell Biol.*, **1**, 399.
[ii]Dent, E.W., Callaway, J.L., Szebenyi, G., Baas, P.W. and Kalil, K. (1999) *J. Neurosci.*, **19**, 8894.
[jj]Kabir, N., Schaefer, A.W., Nakhost, A., Sossin, W.S. and Forscher, P. (2001) *J. Cell Biol.*, **5**, 1033.
[kk]Zhou, F.-Q., Waterman-Storer, C.M. and Cohan, C.S. (2002) *J. Cell Biol.* **2002**, **157**, 839.
[ll]Maddox, P., Chin, E., Mallavarapu, A., Yeh, E., Salmon, E.D. and Bloom, K. (1999) *J. Cell Biol.*, **144**, 977.
[mm]Maddox, P.S., Bloom, K.S. and Salmon, E.D. (2000) *Nat. Cell Biol.*, **2**, 36.
[nn]Tran, P.T., Marsh, L., Doye, V., Inoue, S. and Chang, F. (2001) *J. Cell Biol.*, **153**, 397.
[oo]Steinberg, G., Wedlich-Soldner, R., Brill, M. and Schulz, I. (2001) *J. Cell Sci.*, **114**, 609.
[pp]Finley, K.R. and Berman, J. (2005) *Eukaryotic Cell*, **4**, 1697.
[qq]Perez, F., Diamantopoulos, G.S., Stalder, R. and Kreis, T.E. (1999) *Cell*, **96**, 517.
[rr]Tirnauer, J.S., Salmon, E.D. and Mitchison, T.J. (2004) *Mol. Biol. Cell*, **15**, 1776.
[ss]Kapoor, T.M. and Mitchison, T.J. (2001) *J. Cell Biol.*, **154**, 1125.
[tt]Faire, K., Waterman-Storer, C.M., Gruber, D., Masson, D., Salmon, E.D. and Bulinski, J.C. (1999) *J. Cell Sci.*, **112**, 4243.
[uu]Bulinski, J.C., Odde, D.J., Howell, B.J., Salmon, T.D. and Waterman-Storer, C.M. (2001) *J. Cell Sci.*, **114**, 3885.
[vv]Salmon, W.C., Adams, M.C. and Waterman-Storer, C.M. (2002) *J. Cell Biol.*, **158**, 31.
[ww]Watanabe, Y. and Mitchison, T.J. (2002) *Science* **2002**, **295**, 1083.
[xx]Ponti, A., Machacek, M., Gupton, S.L., Waterman-Storer, C.M. and Danuser, G. (2004) *Science*, **305**, 1782.
[yy]Ponti, A., Matov, A., Adams, M., Gupton, S. Waterman-Storer, C.M. and Danuser, G. (2005) *Biophys. J.*, **89**, 3456.
[zz]Gupton, S.L., Anderson, K.L., Kole, T.P., Fischer, R.S., Ponti, A., Hitchcock-DeGregori, S.E., Danuser, G., Fowler, V.M., Wirtz, D., Hanein, D. and Waterman-Storer, C.M. (2005) *J. Cell Biol.*, **168**, 619.
[aaa]Miyoshi, T., Tsuji, T., Higashida, C., Hertzog, M., Fujita, A., Narumiya, S., Scita, G. and Watanabe, N. (2006) *J. Cell Biol.*, **175**, 947.
[bbb]Iwasa, J.H. and Mullins, R.D. (2007) *Curr. Biol.*, **17**, 395.
[ccc]Medeiros, N.A., Burnette, D.T. and Forscher, P. (2006) *Nat. Cell Biol.*, **8**, 215.
[ddd]Verkhovsky, A.B., Svitkina, T.M., Borisy, G.G. (1999) *Curr. Biol.*, **9**, 11.
[eee]Vallotton, P., Danuser, G., Bohnet, S., Meister, J.J. and Verkhovsky, A. (2005) *Mol. Biol. Cell*, **16**, 1223.
[fff]Jurado, C., Haserick, J.R. and Lee, J. (2005) *Mol. Biol. Cell*, **16**, 507.
[ggg]Schaub, S., Bohnet, S., Laurent, V.M., Meister, J.-J. and Verkhovsky, A.B. (2007) *Mol. Biol. Cell*, E06.
[hhh]Yam, P.T., Wilson, C.A., Ji, L., Hebert, B. Barnhart, E.L., Dye, N.A., Wiseman, P.W., Danuser, G. and Theriot, J.A. (2007) *J. Cell Biol.*, **178**, 1207.
[iii]Waterman-Storer, C.M., Salmon, W.C. and Salmon, E.D. (2000) *Mol. Biol. Cell* **2000**, **11**, 2471.
[jjj]Ponti, A., Vallotton, P., Salmon, W.C., Waterman-Storer, C.M. and Danuser, G. (2003) *Biophys. J.*, **84**, 3336.
[kkk]Yang, H.-C. and Pon, L.A. (2002) *Proc. Natl Acad. Sci. USA*, **99**, 751.
[lll]Waterman-Storer, C., Duey, D.Y., Weber, K.L., Keech, J., Cheney, R.E., Salmon, E.D. and Bement, W. M. (2000) *J. Cell Biol.*, **150**, 361.
[mmm]Dent, E.W. and Kalil, K. (2001) *J. Neurosci.*, **15**, 9757.
[nnn]Schaefer, A.W., Kabir, N. and Forscher, P. (2002) *J. Cell Biol.*, **158**, 139.

[ooo]Gupton, S.L., Salmon, W.C. and Waterman-Storer, C.M. (2002) *Curr. Biol.*, **12**, 1891.
[ppp]Zhang, X.-F., Schaefer, A.W., Burnette, D.T., Schoonderwoert, V.T. and Forscher, P. (2003) *Neuron*, **40**, 931.
[qqq]Wittmann, T., Bokoch, G.M. and Waterman-Storer, C.M. (2003) *J. Cell Biol.*, **161**, 845.
[rrr]Hu, K., Ji, L., Applegate, K., Danuser, G. and Waterman-Storer, C.M. (2007) *Science*, **315**, 111.

7.9
Results from Studying Cytoskeleton Dynamics

7.9.1
F-Actin in Cell Migration

Over the past few years, qFSM has critically driven our understanding of actin cytoskeleton dynamics in cell migration. It has provided unprecedented details of the spatial organization of F-actin turnover and the transport and deformation of F-actin networks *in vivo*. In the following sections, we review a few key discoveries, enabled by qFSM, that have defined a new paradigm for the functional linkage between actin cytoskeleton regulation and epithelial cell migration.

7.9.1.1 F-Actin in Epithelial Cells is Organized Into Four Dynamically Distinct Regions

Figures 7.5 and 7.6 indicate the steady-state organization of the F-actin cytoskeleton in four kinematically and kinetically distinct zones:

- The lamellipodium, characterized by fast retrograde flow and two narrow bands of assembly and disassembly resulting from the fast treadmilling of actin between its polymeric and monomeric states [54, 82].
- The lamella, characterized by reduced retrograde flow and assembly and disassembly in random punctate patterns.
- The cell body, characterized by anterograde flow and turnover patterns similar to those of the lamella.
- The convergence zone, where the flows of the lamella and cell body meet and where strong depolymerization suggests that the lamella and cell body are materially separate structures.

qFSM also delivers nonsteady-state measurements of flow and turnover, revealing distinct variations in the periodicity of turnover between these regions [78]. In combination with pharmacological perturbation, qFSM was used to dissect the mechanisms of retrograde flow. Thus, lamellipodium flow was found to be independent of myosin motor contraction, whereas lamella flow was blocked by the specific inhibition of myosin II activity [83]. The lamellipodium and the lamella also exhibited different sensitivity to the disruption of filament assembly, disassembly and severing, which suggested that these regional differences might be associated with differential molecular regulation [83]. This hypothesis has thus far been confirmed by immunostaining studies [83, 84] and by the expression of constitutively active and dominant-negative constructs of regulatory proteins [84, 85]. Here, qFSM provides a critical insight into cytoskeleton dynamic responses to shifted activation of regulatory factors.

In summary, these data demonstrate how qFSM can be used to quantify spatiotemporal modulations of the kinetics and kinematics of molecular assemblies and to identify dynamically distinct structural modules, even when they are composed of the same base protein.

7.9.1.2 Actin Disassembly and Contraction are Coupled in the Convergence Zone

A similar spatiotemporal correlation analysis was performed to examine the relationship of F-actin network depolymerization and contraction in the convergence zone [81]. It was first established that transient increases in speckle flow convergence are coupled to transient increases in disassembly. This begged the question whether the rate of speckle flow convergence increases because disassembly boosts the efficiency of myosin II motors in contracting a more compliant network, or because motor contraction mediates network disassembly. To address this question, we transiently perfused cells with Calyculin A, a type II phosphatase inhibitor that increases myosin II activity. Unexpectedly, we reproducibly measured a strong burst of disassembly long before flow convergence was affected. This evidence suggested that myosin II contraction can actively promote the depolymerization of F-actin, for example, by breaking filaments. The link between F-actin contractility and turnover has since been confirmed by fluorescence recovery after photobleaching measurements in the contractile ring required for cytokinesis [86]. In summary, these data demonstrate the correlation of two qFSM parameters to decipher the relationship between deformation and plasticity of polymer networks inside cells.

7.9.1.3 Two Distinct F-Actin Structures Overlap at the Leading Edge

The transition between the lamellipodium and the lamella is characterized by a narrow band of strong disassembly adjacent to a region of mixed assembly and disassembly and a sharp decrease in retrograde flow velocity (see Figure 7.6). Together, these features defined a unique mathematical signature for tracking the boundary between the two regions over time (Figure 7.8a). In view of the different speckle velocities and lifetimes between the two regions, it was speculated that the same boundary could be tracked by spatial clustering of speckle properties. It was predicted that fast, short-living speckles (class 1) would preferentially localize in the lamellipodium, whereas slow, longer-living speckles (class 2) would be dominant in the lamella. To test this hypothesis, we solved a multiobjective optimization problem in which the thresholds of velocity v_{th} and lifetime τ_{th} separating the two classes, as well as the boundary ∂Lp between lamellipodium and lamella, were determined subject to the rule $\{\partial \mathrm{Lp}, \tau_{th}, v_{th}\} = \max(N_1/(N_1 + N_2) \in \mathrm{Lp}) \& \min(N_1/(N_1 + N_2) \in \mathrm{La})$ (Figure 7.8b and c), where N_1 and N_2 denote the number of speckles in classes 1 and 2, respectively. The prediction was confirmed in the lamella, where class 1 speckles occupied a statistically insignificant fraction. However, class 2 speckles made up 30–40% of the lamellipodium, indicating that in this region speckles with different kinetic and kinematic behavior colocalize. This information was previously lost in the averaged analysis of single-speckle trajectories. When mapping the scores of class 1 and class 2 speckles separately (Figure 7.8di-dii), it was discovered that class 1 speckles define the bands of polymerization and depolymerization characteristic

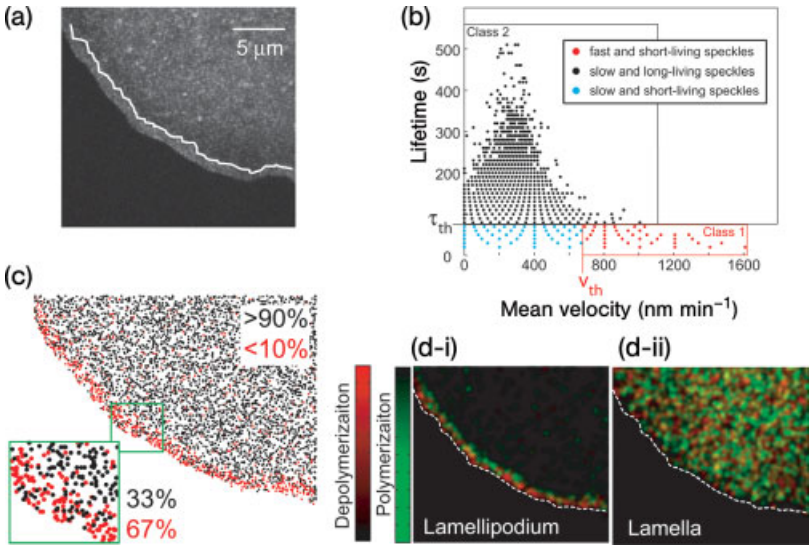


Figure 7.8 Distinction of two spatially overlapping actin networks based on heterogeneity of single-speckle properties. (a) Raw FSM image overlaid with the boundary between lamellipodium and lamella computed from spatial gradients in F-actin turnover and flow velocity; (b,c) Cluster analysis of speckle lifetime and velocity (see text); (d) Class 1 speckles constitute the rapidly treadmilling lamellipodium. Class 2 speckles constitute the lamella with a punctate pattern of random actin turnover. Both networks spatially overlap in the first 2 µm from the cell edge. Figure reproduced in parts with permission from Refs [83] and [23].

of the lamellipodium, and that class 2 speckles define the puncta of assembly and disassembly characteristic of the lamella, which reaches all the way to the leading edge. Subsequent experiments specifically disrupting actin treadmilling in the lamellipodium confirmed the finding that the lamellipodium and lamella form two spatially overlapping, yet kinetically, kinematically and molecularly different, F-actin networks [83, 84].

7.9.2 Architecture of *Xenopus laevis* Egg Extract Meiotic Spindles

During cell division, MTs form a spindle, which maintains stable bipolar attachment to chromosomes over tens of minutes. A sophisticated checkpoint system senses the status of attachment and generates a signal to progress with symmetric segregation of the replicated sister chromatids into the newly forming mother and daughter cells [6]. The minus ends of polar MTs are preferentially located at the spindle poles, whereas the plus ends continually switch between growth and shrinkage, a process known as 'MT dynamic instability' [87]. Strikingly, dynamic instability in vertebrate spindles occurs within a few tens of seconds, a time scale at least an order of magnitude shorter than the existence of the spindle [88]. In addition to MT dynamic instability at the MT plus end, individual MTs are transported toward the spindle poles, a behavior known as 'poleward flux'. Poleward flux has only been observed in higher eukaryotic

spindles, including the *Xenopus laevis* extract spindle system [89]. How the overall stability of spindle architecture is maintained under the much faster dynamics of its building blocks is largely unknown. qFSM has made several critical contributions to the mechanistic analysis of spindle architecture (Table 7.1). For example, it has revealed detailed maps of the organization of heterogeneous MT poleward flux (Figure 7.9a–c) [74, 90], and it was also used to show that MTs form distinct types of bundles with different flux dynamics, depending on whether they are attached to chromosomes (kinetochore fibers) or form a scaffold of overlapping fibers emanating from opposite poles (interpolar MT fibers) [38]. Together, these data have indicated an enormous architectural complexity, which requires fine regulation of the dynamics of each MT. However, the high MT density in the spindle has precluded measurement of the dynamics of individual MTs within bundles. Speckles generally consist of multiple fluorophores distributed over many different MTs.

Recently, this difficulty has been overcome by single-fluorophore speckle imaging of *Xenopus laevis* extract spindles [62]. As a cell-free spindle model, the extract spindle allows convenient control over fluorescent tubulin levels to achieve sparse labeling of MTs (Figure 7.9d). For low labeling ratios *f*, speckle intensities cluster in multiples of ~500 AU, indicating that speckles are composed of a discrete low number (e.g. one, two, three or four) of fluorophores (Figure 7.9e and f) [62]. At the lowest concentrations of labeled tubulin, only one intensity cluster with a mean value of ~500 AU was found. Furthermore, the average intensity of a detectable speckle remained constant over time at ~500 AU (Figure 7.9g), although the speckle number decreased due to photobleaching, Together, the cluster analysis of speckle intensities and the photobleaching analysis confirmed that >98% of the speckles reflected the image of a single fluorophore.

7.9.2.1 Individual MTs within the Same Bundle Move at Different Speeds

Single-fluorophore speckles were then used to investigate how individual MTs in close proximity move relative to one another. In order to avoid any *a priori* assumptions, the spatial organization of spindle MTs was mapped using the dense flow field measured in a spectrally distinct channel displaying multifluorophore speckles. Path integration of the flow field allowed the construction of equally spaced bands of uniform width (480 nm), which reflects the average position of MT bundles within the spindle (Figure 7.9h). The band width was chosen to match the diffraction limit of the microscope, and is consistent with electron microscopy studies which showed that MTs form bundles typically a few hundred nanometers wide, with individual MTs 20–50 nm apart [91, 92]. Next, the pairwise difference between the velocities of speckles located in the same band showed that MTs spaced at a distance comparable to the width of MT bundles exhibit remarkably heterogeneous movement (Figure 7.9i). Thus, individual MTs appear to slide past one another over very short distances, suggesting that the spindle is a MT scaffold that is continuously restructured at the scale of tens of seconds.

7.9.2.2 The Mean Length of Spindle MTs is 40% of the Total Spindle Length

Despite the heterogeneous movement of the majority of speckles within one band, a small percentage (~1% of all speckles) moved in synchronized pairs: not only did

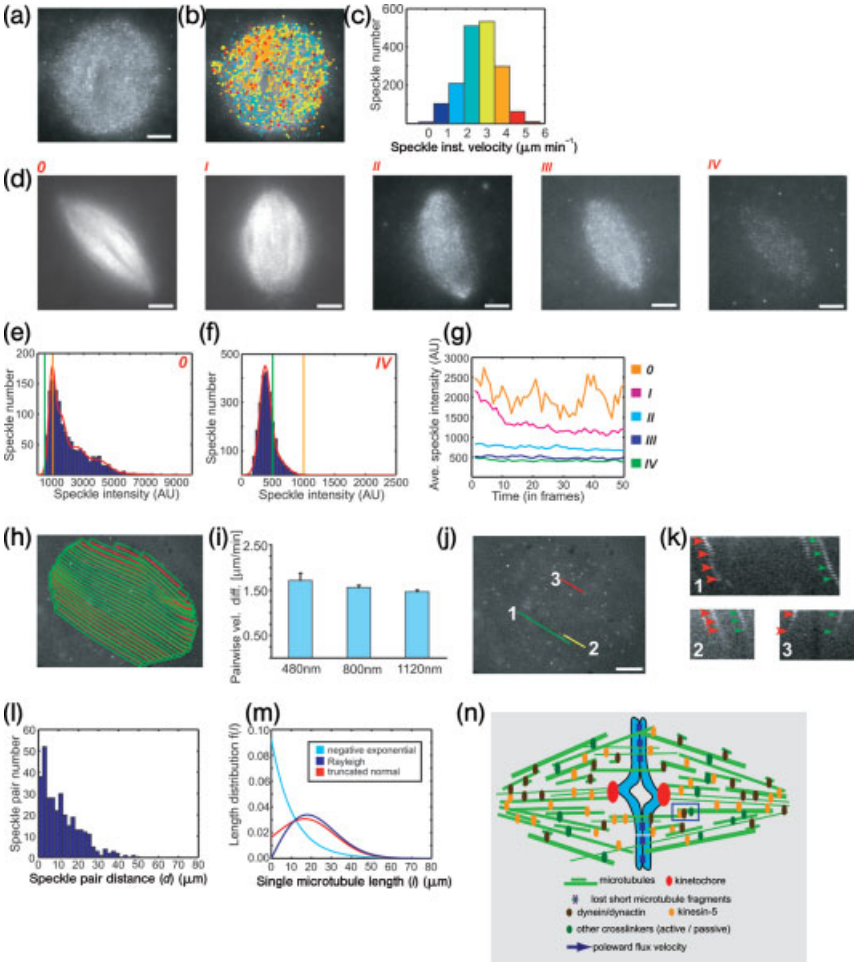


Figure 7.9 Analyzing the dynamic MT architecture of the *Xenopus laevis* egg extract spindle using single-fluorophore imaging. (a) Spindle imaged using multifluorophore tubulin speckles. Scale bar = 10 μm; (b,c) Speckle trajectories (b) overlaid onto the spindle from (a) were recovered using particle tracking. Color-coding by velocity range is specified in the velocity histogram (c). Average velocity mean ± SD = 2.68 ± 0.95 μm min^{-1} (n = 1699 tracks); (*d*) Single-fluorophore imaging conditions were determined by sequentially reducing labeled tubulin concentrations (***0***, 3.3 n*M*; ***I***, 1.1 n*M*; ***II***, 0.33 n*M*; ***III***, 0.11 n*M*; ***IV***, 0.033 n*M*; ***V***, 0.011 n*M*). Scale bars = 10 μm; (e, f) Speckle intensity distributions of spindles in (d) (concentrations ***0*** and ***V***), each fitted by a mixture of normal distributions (red lines) calculated from cluster analysis [103]. Lines mark 500 AU (green) and 1000 AU (brown); (g) Changes of average speckle intensity over time due to photobleaching within spindles in (d); (h) Based on two-color speckle imaging and path integration of the MT poleward flux vector field, uniform bands were generated at equal distances to trace MTs within the spindle [62]; (i) Difference in instantaneous velocities between pairs of speckles within bands constructed as in (h). Error bars represent standard errors of the mean (480 nm: n = 78 pairs; 800 nm: n = 659 pairs; 1120 nm: 1605 pairs); (j) Examples of synchronously moving speckle pairs identified within bands constructed as in (h). Each pair resided in the

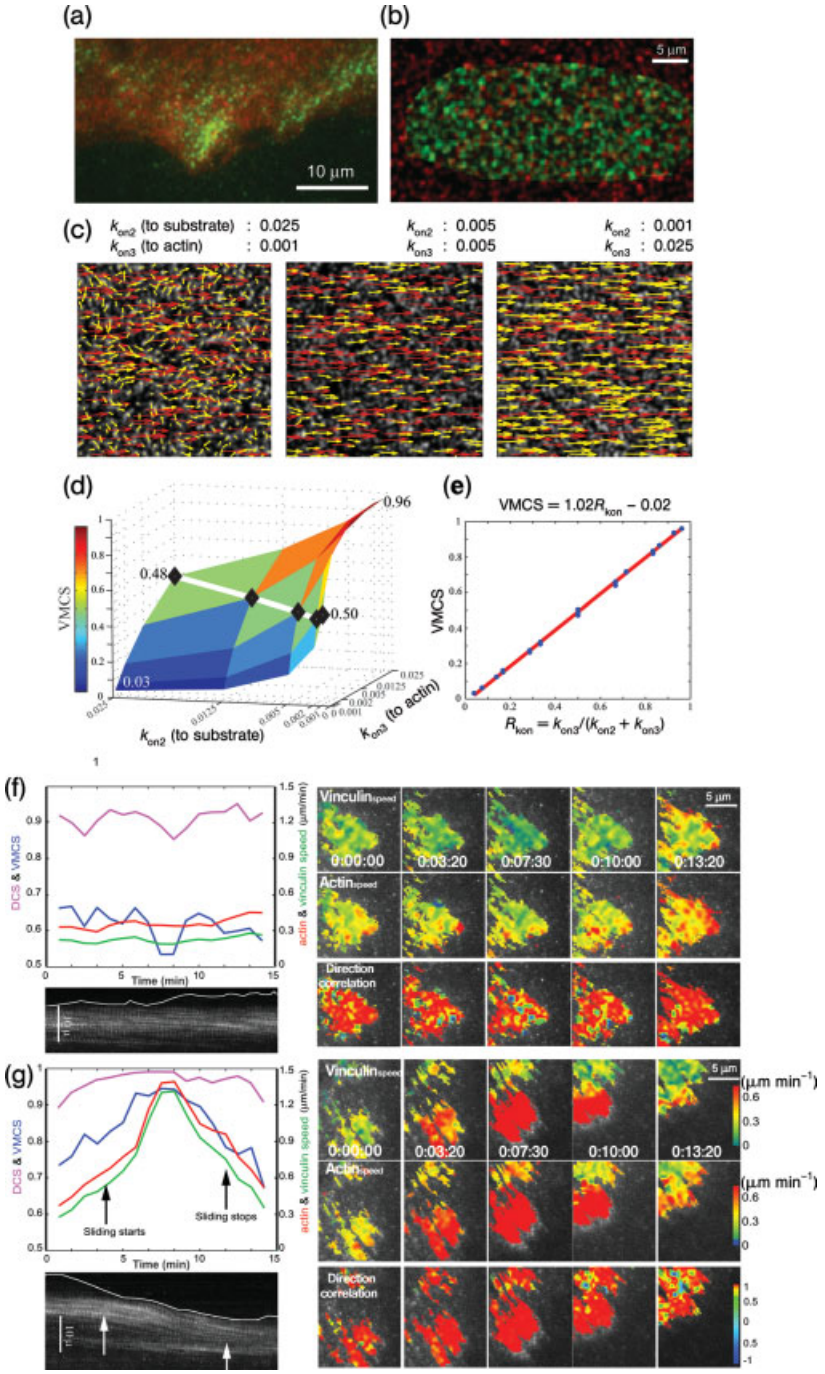
they stay within the same band and move in the same direction at the same time (Figure 7.9j and k), but they also concurrently changed velocities (Figure 7.9k, 1–3). Clearly, these single-fluorophore speckle pairs must reside on the same MT. By applying stringent detection criteria on spatial colocalization, temporal overlap, relative distance change and relative velocity change, a total of 328 synchronous speckle pairs was identified from 13 spindles. Interestingly, 90% of the speckle distances were less than half the length of the spindle (Figure 7.9l). To estimate the lengths of spindle MTs from the measured distances between synchronously moving speckle pairs, a mathematical model was developed of the stochastic incorporation of labeled tubulin into a population of MTs with an *a priori* unknown length distribution $f(l)$ [62]. Assuming a hypothetical function $f(l)$, the model defined the expected cumulative distribution P of distances d between two speckles, given the event A that the speckles reside on the same MT:

$$P(D<d|A) = \frac{\int_0^d l^2 e^{-c\cdot l\cdot r} f(l)dl + \int_d^{+\infty}(2dl-d^2)e^{-c\cdot l\cdot r} f(l)dl}{\int_0^{+\infty} u^2 e^{-c\cdot u\cdot r} f(u)du}.$$

Here, l denotes the steady-state length of individual MTs in microns, c is the number of tubulin dimers per micron (1625), and r is the fraction of labeled tubulin ($\sim 2.86\times 10^{-6}$ for 0.066–0.033 nM labeled tubulin). Fitting the above formula for the expected cumulative distribution to the measured cumulative histogram of distances between speckle pairs made it possible to estimate parameters of $f(l)$ (Figure 7.9m). For instance, it was estimated that the ratio between the mean length of MTs and the spindle length is ~0.4.

In summary, by integrating single-fluorophore imaging with computational image analysis, it was found that spindle MTs in close proximity move at highly heterogeneous velocities, and that the majority have a length shorter than the spindle pole-to-metaphase plate distance. These results, along with molecular perturbation data (not shown), suggest that MTs in the vertebrate meiotic spindle are dynamically organized as a crosslinked 'tiled-array' in a way similar to how the actin network is organized in motile cells (Figure 7.9n) [62]. This model challenges longstanding textbook models, which assume that the majority of MTs emanate from the two poles. The mechanical stability of the tiled-array is maintained by dynamic crosslinks. Thus, a structure can be formed where the stability of the ensemble is much higher than the stability of its individual building blocks. It is speculated that the design of cytoskeleton structures

same band, coexisted over a time interval of at least 10 s, and varied synchronously in flux velocity. Scale bar = 10 µm; (k) Kymograph representation of the synchronous movement of the speckle pairs shown in (j); (l) Histogram of the measured distances between speckle pairs (328 pairs from $n = 13$ spindles); (m) Estimated length distributions under different models: exponential distribution (mean ± SD: 11.75 ± 11.75 µm) (light blue), Rayleigh distribution (mean ± SD: 22.00 ± 11.50 µm) (blue), truncated normal distribution (TND; mean ± SD: 20.11 ± 12.23 µm). TND was selected based on its minimal fitting error and statistically validated [62]. Spindle length: mean ± SD: 49.0 ± 5.0 µm ($n = 13$ spindles); (n) A tile-array architectural model of the *Xenopus* extract spindle. Figure reproduced with permission from Ref. [62].


(a)
(b)
10 μm
5 μm
(c)
k_{on2} (to substrate) : 0.025
k_{on3} (to actin) : 0.001
k_{on2} : 0.005
k_{on3} : 0.005
k_{on2} : 0.001
k_{on3} : 0.025
(d)
VMCS
0.96
0.48
0.50
0.03
k_{on2} (to substrate)
k_{on3} (to actin)
(e)
VMCS = 1.02R_{kon} − 0.02
VMCS
$R_{kon} = k_{on3}/(k_{on2} + k_{on3})$
(f)
DCS & VMCS
actin & vinculin speed (μm/min)
Time (min)
Vinculin$_{speed}$
Actin$_{speed}$
Direction correlation
0:00:00
0:03:20
0:07:30
0:10:00
0:13:20
5 μm
(g)
Sliding starts
Sliding stops
(μm min^{-1})

follows the general principle of coupling many short and dynamic components into larger, longlasting ensembles to achieve both the flexibility and stability needed for cellular life under constantly changing conditions.

7.9.3
Hierarchical Transmission of F-Actin Motion Through Focal Adhesions

Cell migration requires a delicate spatial balance of cell adherence to the substrate. Dynamic structures called focal complexes assemble next to the leading edge and mature over time into FAs, macromolecular assemblies of more than 100 different proteins. Focal adhesions tether the F-actin network to integrin receptors, which in turn bind to the substrate. Forces generated by F-actin polymerization and/or contraction are transmitted to the extracellular matrix (ECM) via the coupling of F-actin to FAs. It has long been known that F-actin and FA proteins are coupled; many FA proteins bind directly or indirectly to F-actin [93–95] or integrin receptors [96–98], and the ends of contractile actin bundles often appear to be embedded in FAs [57, 99].

However, despite many years of intensive research aimed at identifying the molecular parts list of FAs, it has been impossible to determine the hierarchy of interactions between specific FA proteins and the F-actin cytoskeleton in living cells. Hu *et al*. used two-color total internal reflection fluorescent speckle microscopy (TIR-FSM) to simultaneously image X-rhodamine actin and various GFP-tagged FA proteins (Figure 7.10a) [25]. As expected, F-actin retrograde flow slowed down directly

Figure 7.10 Measuring the coupling between F-actin and FA proteins. (a) F-actin (red) and FA protein vinculin (green) in a live cell; (b) Simulated F-actin (red) and FA protein vinculin (green) from a Monte Carlo simulation. (c–e) Varying the association and dissociation rate constants of FA proteins in Monte Carlo simulations affects the speed of FA speckle motion and the coupling to F-actin. In these simulations, FA proteins switched between unbound (1), FA platform-bound (2) and F-actin-bound (3) states to allow coupling to F-actin flow. The VMCS and DCS were calculated using a simulation of F-actin flowing from left to right at $v = 0.25\,\mu m\,min^{-1}$; (c) Tracked motion of FA speckles (yellow vectors) for three representative FSM movies out of 25. VMCS increased from left to right as FA protein speckle flow became more aligned with F-actin speckle flow and more FA proteins were bound to the F-actin network; (d) Surface plot of VMCS determined by tracking 25 simulated FSM movies with the same dissociation rate constant ($k_{off} = k_{off21} = k_{off31} = 0.005$) and variable association rate constants to the FA platform (k_{on2}) and to F-actin (k_{on3}). For conditions $k_{on2} = k_{on3}$ (black diamonds), the VMCS was ~0.5, in agreement with the notion that half the FA proteins are stationary while the other half move at velocity v; (e) Scatter plot of VMCS versus $R_{kon} = k_{on3}/(k_{on2} + k_{on3})$. R_{kon} is a measure of the fraction of total FA proteins bound to F-actin. (f,g) Temporal variation of F-actin and vinculin speckle speeds, DCS, and VMCS within a stable (f) and a sliding (g) FA. The top left panels show graphs of average speeds of F-actin (red) and vinculin (green) speckles, vinculin-actin VMCS (blue) and vinculin-actin DCS (pink). The bottom left panels show kymographs of GFP-vinculin taken in the direction parallel to actin retrograde flow. The position of the cell edge (white) shows that the FA remains stationary in (f), whereas in (g) the FA initiates sliding at ~4 min (left arrow) and stops at ~12 min (right arrow). Right panels show maps of vinculin and actin speckle speeds and DCS. During retraction and FA sliding, vinculin alters its binding to F-actin. Times are shown as h:min:s. Figure reproduced with permission from Ref. [25].

over the FAs, suggesting that the latter may dampen flow by engaging F-actin to the ECM. Furthermore, when the motions of three classes of FA proteins were compared with F-actin, major differences in the speeds and visual coherence of the flow fields were observed. The ECM-binding α_V integrin exhibited slow, incoherent retrograde flow compared to actin, while the FA 'core' proteins paxilin, zyxin and focal-adhesion kinase (FAK), which do not bind F-actin or the ECM directly but have structural or signaling roles, moved slightly faster and more coherently. The third class, composed of the actin-binding proteins α-actinin, vinculin and talin, moved significantly faster (close to the speed of actin) and with the highest coherence.

The next step was to estimate coupling by quantifying the degree of correlated motion between FA and F-actin speckles. Correlated motion would strongly indicate that FA proteins help to transmit the force generated during actin polymerization and myosin II-mediated contraction to the ECM. To quantify this coupling, both speckle flow maps were interpolated to a common grid and two parameters were calculated: the direction coupling score (DCS) and the velocity magnitude coupling score (VMCS). The DCS measures the level of directional similarity between F-actin and FA speckle motion, while the VMCS measures the component of FA motion in the direction of actin, thus taking into account both direction and speed. The quantitative interpretation of these parameters, in terms of the kinetics of molecular interactions between F-actin and FA components, required mathematical modeling. Monte Carlo simulations were used to generate synthetic two-color movies of speckles associated with two transiently coupled protein structures (Figure 7.10b–d). Relating the binding/unbinding events at the molecular level to the relative movement of speckles in the two structures, along with an analysis of the noise characteristics of such movies, revealed that the VMCS is a linear reporter of the ratio between the time that the FA component is bound to F-actin alone and the time it is bound to both F-actin and the substrate (Figure 7.10e). Thus, the relative movement of two speckle fields directly reflects the degree of interaction between two protein structures in living cells.

When the DCS and VMCS were calculated for the three classes of FA proteins, integrin had the lowest coupling to actin, and F-actin-binding proteins the highest. The core proteins showed intermediate coupling. Such quantitative analysis provides even more insight when scores are compared over time. For example, coupling scores stayed constant for stationary FAs in a protrusive area of the cell edge, but F-actin–vinculin coupling increased in FAs that slid backwards in a retracting area of the cell edge (Figure 7.10f and g). Multicolor qFSM analysis therefore suggests a hierarchical molecular clutch model of force transmission, in which the efficiency of force transmission depends on the make-up of the FAs [25].

7.10
Outlook: Speckle Fluctuation Analysis to Probe Material Properties

Speckle trajectories probe different dynamic phenomena at different spatial and temporal scales. So far, by using the long-range directed components of speckle trajectories, qFSM has been used to measure the flow and deformation of F-actin and

MT networks. These movements are induced by molecular forces coordinated over several microns (e.g. by the activity of a large number of molecular motors or the concerted polymerization of many filaments). On a shorter spatiotemporal scale, speckle trajectories contain components associated with the microscopic deformations of polymer scaffolds that are induced by less-coordinated local actions of individual motors and thermal forces. The positional fluctuations of speckles can also be attributed to the sliding of locally decoupled filaments, to filament bending inside the network, and to photometric shifts of the speckle centroids due to local fluorophore exchange. These fluctuations occur at a length scale shorter than the mesh size of the polymer scaffold, and are independent between speckles. When calculating the cross-correlation of trajectories of two speckles separated by a distance greater than the mesh size, these fluctuations cancel out. However, even after directional components are eliminated, the cross-correlation between two-speckle trajectories decays with $1/r$, where r denotes the distance between them. The magnitude of the correlation indicates how much of the fluctuations are spatially transmitted through the material. Soft materials have a higher rate of transmission, and hence a higher correlation magnitude, than stiff materials. This is shown in Figure 7.11a for the example of a soft and a stiff F-actin network

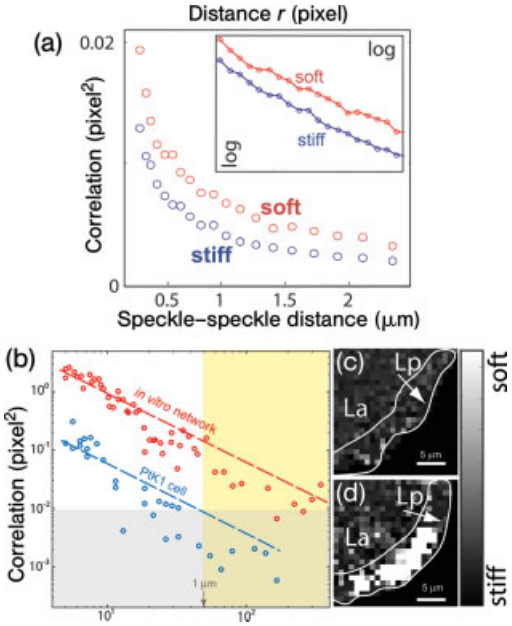


Figure 7.11 Probing stiffness of F-actin networks inside cells. (a) Correlation of random motion of two speckles as a function of their interspeckle distance r. The curves follow a $1/r$ decay (see inset), as predicted for a viscoeleastic medium; (b) Log-log plot of the correlation of random speckle motion as a function of r. *In vitro* networks are at least one order of magnitude softer than plated PtK1 cells. The gray area indicates the noise floor. Yellow area: for $r > 1\,\mu m$ the correlation is insignificant. Compliance of F-actin networks in a control cell (c) and in a cell expressing constitutively active cofilin (S3A) (d) which softens the lamellipodium network, most likely due to its selective severing activity on Lp filaments. Panel (a) reproduced with permission from Ref. [23].

measured *in vitro*. The $1/r$ decay of the fluctuation correlation is known from two-point microrheology, in which embedded beads instead of speckles are used to track thermal fluctuations in polymer networks [100, 101]. Thus, spatially correlated yet undirected components of speckle motion could be used to probe material properties of polymer networks inside a cell at the scale of the interspeckle distance. Figure 7.11b compares the stiffness between an *in vitro* network of entangled actin filaments and a cortical F-actin network in an epithelial cell. The marked difference originates in the dense crosslinking of *in vivo* networks, both intracellularly and extracellularly.

The stiffness of cellular networks is so high that correlations above noise are measurable only for speckles at a distance <1 μm (Figure 7.11b, gray zone). The possibility to extract meaningful information from speckle fluctuations over these short distances relies on recent enhancements of speckle tracking to an accuracy of approximately one-tenth of a pixel, even when speckles overlap. A module was also implemented that performs correlation analysis in small windows to map out the spatial modulation of material properties. Figure 7.11c and d compare the stiffness maps of a control epithelial cell and a cell expressing constitutively active cofilin(S3A). While the control cell has minimal spatial variation, the cell with cofilin(S3A) has a much softer lamellipodium (Lp) network but an unchanged lamella (La). These data show that cofilin – a severing factor and promoter of F-actin depolymerization – acts selectively in the lamellipodium network. High cofilin activity eventually eliminates the crosslink between the lamellipodium and lamella, resulting in a substantial softening of the lamellipodium network structure [85]. This example illustrates the potential of qFSM to derive spatial maps of the mechanical properties of cytoskeleton structures from speckle fluctuations.

7.11 Conclusions

Over the past few years, FSM has become a versatile tool for simultaneously probing the motion, deformation, turnover and materials properties of macromolecular assemblies. Despite the many exciting discoveries already made using FSM (see Table 7.1), it is a technology still in its infancy. In a next step, FSM measurements will be combined with correlational analyses to establish how assemblies operate as dynamic and plastic structures, enabling a broad variety of cell functions. In parallel, FSM will continue to go multispectral, so that these parameters can be correlated among different macromolecular structures. This requires major modifications to the current qFSM software to cope with the explosion of combinatorial data in two or more simultaneously imaged speckle channels.

With regards to future applications, FSM has the potential to uncover new biology outside the cytoskeleton field, and the analyses of FA dynamics have made some initial steps in this direction. Projects are also under way to apply qFSM to studies of the dynamic interaction of clathrin, dynamin and actin structures during endocytosis; of individual interphase MTs, MT-associated proteins and F-actin; and of DNA repair [102].

In addition to advancing basic research, FSM will hopefully become an important tool in drug discovery, particularly in the area of cancer. Already, by measuring MT dynamics in FSM-amenable cell lines transfected with patient-derived mutations in tumor suppressor genes, it has been observed that differential disease phenotypes are reproducibly replicated by different phenotypes of MT dynamics in nondividing cells (unpublished data). These subtle – but statistically highly significant – shifts in MT dynamics, that are not resolvable in static images of fixed cells, may disrupt the balance of the MT dynamics-mediated organization of signals within the cell and/or cell morphological functions. At the scale of multicellular tissues, these defects may result in detrimental responses that trigger tumor formation and metastatic behavior. Thus, FSM could enable the development of a screen for tumor-specific and patient-specific cancer drugs that would reverse the differences between cancer and control cells in terms of cytoskeleton dynamics. Such specific diagnostic tools at the subcellular scale may, at an early stage, allow the identification of efficient compounds and compound combinations with less harsh side effects than the current antimitotic chemotherapies.

Acknowledgments

These studies were supported by NIH through grants R01 GM67230 and NIH R01 GM60678 to the Danuser laboratory. Fellowship support from the Burroughs-Wellcome LJIS program (G.Y.) and the National Science Foundation (K.A.) is also acknowledged. We thank our collaborators, Clare Waterman-Storer, Edward Salmon, Tarun Kapoor, Julie Theriot, Paul Forscher and their laboratory members, for image data and uncountable discussions, without which the development of qFSM would not have been possible. We also thank James Lim and Dinah Loerke for sharing their unpublished data.

References

1 Pauling, L., Itano, H., Singer, S.J. and Wells, I. (1949) *Science*, **110**, 543.
2 Nahta, R. and Esteva, F.J. (2003) *Clinical Cancer Research*, **9**, 5078.
3 Garg, U. and Dasouki, M. (2006) *Clinical Biochemistry*, **39**, 315.
4 Eggert, U.S. and Mitchison, T.J. (2006) *Current Opinion in Chemical Biology*, **10**, 232.
5 Dorn, J.F., Danuser, G. and Yang, G. (2008) in *Fluorescent Proteins*, 2nd edn, Academic Press, Elsevier, Vol. 85, pp. 497.
6 Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. and Walter, P. (2002) *Molecular Biology of the Cell*, 4th edn, Garland Science, New York.
7 Ramaekers, F.C. and Bosman, F.T. (2004) *The Journal of Pathology*, **204**, 351.
8 http://www.cellmigration.org (2007) (accessed 4 October 2007).
9 Abercrombie, M. (1978) *Proceedings of the Royal Society of London. Series, B, Biological Sciences*, **207**, 129.
10 Lauffenburger, D.A. and Horwitz, A.F. (1996) *Cell*, **84**, 359.

11 Ridley, A.J., Schwartz, M.A., Burridge, K., Firtel, R.A., Ginsberg, M.H., Borisy, G.B., Parsons, J.T. and Horwitz, A.R. (2003) *Science*, **302**, 1704.

12 Small, J.V., Geiger, B., Kaverina, I. and Bershadsky, A. (2002) *Nature Reviews Molecular Cell Biology*, **3**, 957.

13 Rodriguez, O.C., Schaefer, A.W., Mandato, C.A., Forscher, P., Bement, W.M. and Waterman-Storer, C.M. (2003) *Nature Cell Biology*, **5**, 599.

14 Lewis, E.M. (1991) *Journal of Surgical Oncology*, **47**, 243.

15 Ganem, N.J., Storchova, Z. and Pellman, D. (2007) *Current Opinion in Genetics and Development*, **17**, 157.

16 Skop, A.R., Liu, H., Yates, J. III, Meyer, B.J. and Heald, R. (2004) *Science*, **305**, 61.

17 Yang, Y., Varvel, N.H., Lamb, B.T. and Herrup, K. (2006) *The Journal of Neuroscience*, **26**, 775.

18 Wang, N., Butler, J.P. and Ingber, D.E. (1993) *Science*, **260**, 1124.

19 Chen, C.S., Mrksich, M., Huang, S., Whitesides, G.M. and Ingber, D.E. (1997) *Science*, **276**, 1425.

20 Engler, A.J., Sen, S., Sweeney, H.L. and Discher, D.E. (2006) *Cell*, **126**, 677.

21 Ruch, R.J. (2002) *Toxicological Sciences*, **68**, 265.

22 Stevens, J.M., Galyov, E.E. and Stevens, M.P. (2006) *Nature Reviews Microbiology*, **4**, 91.

23 Danuser, G. and Waterman-Storer, C.M. (2006) *Annual Review of Biophysics and Biomolecular Structure*, **35**, 361.

24 Waterman-Storer, C.M. and Salmon, E.D. (1997) *The Journal of Cell Biology*, **139**, 417.

25 Hu, K., Ji, L., Applegate, K., Danuser, G. and Waterman-Storer, C.M. (2007) *Science*, **315**, 111.

26 Salmon, W.C., Adams, M.C. and Waterman-Storer, C.M. (2002) *The Journal of Cell Biology*, **158**, 31.

27 Schaefer, A.W., Kabir, N. and Forscher, P. (2002) *The Journal of Cell Biology*, **158**, 139.

28 Ji, L., Loerke, D., Gardel, M. and Danuser, G. (2007) *Methods in Cell Biology*, **83**, 199–235.

29 Lippincott-Schwartz, J. and Patterson, G.H. (2003) *Science*, **300**, 87.

30 Mitchison, T.J. (1989) *The Journal of Cell Biology*, **109**, 637.

31 Theriot, J.A. and Mitchison, T.J. (1991) *Nature*, **352**, 126.

32 Wadsworth, P. and Salmon, E. (1986) *The Journal of Cell Biology*, **102**, 1032.

33 Wang, Y. (1985) *The Journal of Cell Biology*, **101**, 597.

34 Wolf, D.E. (1989) *Methods in Cell Biology*, **30**, 271.

35 Dunn, G.A., Dobbie, I.M., Monypenny, J., Holt, M.R. and Zicha, D. (2002) *Journal of Microscopy, Oxford*, **205**, 109.

36 Zicha, D., Dobbie, I.M., Holt, M.R., Monypenny, J., Soong, D.Y.H., Gray, C. and Dunn, G.A. (2003) *Science*, **300**, 142.

37 Maddox, P., Desai, A., Oegema, K., Mitchison, T.J. and Salmon, E.D. (2002) *Current Biology*, **12**, 1670.

38 Maddox, P., Straight, A., Coughlin, P., Mitchison, T.J. and Salmon, E.D. (2003) *The Journal of Cell Biology*, **162**, 377.

39 Waterman-Storer, C. Desai, A. and Salmon, E.D. (1999) *Methods in Cell Biology*, **61**, 155.

40 Adams, M., Matov, A., Yarar, D., Gupton, S., Danuser, G. and Waterman-Storer, C.M. (2004) *Journal of Microscopy*, **216**, 138.

41 Adams, M.C., Salmon, W.C., Gupton, S.L., Cohan, C.S., Wittmann, T., Prigozhina, N. and Waterman-Storer, C.M. (2003) *Methods (San Diego, Calif.)*, **29**, 29.

42 Grego, S., Cantillana, V. and Salmon, E.D. (2001) *Biophysical Journal*, **81**, 66.

43 Waterman-Storer, C.M. and Salmon, E.D. (1998) *Biophysical Journal*, **75**, 2059.

44 Desai, A. and Mitchison, T.J. (1997) *Annual Review of Cell and Developmental Biology*, **13**, 83.

45 Inoue, S. and Spring, K.R. (1997) *Video Microscopy: The Fundamentals*, 2nd edn, Plenum, New York and London.

46 Danuser, G. and Waterman-Storer, C.M. (2003) *Journal of Microscopy*, **211**, 191.

47 Verkhovsky, A.B., Svitkina, T.M. and Borisy, G.G. (1999) *Current Biology*, **9**, 11.

48 Watanabe, Y. and Mitchison, T.J. (2002) *Science*, **295**, 1083.

49 Waterman-Storer, C.M., Desai, A., Bulinski, J.C. and Salmon, E.D. (1998) *Current Biology*, **8**, 1227.

50 Waterman-Storer, C.M. Salmon, W.C. and Salmon, E.D. (2000) *Molecular Biology of the Cell*, **11**, 2471.

51 Jurado, C., Haserick, J.R. and Lee, J. (2005) *Molecular Biology of the Cell*, **16**, 507.

52 Vallotton, P., Danuser, G., Bohnet, S., Meister, J.J. and Verkhovsky, A. (2005) *Molecular Biology of the Cell*, **16**, 1223.

53 Zhang, X.-F., Schaefer, A.W., Burnette, D.T., Schoonderwoert, V.T. and Forscher, P. (2003) *Neuron*, **40**, 931.

54 Pollard, T.D., Blanchoin, L. and Mullins, R.D. (2000) *Annual Review of Biophysics and Biomolecular Structure*, **29**, 545.

55 Small, V. (1981) *The Journal of Cell Biology*, **91**, 695.

56 Svitkina, T.M., Verkhovsky, A.B., McQuade, K.M. and Borisy, G.G. (1997) *The Journal of Cell Biology*, **139**, 397.

57 Geiger, B., Bershadsky, A., Pankov, R. and Yamada, K.M. (2001) *Nature Reviews Molecular Cell Biology*, **2**, 793.

58 Bulinski, J.C., Odde, D.J., Howell, B.J., Salmon, T.D. and Waterman-Storer, C.M. (2001) *Journal of Cell Science*, **114**, 3885.

59 Kapoor, T.M. and Mitchison, T.J. (2001) *The Journal of Cell Biology*, **154**, 1125.

60 Ponti, A., Vallotton, P., Salmon, W.C., Waterman-Storer, C.M. and Danuser, G. (2003) *Biophysical Journal*, **84**, 3336.

61 Ponti, A. (2004) High-resolution analysis of F-actin meshwork kinetics and kinematics using computational fluorescent speckle microscopy. Dissertation No. 15286, *ETH Zurich (Zurich)*.

62 Yang, G., Houghtaling, B.R., Gaetz, J., Liu, J.Z., Danuser, G. and Kapoor, T.M. (2007) *Nature Cell Biology*, **9**, 1233.

63 Mikhailov, A.V. and Gundersen, G.G. (1995) *Cell Motility and the Cytoskeleton*, **32**, 173.

64 Waterman-Storer, C.M. Sanger, J. and Sanger, J. (1993) *Cell Motility and the Cytoskeleton*, **26**, 19.

65 Waterman-Storer, C.M. (2002) *Current Protocols in Cell Biology* (eds J.S. Bonifacino, M. Dasso, J.B. Harford, J. Lippincott-Schwartz and K.M. Yamada), John Wiley & Sons, New York.

66 Gupton, S.L. and Waterman-Storer, C.M. (2006) *Cell Biology: A Laboratory Handbook*, 3rd edn, Vol. **3**, (eds J. Celis, N. Carter, K. Simons, J.V. Small, T. Hunter and D. Shotton), Academic Press, San Diego, pp. 137.

67 Brust-Mascher, I. and Scholey, J.M. (2002) *Molecular Biology of the Cell*, **13**, 3967.

68 Gaetz, J. and Kapoor, T.M. (2004) *The Journal of Cell Biology*, **166**, 465.

69 Gupton, S.L., Salmon, W.C. and Waterman-Storer, C.M. (2002) *Current Biology*, **12**, 1891.

70 Wittmann, T., Bokoch, G.M. and Waterman-Storer, C.M. (2003) *The Journal of Cell Biology*, **161**, 845.

71 Jepson, A.D., Fleet, D.J. and El-Maraghi, T.F. (2003) *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **25**, 1296.

72 Micheli, E.D., Torre, V. and Uras, S. (1993) *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **15**, 434.

73 Ye, M., Haralick, R.M. and Shapiro, L.G. (2003) *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **25**, 1625.

74 Vallotton, P., Ponti, A., Waterman-Storer, C.M., Salmon, E.D. and Danuser, G. (2003) *Biophysical Journal*, **85**, 1289.

75 Ahuja, R.K., Magnanti, T.M. and Orlin, J.B. (1993) *Network Flows: Theory, Algorithms and Optimization*, Prentice-Hall, Inc., New Jersey.

76 Miyamoto, D.T., Perlman, Z.E., Burbank, K.S., Groen, A.C. and Mitchison, T.J. (2004) *The Journal of Cell Biology*, **167**, 813.

77 Ji, L. and Danuser, G. (2005) *Journal of Microscopy*, **220**, 150.

78 Ponti, A., Matov, A., Adams, M., Gupton, S., Waterman-Storer, C.M. and Danuser, G. (2005) *Biophysical Journal*, **89**, 3456.

79 Blackman, S.S. and Popoli, R. (1999) *Design and Analysis of Modern Tracking Systems*, Artech House, Norwood, MA.

80 Burkard, K.E. and Cela, E. (1999) in *Handbook of Combinatorial Optimization*, Vol. Supp. A, (eds D.Z. Du and P.M. Pardalos), Kluwer Academic Publishers, Dordrecht, NL, p. 75.

81 Vallotton, P., Gupton, S.L., Waterman-Storer, C.M. and Danuser, G. (2004) *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 9660.

82 Pollard, T.D. and Borisy, G.B. (2003) *Cell*, **112**, 453.

83 Ponti, A., Machacek, M., Gupton, S.L., Waterman-Storer, C.M. and Danuser, G. (2004) *Science*, **305**, 1782.

84 Gupton, S.L., Anderson, K.L., Kole, T.P., Fischer, R.S., Ponti, A., Hitchcock-DeGregori, S.E., Danuser, G., Fowler, V.M., Wirtz, D., Hanein, D. and Waterman-Storer, C.M. (2005) *The Journal of Cell Biology*, **168**, 619.

85 Delorme, V., Machacek, M., DerMardirossian, C., Andersen, K.L., Wittmann, T., Hanein, D., Waterman-Storer, C.M., Danuser, G. and Bokoch, G. (2007) *Developmental Cell*, **13** (5), 646–662.

86 Murthy, K. and Wadsworth, P. (2005) *Current Biology*, **15**, 724.

87 Mitchison, T. and Kirschner, M. (1984) *Nature*, **312**, 237.

88 Kinoshita, K., Arnal, I., Desai, A., Drechsel, D.N. and Hyman, A.A. (2001) *Science*, **294**, 1340.

89 Sawin, K.E. and Mitchison, T.J. (1991) *The Journal of Cell Biology*, **112**, 941.

90 Burbank, K.S., Groen, A.C., Perlman, Z.E., Fisher, D.D. and Mitchison, T.J. (2006) *The Journal of Cell Biology*, **175**, 369.

91 Mastronarde, D.N., McDonald, K.L., Ding, R. and McIntosh, J.R. (1993) *The Journal of Cell Biology*, **123**, 1475.

92 Mitchison, T.J., Maddox, P., Groen, A., Cameron, L., Perlman, Z., Ohi, R., Desai, A., Salmon, E.D. and Kapoor, T.M. (2004) *Molecular Biology of the Cell*, **15**, 5603.

93 Maruyama, K. and Ebashi, S. (1965) *Journal of Biochemistry*, **58**, 13.

94 Muguruma, M., Matsumura, S. and Fukazawa, T. (1990) *Biochemical and Biophysical Research Communications*, **171**, 1217.

95 Johnson, R.P. and Craig, S.W. (1995) *Nature*, **373**, 261.

96 Tanaka, T., Yamaguchi, R., Sabe, H., Sekiguchi, K. and Healy, J.M. (1996) *FEBS Letters*, **399**, 53.

97 Calderwood, D.A., Zent, R., Grant, R., Rees, D.J.G., Hynes, R.O. and Ginsberg, M.H. (1999) *The Journal of Biological Chemistry*, **274**, 28071.

98 Burridge, K. and Mangeat, P. (1984) *Nature*, **308**, 744.

99 Burridge, K. and Chrzanowska-Wodnicka, M. (1996) *Annual Review of Cell and Developmental Biology*, **12**, 463.

100 Crocker, J.C., Valentine, M.T., Weeks, E.R., Gisler, T., Kaplan, P.D., Yodh, A.G. and Weitz, D.A. (2000) *Physical Review Letters*, **85**, 888.

101 Gardel, M.L., Shin, J.H., MacKintosh, F.C., Mahadevan, L., Matsudaira, P. and Weitz, D.A. (2004) *Science*, **304**, 1301.

102 Soutoglou, E., Dorn, J.F., Sengupta, K., Jasin, M., Nussenzweig, A., Ried, T., Danuser, G. and Misteli, T. (2007) *Nature Cell Biology*, **9**, 675.

103 Fraley, C. and Raftery, A.E. (2002) *Journal of the American Statistical Association*, **97**, 611.

8
Harnessing Biological Motors to Engineer Systems for Nanoscale Transport and Assembly*

Anita Goel and Viola Vogel

By considering how the biological machinery of our cells carries out many different functions with a high level of specificity, we can identify a number of engineering principles that can be used to harness these sophisticated molecular machines for applications outside their usual environments. Here, we focus on two broad classes of nanomotors that burn chemical energy to move along linear tracks: assembly nanomotors and transport nanomotors.

8.1
Sequential Assembly and Polymerization

The molecular machinery found in our cells is responsible for the sequential assembly of complex biopolymers from their component building blocks (monomers): polymerases make DNA and RNA from nucleic acids, and ribosomes construct proteins from amino acids. These assembly nanomotors operate in conjunction with a master DNA or RNA template that defines the order in which individual building blocks must be incorporated into a new biopolymer. In addition to recognizing and binding the correct substrates (from a pool of many different ones), the motors must also catalyze the chemical reaction that joins them into a growing polymer chain. Moreover, both types of motors have evolved highly sophisticated mechanisms so that they are able not only to discriminate the correct monomers from the wrong ones, but also to detect and repair mistakes as they occur [1].

Molecular assembly machines or nanomotors (Figure 8.1a) must effectively discriminate between substrate monomers that are structurally very similar. Polymerases must be able to distinguish between different nucleosides, and ribosomes need to recognize particular transfer RNAs (tRNAs) that carry a specific amino acid. These well-engineered biological nanomotors achieve this by pairing complementary Watson–Crick base pairs and comparing the geometrical fit of the monomers to their respective polymeric templates. This molecular discrimination makes use of the

*Reprinted by Permission from Macmillan Publishers Ltd: nature nanotechnology, Vol 2, August 2008.

Nanotechnology, Volume 5: Nanomedicine. Edited by Viola Vogel

ISBN: 978-3-527-31736-3

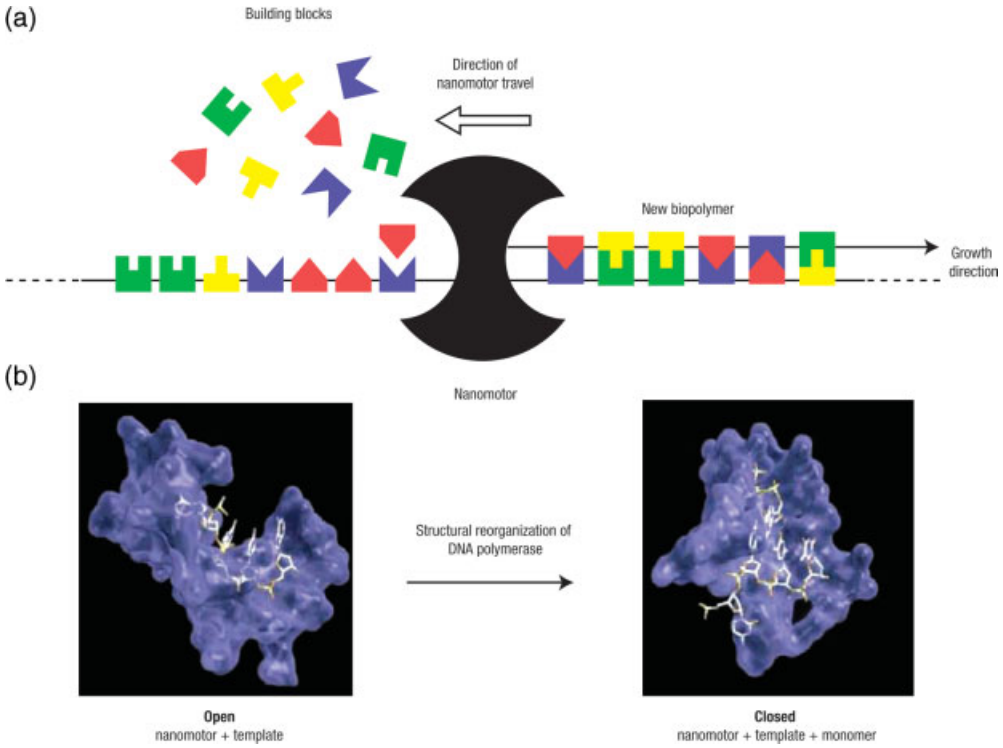


Figure 8.1 Molecular discrimination during sequential assembly. (a), The polymerase nanomotor discriminates between four different building blocks as it assembles a DNA or RNA strand complementary to its template sequence. Molecular discrimination between substrate monomers that are structurally very similar is achieved by comparing the geometrical fit of the monomers to their respective polymeric templates; (b), The T7 DNA polymerase motor undergoes an internal structural transition from an open state (when the active site samples different nucleotides) to a closed state (when the correct nucleotide is incorporated into the nascent DNA strand). Nucleotides are added to the nascent strand one at a time. This structural transition is the rate-limiting step in the replication cycle and is thought to be dependent on the mechanical tension in the template strand [2, 9, 107, 116, 121, 127, 128, 131]. Figure adapted from Ref. [127]; © 2001 PNAS.

differential binding strengths of correctly matched and mismatched substrates, which is determined by the complementarity of the base-pairing between them.

Figure 8.1b illustrates the assembly process used by the DNA polymerase nanomotor. A template of single-stranded DNA binds to the nanomotor with angstrom-level precision, forming an open complex. The open complex can 'sample' the free nucleosides available. Binding of the correct nucleoside induces a conformational change in the nanomotor, which then allows the new nucleoside to be added to the growing DNA strand [1]. The tight-fitting complementarity of shapes between the polymerase binding site and the properly paired base pair guarantees a 'geometric selection' for the correct nucleotide [2]. A similar mechanism is seen in *Escherichia coli* RNA polymerase, where the binding of an incorrect monomer inhibits the

conformational change in the motor from an 'open' (inactive) to a 'closed' (active) conformation [3].

Ribosome motors carry out tasks much more complex than polymerases. Instead of the four nucleotide building blocks used by polymerases to assemble DNA or RNA, ribosomes must recognize and selectively arrange 20 amino acids to synthesize a protein. This fact alone increases the chance of errors. Nevertheless, ribosomes obviously work (and do so along the same principles of geometric fit and conformational change as do polymerases) and are able to build amino acid polymers that are subsequently folded into functional proteins. But ribosomal motors can be tricked, much more easily than DNA motors, into building the 'incorrect' sequences when supplied with synthetic amino acids that resemble real ones [4].

8.1.1
Engineering Principle No. 1: discrimination of similar building blocks

Nanomotors used in the sequential assembly of biopolymers can discriminate efficiently between similar building blocks.

The structure of molecular machines can be visualized with angstrom-level resolution using X-ray crystallography, and the sequential assembly processes they drive can be probed in real time using single-molecule techniques [5–9]. By elucidating nanomotor kinetics under load, such nanoscale techniques provide detailed insights into the single-molecule dynamics of nanomotor-driven assembly processes. Techniques such as optical and magnetic tweezers, for example, have further elucidated the polymer properties of DNA [7, 10–12] and the force-dependent kinetics of molecular motors [13–18]. Single-molecule fluorescence methods such as fluorescence energy transfer, in conjunction with such biomechanical tools, are illuminating the internal conformational dynamics of these nanomotors [19–21].

As the underlying design principles of assembly nanomotors are revealed, it will become increasingly possible to use these biomachines for *ex vivo* tasks. Sequencing and PCR are two such techniques that already harness polymerase nanomotors for the *ex vivo* replication of nucleic acids. The polymerase chain reaction, or PCR, is a landmark, Nobel prize-winning technique [22] invented in the 1980s that harnessed polymerase nanomotors to amplify a very small starting sample of DNA to billions of molecules. Likewise, there are many conceivable future applications that either use assembly nanomotors *ex vivo* or mimic some of their design principles. Efforts are already under way to control these nanomotors better, and thus to improve such *ex vivo* sequential assembly processes for industrial use (see, for example, the websites www.cambrios.com; www.helicosbio.com; www.nanobiosym.com; www.pacificbiosciences.com).

In contrast, current *ex vivo* methods to synthesize block copolymers rely primarily on random collisions, resulting in a wide range of length distributions and much less control over the final sequence [23]. Sequential assembly without the use of nanomotors remains limited to the synthesis of comparatively short peptides, oligonucleotides and oligosaccharides [24–26]. Common synthesizers still lack both

the precision of monomer selection and the inbuilt proofreading machinery for monomer repair that nanomotors have. Building such copolymers with polymerase nanomotors *ex vivo* would yield much more homogeneous products of the correct sequence and precise length. Natural (e.g., nanomotor-enabled) designs could inspire new technologies to synthesize custom biopolymers precisely from a given blueprint.

Ribosome motors have likewise been harnessed *ex vivo* to drive the assembly of new bioinorganic heterostructures [27] and peptide nanowires [28, 29] with gold-modified amino acids inserted into a polypeptide chain. These ribosomes are forced to use inorganically modified tRNAs to sequentially assemble a hybrid protein containing gold nanoparticles wherever the amino acid cysteine was specified by the messenger RNA template. Such hybrid gold-containing proteins can then attach themselves selectively to materials used in electronics, such as gallium arsenide [28]. This application illustrates how biomotors could be harnessed to synthesize and assemble even nonbiological constructs such as nanoelectronic components (see www.cambrios.com).

Assembly nanomotors achieve such high precision in sequential assembly by making use of three key features: (i) geometric shape-fitting selection of their building blocks (e.g., nucleotides); (ii) motion along a polymeric template coupled to consumption of an energy source (e.g., hydrolysis of ATP molecules); and (iii) intricate proofreading machinery to correct errors as they occur. Furthermore, nanomotor-driven assembly processes allow much more stable, precise and complex nanostructures to be engineered than can be achieved by thermally driven self-assembly techniques alone [30–32].

We should also ask whether some of these principles, which work so well at the nanoscale, could be realized at the micrometer scale as well. Whitesides and coworkers, for example, have used simple molecular self-assembly strategies, driven by the interplay of hydrophobic and hydrophilic interactions, to assemble microfabricated objects at the mesoscale [33, 34]. Perhaps the design principles used by nanomotors to improve precision and correct errors could also be harnessed to engineer future *ex vivo* systems at the nanoscale, as well as on other length scales. Learning how to engineer systems that mimic the precision and control of nanomotor-driven assembly processes may ultimately lead to efficient fabrication of complex nanoscopic and mesoscopic structures.

8.2 Cargo Transport

Cells routinely use another set of nanomotors (i.e., transport nanomotors) to recognize, sort, shuttle and deliver intracellular cargo along filamentous freeways to well-defined destinations, allowing molecules and organelles to become highly organized (for reviews, see Refs. [35–44]). This is essential for many life processes. Motor proteins transport cargo along cytoskeletal filaments to precise targets, concentrating molecules in desired locations. In intracellular transport, myosin motors are guided by actin

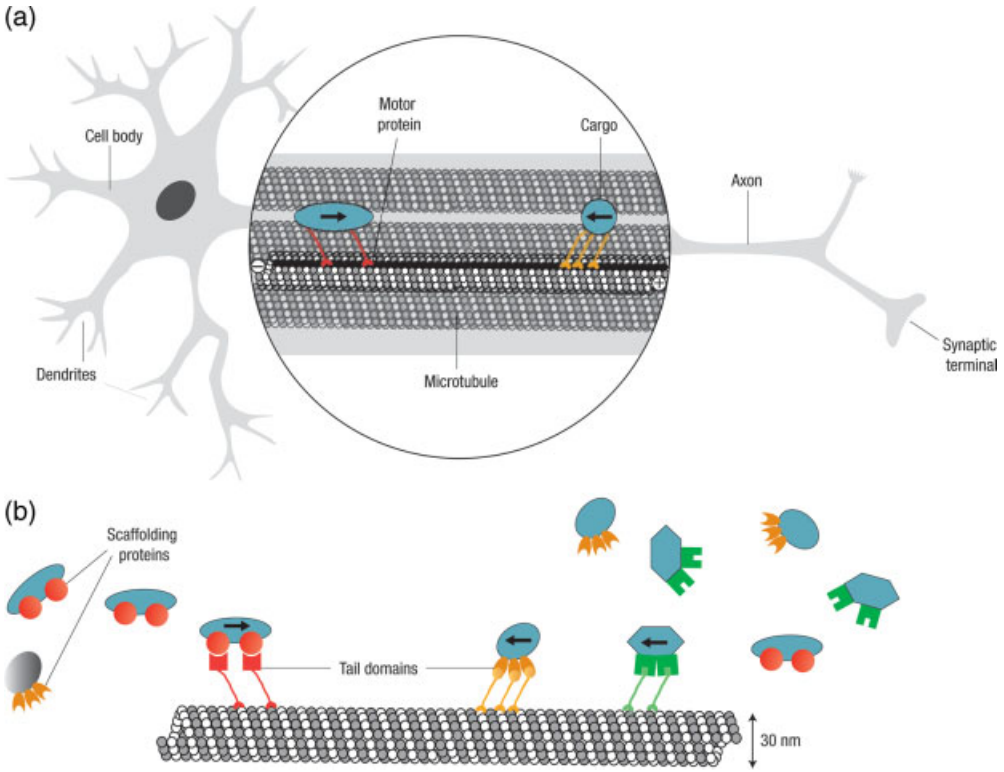


Figure 8.2 Motor-specific cargo transport in neurons. (a), The axon of neurons consists of a bundle of highly aligned microtubules along which cargo is trafficked from the cell body to the synapse and *vice versa*. Most members of the large kinesin family (red) transport cargo towards the periphery, while other motors, including dyneins (yellow), transport cargo in the opposite direction. Motors preferentially move along a protofilament rather then side-stepping (one randomly selected protofilament is shown in dark gray). Protofilaments are assembled from the dimeric protein tubulin (white and gray spheres) which gives microtubules their structural polarity. The protofilaments then form the hollow microtubule rod. When encountering each other on the same protofilament, the much more tightly bound kinesin has the 'right of way', perhaps even forcing the dynein to step sidewise to a neighboring protofilament [52–55]. (b), Each member of a motor family selects its own cargo (blue shapes) through specific binding by scaffolding proteins (colored symbols) or directly by the cargo's tail domains.

filaments, whereas dynein and kinesin motors move along rodlike microtubules. Figure 8.2a illustrates how conventional kinesins transport molecular cargo along nerve axons towards the periphery, efficiently transporting material from the cell body to the synaptic region [45]. Dyneins, in contrast, move cargo in the opposite direction, so that there is active communication and recycling between both ends (see reviews [42, 46]). In fact, the blockage of such bidirectional cargo transport along nerve axons can give rise to substantial neural disorders [47–50].

The long-range guidance of cargo is made possible by motors pulling their cargo along filamentous rods. Microtubules, for example, are polymerized from the dimeric

tubulin into protofilaments that assemble into rigid rods around 30 nm in diameter [36]. These polymeric rods are inherently unstable: they polymerize at one end (plus) while depolymerizing from the other (minus) end, giving rise to a structural polarity. The biological advantage of using transient tracks is that they can be rapidly reconfigured on demand and in response to changing cellular needs, or to various external stimuli. Highly efficient unidirectional cargo transport is realized in cells by bundling microtubules into transport highways where all microtubules are oriented in the same direction. Excessively tight bundling of microtubules, however, can greatly impair the efficiency of cargo transport, by blocking the access of motors and cargo to the microtubules in the bundle interior. Instead, microtubule-associated proteins are thought to act as repulsive polymer-brushes, thereby regulating the proximity and interactions between neighboring microtubules [51].

Traffic control is an issue when using the filaments as tracks on which kinesin and dynein motors move in opposite directions. Although different cargoes can be selectively recognized by different members of the motor protein families and shuttled to different destinations, what happens if motors moving in opposite directions encounter each other on the same protofilament (Figure 8.2b)? If two of these motors happen to run into each other, kinesin seems to have the 'right of way'. As kinesin binds the microtubule much more strongly, it is thought to force dynein to step sideways to a neighboring protofilament [52]. Dynein shows greater lateral movement between protofilaments than kinesin [52–54] as there is a strong diffusional component to its steps [55]. When a microtubule becomes overcrowded with only kinesins, the runs of individual kinesin motors are minimally affected. But when a microtubule becomes overloaded with a mutant kinesin that is unable to step efficiently, the average speed of wild-type kinesin is reduced, whereas its processivity is hardly changed. This suggests that kinesin remains tightly bound to the microtubule when encountering an obstacle and waits until the obstacle unbinds and frees the binding site for kinesin's next step [56].

8.2.1
Engineering Principle No. 2: various track designs

> *Various track designs enable motors to pull their cargo along filamentous tracks, whereas others allow motors bound to micro- or nanofabricated tracks to propel the filaments which can then serve as carriers.*

It is not a trivial task to engineer transport highways *ex vivo*, particularly in versatile geometries with intersections and complex shapes. Individual filaments typically allow only one-dimensional transport, as the motor-linked cargo drops off once the end of the filament is reached. Furthermore, conventional kinesin makes only a few hundred 8 nm-sized steps before dissociating from the microtubule [57, 58], further limiting the use of such a system for *ex vivo* applications.

Instead of having the motors transport their cargo along filaments, motors have been immobilized on surfaces in an inverted geometry that enables the filaments to

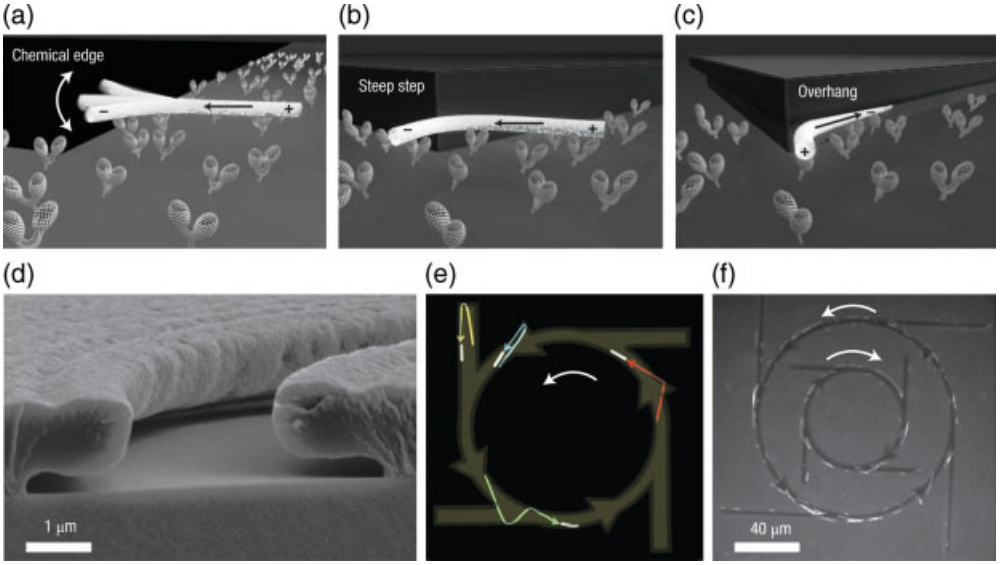


Figure 8.3 Track designs to guide nanomotor-driven filaments *ex vivo*. A variety of track designs have been used. (a), A chemical edge (adhesive stripes coated with kinesin surrounded by nonadhesive areas). The filament crosses the chemical edge and ultimately falls off as it does not find kinesins on the nonadhesive areas [61]; (b), Steep channel walls keep the microtubule on the desired path as they are forced to bend [61,65]; (c), Overhanging walls have been shown to have the highest guidance efficiency [64]; (d), Electron micrograph of a microfabricated open channel with overhanging walls [64]; (e), Breaking the symmetry of micropatterns can promote directional sorting of filament movement [63, 65, 69, 138]. The trajectories of four microtubules are shown: movement into reflector arms causes the tubule to turn around (yellow), an arrow-shaped direction rectifier allows those travelling in the desired direction to continue (red) and forces others to turn around (blue). At intersections, tubules preferentially continue straight on (green); (f), The complex microfabricated circuit analysed in (e) with open channels and overhanging walls, demonstrating unidirectional movement of microtubules.

be collectively propelled forward [45]. The head domains of the kinesin and myosin motors can rotate and swivel with respect to their feet domains, which are typically bound in random orientations to the surface. These motor heads detect the structural anisotropy of the microtubules and coherently work together to propel a filament forward [59, 60].

Various examples of such inverted designs for motor tracks have been engineered to guide filaments efficiently. Some of these are illustrated in Figure 8.3. Inverted motility assays can be created, for example, by laying down tracks of motor proteins in microscopic stripes of chemical adhesive on an otherwise flat, protein-repellent surface, surrounded by nonadhesive surface areas. Such chemical patterns (Figure 8.3a) have been explored to guide actin filaments or microtubules. The loss rate of guiding filaments increases exponentially with the angle at which they approach an adhesive/nonadhesive contact line [61]. The passage of the contact line by filaments at nongrazing angles, followed by their drop off, can be prevented by using much narrower lanes whose size is of the order of the diameter of the moving

object. Such nanoscale kinesin tracks provide good guidance and have been fabricated by nanotemplating [62].

Alternatively, considerably improved guidance has been accomplished by topographic surface features (Figure 8.3b). Microtubules hitting a wall are forced to bend along this obstacle and will continue to move along the wall [63–66]. The rigidity of the polymeric filaments used as shuttles thus greatly affects how tracks should be designed for optimal guidance. Whereas microtubules with a persistence length of a few millimeters can be effectively guided in channels a few micrometers wide as they are too stiff to turn around [61], the much more flexible actin filaments require channel widths in the submicrometer range [67, 68]. Finally, the best long-distance guidance of microtubules has been obtained so far with overhanging walls [64, 69] (Figure 8.3c). The concept of topographic guidance in fact works so well that swarms of kinesin-driven microtubules have been used as independently moving probes to image unknown surface topographies. After averaging all their trajectories in the focal plane for an extended time period, the image grayscale is determined by the probability of a surface pixel being visited by a microtubule in a given time frame [70].

But how can tracks be engineered to produce *uni*directional cargo transport? All the motor-propelled filaments must move in the same direction to achieve effective long-distance transport. When polar filaments land from solution onto a motor-covered surface, however, their orientations and initial directions of movement are often randomly distributed. Initially, various physical means, such as flow fields [71], have been introduced to promote their alignment. Strong flows eventually either force gliding microtubules to move along with the flow, or force microtubules, if either their plus or minus end is immobilized on a surface [72], to rotate around the anchoring point and along with the flow. The most universal way to control the local direction in which the filamentous shuttles are guided is to make use of asymmetric channel features. Figure 8.3d–f illustrates how filaments can be actively sorted according to their direction of motion by breaking the symmetry of the engineered tracks. This 'local directional sorting' has been demonstrated on surfaces patterned with open-channel geometries, where asymmetric intersections are followed by dead-ended channels (that is, reflector arms), or where channels are broadened into arrow heads. Both of these topographical features not only selectively pass filaments moving in the desired direction, but can also force filaments moving in the opposite direction to turn around [65, 69, 73, 74]. Once directional sorting has been accomplished, electric fields have been used to steer the movement of individual microtubules as they pass through engineered intersections [75, 76].

In addition to using isolated nanomotors, hybrid biodevices and systems that harness self-propelling microbes could be used to drive transport processes along engineered tracks. Flagellated bacteria, for example, have been used to generate both translational and rotational motion of microscopic objects [77]. These bacteria can be attached head-on to solid surfaces, either via polystyrene beads or polydimethylsiloxane, thereby enabling the cell bodies to form a densely packed monolayer, while their flagella continue to rotate freely. In fact, a microrotary motor, fuelled by glucose and comprising a 20 μm-diameter silicon dioxide rotor, can be driven along a silicon track by the gliding bacterium *Mycoplasma* [78]. Depending on the specific applica-

tion and the length scale on which transport needs to be achieved, integrating bacteria into such biohybrid devices (that work under physiological conditions) might ultimately prove more robust than relying solely on individual nanomotors.

8.3
Cargo Selection

To maintain intracellular contents in an inhomogeneous distribution far from equilibrium, the intracellular transport system must deliver molecular cargo and organelles on demand to precise destinations. This tight spatiotemporal control of molecular deliveries is critical for adequate cell function and survival. Molecular cargo or organelles are typically barcoded so that they can be recognized by their specific motor protein (Figure 8.4). Within cells, motors recognize cargo either from the cargo's tail domains directly, or via scaffolding proteins that link cargo to their tail domain [43].

8.3.1
Engineering Principle No. 3: barcoding

Engineered molecular recognition sites enable cargo to be selectively bonded to moving shuttles.

Although most cargo shuttled around by motors can be barcoded using the existing repertoire of biological scaffolding proteins, synthetic approaches are needed for all those *ex vivo* applications where the cargo has to be specifically linked to moving filaments. The loading and transport of biomedically relevant or engineered cargo has already been demonstrated (Figure 8.4) [79–83]. Typical approaches are to tag the cargo

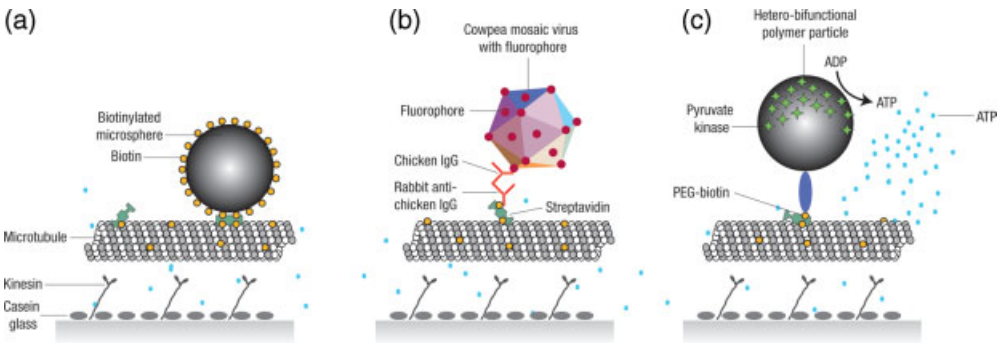


Figure 8.4 Selecting specific cargo by molecular recognition. A versatile toolbox exists by which synthetic and biological cargo can be coupled to microtubules. (a) Biotinylated objects are coupled via avidin or streptavidin to biotinylated microtubules. (b) Biological molecules, viruses [79, 81] or cells can be coupled by antibody recognition. (c) Backpacks of chemically or biologically active reagents can be shuttled around, including bioprobes [80] or tiny ATP factories [93] as shown here.

with antibodies or to biotinylate microtubules and coat the cargo with avidin or streptavidin (Figure 8.4) (for reviews, see Refs. [74, 79]), as done for polymeric and magnetic beads [84, 85] (Figure 8.4a), gold nanoparticles [86–88], DNA [87, 89, 90] and viruses [79, 81] (Figure 8.4b), and finally mobile bioprobes and sensors [80, 81, 91] (Figure 8.4c). However, if too much cargo is loaded onto the moving filaments and access of the propelling motors is even partially blocked, the transport velocity can be significantly impaired [92]. Finally, the binding of cargo to a moving shuttle can be used to regulate its performance. In fact, microtubules have recently been furnished with a backpack that selfsupplies the energy source ATP. Cargo particles bearing pyruvate kinase have been tethered to the microtubules to provide a local ATP source [93] (Figure 8.4c). The coupling of multiple motors to cargo or other scaffold materials can affect the motor performance. If single-headed instead of double-headed kinesins are used, cooperative interactions between the monomeric motors attached to protein scaffolds increase hydrolysis activity and microtubule gliding velocity [59].

At the next level of complexity – successful cargo tagging – sorting and delivery will depend on the engineering of integrated networks of cargo loading, cargo transport and cargo delivery zones. Although the construction of integrated transport circuits is still in its infancy, microfabricated loading stations have been built [88] (Figure 8.5). The challenge here is to immobilize cargo on loading stations such that it is not easily detached by thermal motion, yet to allow for rapid cargo transfer to passing microtubules. By properly tuning bond strength and multivalency, and most importantly by taking advantage of the fact that mechanical strain weakens bonds, cargo can be efficiently stored on micropatches and transferred after colliding with a microtu-

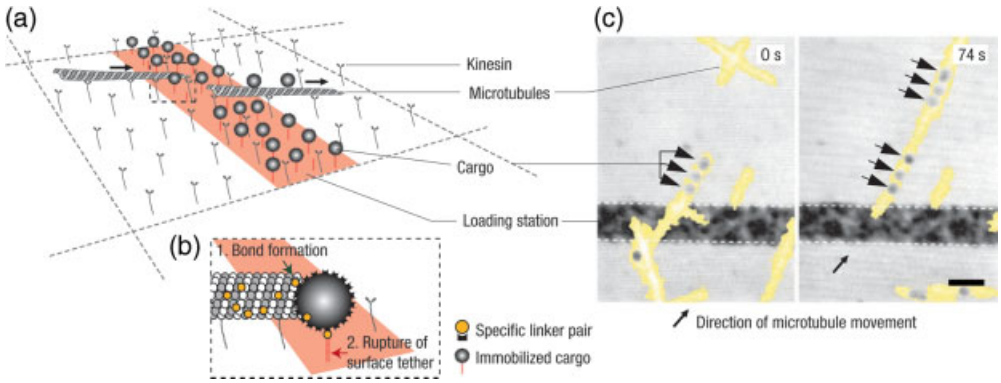


Figure 8.5 Cargo loading stations [93]; (a) Stripes of immobilized cargo are fabricated by binding thiolated oligonucleotides to micropatterned lines of gold. Hybridization with complementary strands exposing antibodies at their terminal ends allows them to immobilize a versatile range of cargos that carry antibodies on their surfaces. (b) The challenge is to tune the bond strength and valency to prevent thermal activation during cargo storage on the loading station. On collision with the shuttle (microtubule), the cargo must rapidly break off the bond it has formed with the station88. Fortunately, however, tensile mechanical force acting on a noncovalent bond shortens its lifetime; (c, d) These concepts are used in the design of the loading stations shown here, where a microtubule moves through a stripe of immobilized gold cargo and picks up a few beads.

bule [88]. Considerable fine-tuning of bond strength can be accomplished by using DNA oligomers hybridized such that the bonds are either broken by force all at once (a strong bond) or in sequence (a weak bond) [94].

As discussed above, filaments are most commonly used to shuttle molecular cargo in most emerging devices that harness linear motors for active transport. Alternatively, if the filamentous tracks could be engineered in versatile geometries, the motors themselves could be used to drag cargo coupled to the molecular recognition sites of their tail domains as in the native systems. We could thus make use of the full biological toolbox of already known or engineered scaffolding proteins that link specific motors to their respective cargoes [40, 43]. So far, assemblies of microtubules organized into complex, three-dimensional patterns such as asters, vortices and networks of interconnected poles [95, 96] have been successfully created in solution, and mesoscopic needles and rotating spools of microtubule bundles held together by noncovalent interactions have been engineered on surfaces [31]. All of these mesoscopic structures are uniquely related to active motor-driven motion, and would not have formed purely by self-assembly without access to an energy source.

To increase the complexity of microtubule track networks, densely packed arrays of microtubules have been grown in confined spaces, consisting of open microfabricated channels with user-defined geometrical patterns [97]. The key to achieving directed transport, however, is for all microtubules within each bundle or array to be oriented in the same direction. This has been accomplished by making use of directed motility in combination with sequential assembly procedures (Figure 8.6). First, microtubule seedlings have been oriented in open microfabricated and kinesin-

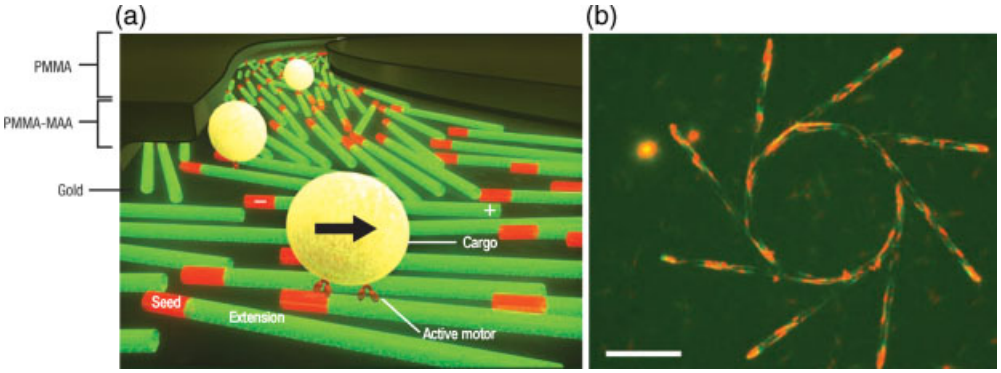


Figure 8.6 Filament tracks made from engineered bundles of microtubules [97]. Active transport is used to produce bundles of microtubules and confine them to user-defined geometries. (a) Sequential assembly procedure: first, microtubule seedlings (labelled in red) are allowed to orient themselves in open kinesin-coated microfabricated channels that contained reflector arms. Second, and after mild fixation, the oriented seedlings are polymerized into mature microtubules through the addition of tubulin into the solution (labelled green) which preferentially binds to the plus-end (polymerizing end) of the microtubules. (b) Fluorescence image of microtubules that have been grown in the confined space provided by the open channels until the channels were filled with dense networks of microtubules all oriented in the same direction [97]. Scale bar-40 μm.

coated channels that contain reflector arms. Once oriented by self-propelled motion, the seedlings were polymerized into mature microtubules that were confined to grow in the open channels until the channels were filled with dense networks of microtubules all oriented in the same direction [97]. Single kinesins take only a few hundred steps before they fall off, but the walking distance can be greatly increased if the cargo is pulled by more than one motor [98]. Such approaches to fabricating networks of microtubule bundles could be further expanded to engineer future devices that use either the full toolbox of native scaffolding proteins or new scaffolding proteins that target both biological and synthetic cargo.

Nanoengineers would not be the first to harness biological motors to transport their cargo. Various pathogens are known to hijack microtubule or actin-based transport systems within host cells (reviewed in Ref. [99]). *Listeria monocytogenes*, for example, propels itself through the host cell cytoplasm by means of a fast-polymerizing actin filament tail [100]. Likewise, the vaccinia virus, a close relative of smallpox, uses actin polymerization to enhance its cell-to-cell spreading [101], and the alpha herpesvirus hijacks kinesins to achieve long-distance transport along the microtubules of neuronal axons [102]. Signaling molecules and pathogens that cannot alter cell function and behavior by simply passing the outer cell membrane can thus hijack the cytoskeletal highways to get transported from the cell periphery to the nucleus.

8.3.2
Engineering Principle No. 4: active transport of tailored drugs and gene carriers

By taking advantage of the existing cytoskeleton, tailored drugs and gene carriers can be actively transported to the cell nucleus.

Indeed, many viruses [37, 103, 104] as well as nonviral therapeutic gene carriers, such as polyethylenimine/DNA or other polymer-based gene transfer systems (i.e., polyplexes) [105, 106] take advantage of nanomotor-driven transport along microtubule filaments to accelerate their way through the cytoplasm towards the nucleus. Nanomotor-driven transport to the nucleus leads to a much more efficient nuclear localization than could ever be achieved by slow random diffusion through the viscous cytoplasm. Active gene carrier transport can lead to more efficient perinuclear accumulation within minutes [37, 105, 106]. In contrast, nonviral gene carriers that depend solely on random diffusion through the cytoplasm move much more slowly and thus have considerably reduced transfection efficiencies. Understanding how to 'hijack' molecular and cellular transport systems, instead of letting a molecule become a target for endosomal degradation [37, 91], will ultimately allow the design of more efficient drug and gene carrier systems.

8.4
Quality Control

Nanomanufacturing processes, much like macroscopic assembly lines, urgently need procedures that offer precise control over the quality of the product, including

the ability to recognize and repair defects. Living systems use numerous quality control procedures to detect and repair defects occurring during the synthesis and assembly of biological nanostructures. As yet, this has not been possible in synthetic nanosystems. Many cellular mechanisms for damage surveillance and error correction rely on nanomotors. Such damage control can occur at two different levels as follows.

8.4.1
Engineering Principle No. 5: error recognition and repair at the molecular level

Certain motor proteins recognize assembly mistakes and repair them at the molecular level.

DNA replication represents one of the most complex sequential assembly processes in a cell. Here, the genetic information stored in the four-base code must be copied with ultra-high precision. Errors generated during replication can have disastrous biological consequences. Figure 8.7 illustrates the built-in mechanism used by the polymerase (DNAp) motor to repair mistakes made during the process of DNA replication [107]. When the DNAp motor misincorporates a base while replicating the template DNA strand, it slows down and switches gears from the polymerase to the exonuclease cycle. Once in exonuclease mode, it will excise the mismatched base pair and then rapidly switch back to the polymerase cycle to resume forward replication. Similar error correction mechanisms, known as 'kinetic proofreading', are conjectured to occur in RNA polymerases and ribosomal machineries [1, 13, 108–113].

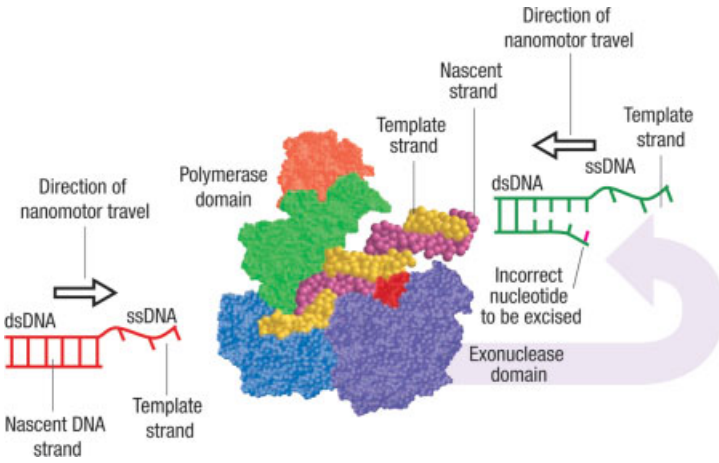


Figure 8.7 Quality control procedures for damage recognition and molecular repair. The DNA polymerase motor (DNAp) contains two active sites. It switches from polymerase (copying) to exonuclease (error correction) activity when it encounters a mismatched base. Mismatched bases are detected as they have weaker bonding interactions—the 'melting' temperature is lower—and this increases the chance of switching from the polymerase to the exonuclease active site [107]. In the exonuclease mode, the motor excises the incorrect base from the nascent DNA strand.

8.4.2
Engineering Principle No. 6: error recognition and repair at the system level

Integrated systems of motors and signaling molecules are needed to recognize and repair damage at the supramolecular level.

Nerve cells have evolved a highly regulated axonal transport system that contains an integrated damage surveillance system [114]. The traffic regulation of motors moving in opposite directions on a microtubule typically occurs in special 'turnaround' zones at the base and tip of an axon [43], but a zone for switching the organelle's direction can also be created when axonal transport is blocked at the site of nerve injury [46] (see Figure 8.2). When irreparable, such blockages are often signatures of neurodegenerative diseases. For example, amyloid precursor protein [47] or tau [115] can give rise to the accumulation of protein aggregates that inhibit anterograde axonal transport, a mechanism potentially implicated in Alzheimer's disease.

At present, there are no synthetic materials that can, in a self-regulated manner, recognize and repair defects at either the molecular or supramolecular level. Molecular recognition and repair is typically attributed to a tightly fitted stereochemical complementarity between binding partners. Nanoscale tools applied to the study of molecular recognition and repair are also elucidating the functional roles of the different structural conformations (and hence three-dimensional shapes) of the motors. For instance, the DNAp motor is in one particular conformation when it binds DNA in its copying (i.e., polymerization) mode and in an entirely different conformation (i.e., the exonuclease mode) when it binds DNA to proofread or excise a mistaken base from the replicated DNA strand [107]. In contrast, damage control at the supramolecular level (e.g., during axonal transport) is achieved by the trafficking of signaling molecules. Deciphering the underlying engineering design principles of damage surveillance and error correction mechanisms in biological systems will inevitably allow better quality-control procedures to be integrated into nanoengineered systems.

8.5
External Control

8.5.1
Engineering Principle No. 7: performance regulation on demand

As with macroscopic engines, external controls can regulate the performance of nanomotors on demand.

Learning how to control and manipulate the performance of nanomotors externally is another critical hurdle in harnessing nanomotors for *ex vivo* applications. By finding or engineering appropriate external knobs in the motor or its environment, its nanoscale movement can be tightly regulated, switched on and off, or otherwise manipulated on demand.

To achieve external control over the nanoscale movement of biological motors, it is important to identify the correct external parameters that can be used to control their dynamics. These external modulators of motor function ('handles') can be either naturally occurring or somehow artificially engineered into the motor to make it susceptible to a particular external control knob or regulator. Because the motion of nanomotors is typically driven by a series of conformational changes in the protein, mechanical load or strain on the motor molecule can also affect the dynamics of the motor. Nanomotors apply mechanical strain to their filaments or substrates as they go through various internal conformational changes. This mechanical strain is intimately related to their dynamics along the substrate and hence their functional performance. Certain interstate transition rates can depend, for example [107], on the amount of intramolecular strain in the motor protein. Applying a mechanical load to a motor perturbs key mechanical transitions in the motor's kinetic pathway, and can thereby affect rates of nucleotide binding, ATP hydrolysis and product release. Single-molecule techniques are beginning to elucidate how mechanical strain on a motor protein might be used to regulate its biological functions (e.g., nanoscale assembly or transport) [13, 55, 107, 116–120].

The single-molecule dynamics of the DNAp motor, as it converts single-stranded (ss) DNA to double-stranded (ds) DNA, has been probed, for example, through the differential elasticity of ssDNA and dsDNA (see Figure 8.8). The T7 DNA polymerase motor replicates DNA at rates of more than 100 bases per second, and this rate steadily decreases with mechanical tension greater than about 5 pN on the DNA template [9]. The motor can work against a maximum of about 34 pN of template tension [9]. The replication rates for the Klenow and Sequenase DNA polymerases also decrease when the ssDNA template tension exceeds 4 pN, and completely ceases at tensions greater than 20 pN [121]. Likewise, single-molecule techniques have allowed direct observation of the RNA polymerase (RNAp) motor moving one base at a time [122], and occasionally pausing and even backtracking [123]. Although RNAp motors are typically five- to tenfold slower than DNAp motors, the effects of DNA template tension on their dynamics are still being investigated [6]. Similarly, ribosome motors, which translate messenger RNA (mRNA) into amino acids at roughly 10 codons per second, have been found to generate about 26.5 ± 1 pN of force [124]. The underlying design principles by which these nanomotors operate are being further elucidated by theoretical models [107, 116, 125–128] that describe nanomachines at a level commensurate with single-molecule data. Furthermore, these molecular assembly machines can be actively directed, driven and controlled by environmental signals [107].

Consequently, an external load or force applied to the substrate or to the motor itself can be used to slow down a motor's action or stall its movement. The stalling forces of kinesin and dynein are 6 and 1 pN, respectively [58, 129]. For example, the binding of two kinesin domains to a microtubule track creates an internal strain in the motor that prevents ATP from binding to the leading motor head. In this way, the two motor domains remain out-of-phase for many mechanochemical cycles and thereby provide an efficient, adaptable mechanism for achieving highly processive movement [130]. Beyond stalling the movement of motors by a mechanical load, other types of perturbations can also influence the dynamics of molecular motors,

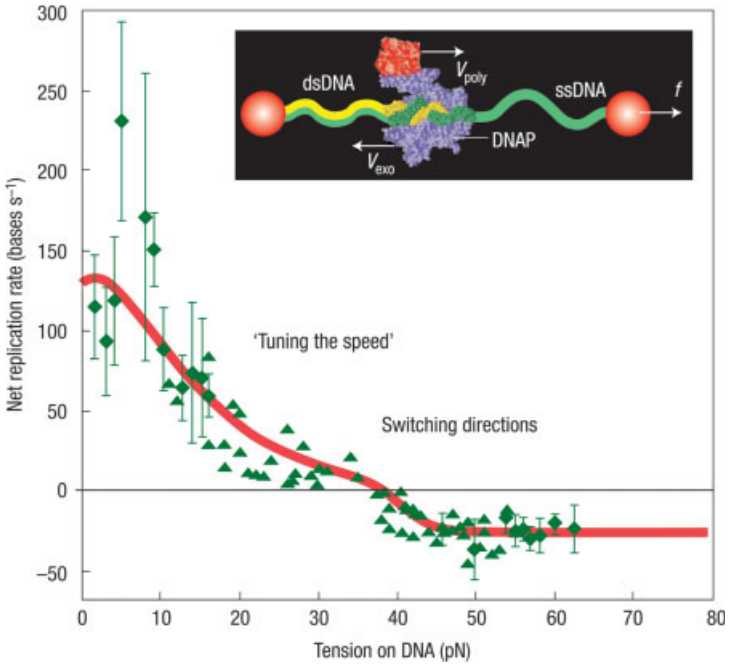


Figure 8.8 Precision control of nanomotors with external control 'knobs'. The net replication rate of a DN Ap motor can be controlled by the mechanical tension on the DNA template strand. Single-molecule data for the motor's force-dependent velocity (two sets of data—diamonds and triangles—are shown, relating to constant force and constant extension measurements) can be described by a network model (red curve) as shown here. The change in net replication rate shows how external controls can change the dynamics of the nanomotor. This model illustrates how environmental control knobs can tune the dynamics of the nanomotor by altering the rate constants associated with its various internal transitions [106]. Tensions between 0 and 35 pN control the net replication rate, whereas tensions above 35 pN actually reverse the velocity of the nanomotor. Inset, experimental setup: a single DNA molecule is stretched between two plastic beads as the motor catalyses the conversion of singlestranded to double-stranded DNA. Figure adapted from Ref. [106].

including the stretching of substrate molecules like DNA [13]. Although this external control over nanomotors has been demonstrated in a few different contexts *ex vivo*, a rich detailed mechanistic understanding of how such external control knobs can modulate the dynamics of the molecular motor is emerging from recent work on the DNA polymerase motor [9, 107, 116, 121, 127, 128, 131].

Remote-controlling the local ATP concentration by the photoactivated release of caged ATP can allow a nanomotor-driven transport system to be accelerated or stopped on demand [84]. External control knobs or regulators can also be engineered into the motors. For instance, point mutations can be introduced into the gene encoding the motor protein, such that it is engineered to respond to light, temperature, pH or other stimuli [43, 85]. Engineering light-sensitive switches into nanomotors enables the rate of ATPase [43, 132] to be regulated, thereby providing an alternate handle for tuning the motor's speed, even while the ATP concentration is

kept constant and high. When additional ATP-consuming enzymes are present in solution, the rate of ATP depletion regulates the distance the shuttles move after being activated by a light pulse and before again coming to a halt [84].

Future applications could require that, instead of all the shuttles being moved at the same time, only those in precisely defined locations be activated, on demand. Some of the highly conserved residues within motors help to determine the motor's ATPase rate [43]. Introducing chemical switches near those locations might provide a handle for chemical manipulation of the motor's speed. In fact, this has already been realized for a rotary motor [132] as well as for a linear kinesin motor, where the insertion of a Ca^{2+}-dependent chemical switch makes the ATPase activity steeply dependent on Ca^{2+} concentrations [133]. In addition to caged ATP, caged peptides that block binding sites could be used to regulate the motility of such systems. Caged peptides derived from the kinesin C-terminus domain have already been used to achieve photo control of kinesin-microtubule motility [134]. Instead of modulating the rate of ATP hydrolysis, the access of microtubules to the motor's head domain can also be blocked in an environmentally controlled manner. In fact, temperature has already been shown to regulate the number of kinesins that are accessible while embedded in a surface-bound film of thermoresponsive polymers [135].

The nanomotor-driven assembly of DNA by the DNA polymerase motor provides an excellent example of how precision control over the nanomotor can be achieved by various external knobs in the motor's environment [107, 116, 127, 128]. The DNAp motor moves along the DNA template by cycling through a given sequence of geometric shape changes. The sequence of shapes or internal states of the nanomachine can be denoted by nodes on a simple network [107, 116, 127, 128]. As illustrated in Figure 8.8, this approach elucidates how mechanical tension on a DNA molecule can precisely control (or 'tune') the nanoscale dynamics of the polymerase motor along the DNA track by coupling into key conformational changes of the motor [107].

Macroscopic knobs to precision-control the motor's movement along DNA tracks can be identified by probing how the motor's dynamics vary with each external control knob (varied one at a time). Efforts are currently under way to control even more precisely the movement of these nanomotors along DNA tracks by tightly controlling the parameters in the motor's environment (see www.nanobiosym.com). Concepts of fine-tuning and robustness could also be extended to describe the sensitivity of other nanomotors (modelled as simple biochemical networks) to various external control parameters [107]. Furthermore, such a network approach [107] provides experimentally testable predictions that could aid the design of future molecular-scale manufacturing methods that integrate nanomotor-driven assembly schemes. External control of these nanomotors will be critical in harnessing them for nanoscale manufacturing applications.

8.6 Concluding Remarks

We have reviewed several key engineering design principles that enable nanomotors moving along linear templates to perform a myriad of tasks. Equally complex

biomimetic tasks have not yet been mastered *ex vivo*, either by harnessing biological motors or via synthetic analogues. Engineering insights into how such tasks are carried out by the biological nanosystems will inspire new technologies that harness nanomotor-driven processes to build new systems for nanoscale transport and assembly.

Sequential assembly and nanoscale transport, combined with features currently attributed only to biological materials, such as self-repair and healing, might one day become an integral part of future materials and biohybrid devices. In the near term, molecular biology techniques could be used to synthesize and assemble nanoelectronic components with more control (www.cambrios.com; see also Ref. [29]). Numerous proof-of-concept experiments using nanomotors integrated into synthetic microdevices have already been demonstrated (for reviews, see Refs. [74, 136]). Among many others, these applications include stretching surface-bound molecules by moving microtubules [87, 90]; probing the lifetime of a single receptor–ligand interaction via a cantilevered microtubule that acts as a piconewton force sensor [85]; topographic surface imaging by self-propelled probes [70]; and cargo pick-up from loading stations [88] as illustrated in Figure 8.5.

Although much progress is being made in the synthesis of artificial motors (see Ref. [137]), it has been difficult, in practice, to synthesize artificial motors that come even close in performance to their natural counterparts (see Ref. [39]). Harnessing biological motors to perform nanoscale manufacturing tasks might thus be the best near-term strategy. Although many individual nanoparts can be easily manufactured, the high-throughput assembly of these nanocomponents into complex structures is still nontrivial. At present, no *ex vivo* technology exists that can actively guide such nanoscale assembly processes. Despite advances in deciphering the underlying engineering design principles of nanomotors, many hurdles still impede harnessing them for *ex vivo* transport and sequential assembly in nanosystems. Although the use of biological nanomotors puts intrinsic constraints on the conditions under which they can be assembled and used in biohybrid devices, many of their sophisticated tasks are still poorly mimicked by synthetic analogues. Understanding the details of how these little nanomachines convert chemical energy into controlled movements will nevertheless inspire new approaches to engineer synthetic counterparts that might some day be used under harsher conditions, operate at more extreme temperatures, or simply have longer shelf lives.

Certain stages of the materials production process might one day be replaced by nanomotor-driven sequential self-assembly, allowing much more control at the molecular level. Biological motors are already being used to drive the efficient fabrication of complex nanoscopic and mesoscopic structures, such as nanowires [31] and supramolecular assemblies. Techniques for precision control of nanomotors that read DNA are also being used to engineer integrated systems for rapid DNA detection and analysis (www.nanobiosym.com). The specificity and control of assembly and transport shown by biological systems offers many opportunities to those interested in assembly of complex nanosystems. Most importantly, the intricate schemes of proofreading and damage repair—features that have not yet been realized in any manmade nanosystems—should provide inspiration for those interested in producing synthetic systems capable of similarly complex tasks.

Acknowledgments

We thank Sheila Luna, Christian Brunner and Jennifer Wilson for the artwork, and all of our collaborators who contributed thoughts and experiments. At the same time, we apologize to all authors whose work we could not cite owing to space limitations.

Correspondence and requests for materials should be addressed to A.G. or V.V.

References

1 Rodnina, M.V. and Wintermeyer, W. (2001) Fidelity of aminoacyl-tRNA selection on the ribosome: kinetic and structural mechanisms. *Annual Review of Biochemistry*, **70**, 415–435.

2 Kunkel, T.A. (2004) DNA replication fidelity. *The Journal of Biological Chemistry*, **279**, 16895–16898.

3 Erie, D.A., Hajiseyedjavadi, O., Young, M.C. and von Hippel, P.H. (1993) Multiple RNA polymerase conformations and GreA: control of the fidelity of transcription. *Science*, **262**, 867.

4 Liu, D.R., Magliery, T.J., Pastrnak, M. and Schultz, P.G. (1997) Engineering a tRNA and aminoacyl-tRNA synthetase for the site-specific incorporation of unnatural amino acids into proteins in vivo. *Proceedings of the National Academy of Sciences of the United States of America*, **94**, 10092–10097.

5 Bustamante, C., Smith, S.B., Liphardt, J. and Smith, D. (2000) Single-molecule studies of DNA mechanics. *Current Opinion in Structural Biology*, **10**, 279–285.

6 Davenport, R.J., Wuite, G.J.L., Landick, R. and Bustamante, C. (2000) Single-molecule study of transcriptional pausing and arrest by *E. coli* RNA polymerase. *Science*, **287**, 2497–2500.

7 Greulich, K.O. (2005) Single-Molecule Studies on DNA and RNA. *ChemPhysChem*, **6**, 2459–2471.

8 Wang, M.D. *et al.* (1998) Force and velocity measured for single molecules of RNA polymerase. *Science*, **282**, 902–907.

9 Wuite, G.J., Smith, S.B., Young, M., Keller, D. and Bustamante, C. (2000) Single-molecule studies of the effect of template tension on T7 DNA polymerase activity. *Nature*, **404**, 103–106.

10 Smith, S.B., Cui, Y. and Bustamante, C. (1996) Overstretching B-DNA: The elastic response of individual double-stranded and single-stranded DNA molecules. *Science*, **271**, 795.

11 Smith, S.B., Finzi, L. and Bustamante, C. (1992) Direct mechanical measurements of the elasticity of single DNA molecules by using magnetic beads. *Science*, **258**, 1122.

12 Williams, M.C. and Rouzina, I. (2002) Force spectroscopy of single DNA and RNA molecules. *Current Opinion in Structural Biology*, **12**, 330–336.

13 Bustamante, C., Bryant, Z. and Smith, S.B. (2003) Ten years of tension: single-molecule DNA mechanics. *Nature*, **421**, 423–427.

14 Jeney, S., Stelzer, E.H., Grubmuller, H. and Florin, E.L. (2004) Mechanical properties of single motor molecules studied by three-dimensional thermal force probing in optical tweezers. *ChemPhysChem*, **5**, 1150–1158.

15 Mehta, A.D. (1999) Single-molecule biomechanics with optical methods. *Science*, **283**, 1689–1695.

16 Mogilner, A. and Oster, G. (2003) Polymer motors: pushing out the front and pulling up the back. *Current Biology*, **13**, 721–733.

17 Schnitzer, M.J., Visscher, K. and Block, S.M. (2000) Force production by single

kinesin motors. *Nature Cell Biology*, **2**, 718–723.

18 Strick, T., Allemand, J.F., Croquette, V. and Bensimon, D. (2001) The manipulation of single biomolecules. *Physics Today*, **54**, 46–51.

19 Ha, T. (2001) Single-molecule fluorescence methods for the study of nucleic acids. *Current Opinion in Structural Biology*, **11**, 287–292.

20 Kapanidis, A.N. *et al.* (2006) Initial transcription by RNA polymerase proceeds through a DNA scrunching mechanism. *Science*, **314**, 1144–1147.

21 Keller, R.A. *et al.* (1996) Single-molecule fluorescence analysis in solution. *Applied Spectroscopy*, **50**, 12A–32A.

22 Mullis, K.B. (1993) *The Polymerase Chain Reaction*, Nobel Lecture.

23 van Hest, J.C.M. and Tirrell, D.A. (2001) Protein-based materials, toward a new level of structural control. *Chemical Communications*, **19**, 1897–1904.

24 Fodor, S.P. *et al.* (1993) Multiplexed biochemical assays with biological chips. *Nature*, **364**, 555–556.

25 Merrifield, R.B. (1965) Automated synthesis of peptides. *Science*, **150**, 178–185.

26 Ratner, D.M., Swanson, E.R. and Seeberger, P.H. (2003) Automated synthesis of a protected N-linked glycoprotein core pentasaccharide. *Organic Letters*, **5**, 4717–4720.

27 Ball, P. (2001) It all falls into place. *Nature*, **413**, 667–668.

28 Pavel, I.S. (2005) *Assembly of gold nanoparticles by ribosomal molecular machines*, PhD thesis, Univ. Texas at Austin.

29 Whaley, S.R., English, D.S., Hu, E.L., Barbara, P.F. and Belcher, A.M. (2000) Selection of peptides with semiconductor binding specificity for directed nanocrystal assembly. *Nature*, **405**, 665–668.

30 Chen, H.L. and Goel, A. (2005) in *DNA Computing. Lecture Notes in Computer Science*, Vol. 3384 Springer, Berlin/Heidelberg, pp. 62–75.

31 Hess, H. *et al.* (2005) Molecular self-assembly of 'nanowires'and 'nanospools' using active transport. *Nano Letters*, **5**, 629–633.

32 Winfree, E. and Bekbolatov, R. (2004) in *DNA Computing. Lecture Notes in Computer Science*, Vol 2943 Springer, Berlin/Heidelberg, pp. 126–144.

33 Choi, I.S., Bowden, N. and Whitesides, G.M. (1999) Macroscopic, hierarchical, two-dimensional self-assembly. *Angewandte Chemie (International Edition in English)*, **38**, 3078–3081.

34 Whitesides, G.M. and Boncheva, M. (2002) Beyond molecules: self-assembly of mesoscopic and macroscopic components. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 4769–4774.

35 Caviston, J.P. and Holzbaur, E.L. (2006) Microtubule motors at the intersection of trafficking and transport. *Trends in Cell Biology*, **16**, 530–537.

36 Howard, J. (2001) *Mechanics of Motor Proteins and the Cytoskeleton*, Sinauer, Sunderland, Massachusetts.

37 Lakadamyali, M., Rust, M. and Zhuang, X. (2006) Ligands for clathrin-mediated endocytosis are differentially sorted into distinct populations of early endosomes. *Cell*, **124**, 997–1009.

38 Lakadamyali, M., Rust, M.J., Babcock, H.P. and Zhuang, X. (2003) Visualizing infection of individual influenza viruses. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 9280–9285.

39 Månsson, A. and Linke, H. (2007) Controlled Nanoscale Motion. *Proc. Nobel Symp. 131*, Vol 711, Springer, Berlin.

40 Miki, H., Okada, Y. and Hirokawa, N. (2005) Analysis of the kinesin superfamily: insights into structure and function. *Trends in Cell Biology*, **15**, 467–476.

41 Rust, M.J., Lakadamyali, M., Zhang, F. and Zhuang, X. (2004) Assembly of

endocytic machinery around individual influenza viruses during viral entry. *Nature Structural & Molecular Biology*, **11**, 567–573.

42 Sotelo-Silveira, J.R., Calliari, A., Kun, A., Koenig, E. and Sotelo, J.R. (2006) RNA trafficking in axons. *Traffic (Copenhagen, Denmark)*, **7**, 508–515.

43 Vale, R.D. (2003) The molecular motor toolbox for intracellular transport. *Cell*, **112**, 467–480.

44 Vallee, R.B. and Sheetz, M.P. (1996) Targeting of motor proteins. *Science*, **271**, 1539–1544.

45 Vale, R.D., Reese, T.S. and Sheetz, M.P. (1985) Identification of a novel force-generating protein, kinesin, involved in microtubule-based motility. *Cell*, **42**, 39–50.

46 Guzik, B.W. and Goldstein, L.S. (2004) Microtubule-dependent transport in neurons: steps towards an understanding of regulation, function and dysfunction. *Current Opinion in Cell Biology*, **16**, 443–450.

47 Gunawardena, S. and Goldstein, L.S. (2001) Disruption of axonal transport and neuronal viability by amyloid precursor protein mutations in Drosophila. *Neuron*, **32**, 389–401.

48 Gunawardena, S. and Goldstein, L.S. (2004) Cargo-carrying motor vehicles on the neuronal highway: transport pathways and neurodegenerative disease. *Journal of Neurobiology*, **58**, 258–271.

49 Mandelkow, E. and Mandelkow, E.-M. (2002) Kinesin motors and disease. *Trends in Cell Biology*, **12**, 585–591.

50 Ström, A.L. *et al.* (23 April 2008) Retrograde axonal transport and motor neuron disease. *Journal of Neurochemistry*, Preprint at ⟨http://www.ncbi.nlm.nih.gov/pubmed/18384644⟩

51 Mukhopadhyay, R. and Hoh, J.H. (2001) AFM force measurements on microtubule-associated proteins: the projection domain exerts a long-range repulsive force. *FEBS Letters*, **505**, 374–378.

52 Mizuno, N. *et al.* (2004) Dynein and kinesin share an overlapping microtubule-binding site. *The EMBO Journal*, **23**, 2459–2467.

53 Vale, R.D. and Toyoshima, Y.Y. (1988) Rotation and translocation of microtubules in vitro induced by dyneins from Tetrahymena cilia. *Cell*, **52**, 459–469.

54 Wang, Z., Khan, S. and Sheetz, M.P. (1995) Single cytoplasmic dynein molecule movements: characterization and comparison with kinesin. *Biophysical Journal*, **69**, 2011–2023.

55 Reck-Peterson, S.L. *et al.* (2006) Single-molecule analysis of dynein processivity and stepping behavior. *Cell*, **126**, 335–348.

56 Seitz, A. and Surrey, T. (2006) Processive movement of single kinesins on crowded microtubules visualized using quantum dots. *The EMBO Journal*, **25**, 267–277.

57 Coppin, C.M., Finer, J.T., Spudich, J.A. and Vale, R.D. (1996) Detection of sub-8-nm movements of kinesin by high-resolution optical-trap microscopy. *Proceedings of the National Academy of Sciences of the United States of America*, **93**, 1913–1917.

58 Svoboda, K., Schmidt, C.F., Schnapp, B.J. and Block, S.M. (1993) Direct observation of kinesin stepping by optical trapping interferometry. *Nature*, **365**, 721–727.

59 Diehl, M.R., Zhang, K., Lee, H.J. and Tirrell, D.A. (2006) Engineering cooperativity in biomotor-protein assemblies. *Science*, **311**, 1468–1471.

60 Hunt, A.J. and Howard, J. (1993) Kinesin swivels to permit microtubule movement in any direction. *Proceedings of the National Academy of Sciences of the United States of America*, **90**, 11653–11657.

61 Clemmens, J. *et al.* (2003) Principles of microtubule guiding on microfabricated kinesin-coated surfaces: chemical and topographic surface patterns. *Langmuir*, **19**, 10967–10974.

62 Reuther, C., Hajdo, L., Tucker, R., Kasprzak, A.A. and Diez, S. (2006) Biotemplated nanopatterning of planar

surfaces with molecular motors. *Nano Letters*, **6**, 2177–2183.

63 Clemmens, J., Hess, H., Howard, J. and Vogel, V. (2003) Analysis of microtubule guidance by microfabricated channels coated with kinesin. *Langmuir*, **19**, 1738–1744.

64 Hess, H. *et al.* (2003) Molecular shuttles operating undercover: a new photolithographic approach for the fabrication of structured surfaces supporting directed motility. *Nano Letters*, **3**, 1651–1655.

65 Hiratsuka, Y., Tada, T., Oiwa, K., Kanayama, T. and Uyeda, T.Q. (2001) Controlling the direction of kinesindriven microtubule movements along microlithographic tracks. *Biophysical Journal*, **81**, 1555–1561.

66 Moorjani, S.G., Jia, L., Kackson, T.N. and Hancock, W.O. (2003) Lithographically patterned channels spatially segregate kinesin motor activity and effectively guide microtubule movements. *Nano Letters*, **3**, 633–637.

67 Bunk, R. *et al.* (2003) Actomyosin motility on nanostructured surfaces. *Biochemical and Biophysical Research Communications*, **301**, 783–788.

68 Sundberg, M. *et al.* (2006) Actin filament guidance on a chip: toward high-throughput assays and lab-on-a-chip applications. *Langmuir*, **22**, 7286–7295.

69 Clemmens, J. *et al.* (2004) Motor-protein 'roundabouts': microtubules moving on kinesin-coated tracks through engineered networks. *Lab Chip*, **4**, 83–86.

70 Hess, H., Clemmens, J., Howard, J. and Vogel, V. (2002) Surface imaging by self-propelled nanoscale probes. *Nano Letters*, **2**, 113–116.

71 Stracke, R., Böhm, K.J., Burgold, J., Schacht, H..-J. and Unger, E. (2000) Physical and technical parameters determining the functioning of a kinesin-based cell-free motor system. *Nanotechnology*, **11**, 52–56.

72 Brown, T.B. and Hancock, W.O. (2005) A polarized microtubule array for kinesin-powered nanoscale assembly and force generation. *Nano Letters*, **28**, 571–576.

73 Nitta, T., Tanahashi, A., Hirano, M. and Hess, H. (2006) Simulating molecular shuttle movements: towards computer-aided design of nanoscale transport systems. *Lab Chip*, **6**, 881–885.

74 Vogel, V. and Hess, H. (2007) in *Lecture Notes Proceedings Nobel Symposium*, Vol. 711 Springer, Berlin/Heidelberg, pp. 367–383.

75 Stracke, R., Bohm, K.J., Wollweber, L., Tuszynski, J.A. and Unger, E. (2002) Analysis of the migration behaviour of single microtubules in electric fields. *Biochemical and Biophysical Research Communications*, **293**, 602–609.

76 van den Heuvel, M.G., de Graaff, M.P. and Dekker, C. (2006) Molecular sorting by electrical steering of microtubules in kinesin-coated channels. *Science*, **312**, 910–914.

77 Darnton, N., Turner, L., Breuer, K. and Berg, H.C. (2004) Moving fluid with bacterial carpets. *Biophysical Journal*, **86**, 1863–1870.

78 Hiratsuka, Y., Miyata, M., Tada, T. and Uyeda, T.Q. (2006) A microrotary motor powered by bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 13618–13623.

79 Bachand, G.D., Rivera, S.B., Carroll-Portillo, A., Hess, H. and Bachand, M. (2006) Active capture and transport of virus particles using a biomolecular motor-driven, nanoscale antibody sandwich assay. *Small*, **2**, 381–385.

80 Hirabayashi, M. *et al.* (2006) Malachite green-conjugated microtubules as mobile bioprobes selective for malachite green aptamers with capturing/releasing ability. *Biotechnology and Bioengineering*, **94**, 473–480.

81 Martin, B.D. *et al.* (2006) An engineered virus as a bright fluorescent tag and scaffold for cargo proteins: capture and transport by gliding microtubules. *Journal of Nanoscience and Nanotechnology*, **6**, 2451–2460.

82 Muthukrishnan, G., Hutchins, B.M., Williams, M.E. and Hancock, W.O. (2006) Transport of semiconductor nanocrystals by kinesin molecular motors. *Small*, **2**, 626–630.

83 Taira, S. *et al.* (2006) Selective detection and transport of fully matched DNA by DNA-loaded microtubule and kinesin motor protein. *Biotechnology and Bioengineering*, **95**, 533–538.

84 Hess, H., Clemmens, J., Qin, D., Howard, J. and Vogel, V. (2001) Light-controlled molecular shuttles made from motor proteins carrying cargo on engineered surfaces. *Nano Letters*, **1**, 235–239.

85 Hess, H., Howard, J. and Vogel, V. (2002) A piconewton forcemeter assembled from microtubules and kinesins. *Nano Letters*, **2**, 1113–1115.

86 Boal, A.K., Bachand, G.D., Rivera, S.B. and Bunker, B.C. (2006) Interactions between cargo-carrying biomolecular shuttles. *Nanotechnology*, **17**, 349–354.

87 Diez, S. *et al.* (2003) Stretching and transporting DNA molecules using motor proteins. *Nano Letters*, **3**, 1251–1254.

88 Brunner, C., Wahnes, C. and Vogel, V. (2007) Cargo pick-up from engineered loading stations by kinesin driven molecular shuttles. *Lab on a Chip*, **7**, 1263–1271.

89 Ramachandran, S., Ernst, K.H., Bachand, G.D., Vogel, V. and Hess, H. (2006) Selective loading of kinesin-powered molecular shuttles with protein cargo and its application to biosensing. *Small*, **2**, 330.

90 Dinu, C.Z. *et al.* (2006) Parallel manipulation of bifunctional DNA molecules on structured surfaces using kinesin-driven microtubules. *Small*, **2**, 1090–1098.

91 Soldati, T. and Schliwa, M. (2006) Powering membrane traffic in endocytosis and recycling. *Nature Reviews. Molecular Cell Biology*, **7**, 897–908.

92 Bachand, M., Trent, A.M., Bunker, B.C. and Bachand, G.D. (2005) Physical factors affecting kinesin-based transport of synthetic nanoparticle cargo. *Journal of Nanoscience and Nanotechnology*, **5**, 718–722.

93 Du, Y.Z. *et al.* (2005) Motor protein nanobiomachine powered by self-supplying ATP. *Chemical Communications*, 2080–2082.

94 Kufer, S.K., Puchner, E.M., Gumpp, H., Liedl, T. and Gaub, H.E. (2008) Single-molecule cut-and-paste surface assembly. *Science*, **319**, 594–596.

95 Chakravarty, A., Howard, L. and Compton, D.A. (2004) A mechanistic model for the organization of microtubule asters by motor and non-motor proteins in a mammalian mitotic extract. *Molecular Biology of the Cell*, **15**, 2116–2132.

96 Surrey, T., Nedelec, F., Leibler, S. and Karsenti, E. (2001) Physical properties determining self-organization of motors and microtubules. *Science*, **292**, 1167–1171.

97 Doot, R.K., Hess, H. and Vogel, V. (2007) Engineered networks of oriented microtubule filaments for directed cargo transport. *Soft Matter*, **3**, 349–356.

98 Beeg, J. *et al.* (2008) Transport of beads by several kinesin motors. *Biophysical Journal*, **94**, 532.

99 Henry, T., Gorvel, J.P. and Meresse, S. (2006) Molecular motors hijacking by intracellular pathogens. *Cellular Microbiology*, **8**, 23–32.

100 Soo, F.S. and Theriot, J.A. (2005) Large-scale quantitative analysis of sources of variation in the actin polymerization-based movement of *Listeria monocytogenes*. *Biophysical Journal*, **89**, 703–723.

101 Rietdorf, J. *et al.* (2001) Kinesin-dependent movement on microtubules precedes actin-based motility of vaccinia virus. *Nature Cell Biology*, **3**, 992–1000.

102 Smith, G.A., Gross, S.P. and Enquist, L.W. (2001) Herpes viruses use bidirectional fast-axonal transport to spread in sensory neurons. *Proceedings of the National*

Academy of Sciences of the United States of America, **98**, 3466–3470.

103 Döhner, K., Nagel, C.-H. and Sodeik, B. (2005) Viral stop-and-go along microtubules: taking a ride with dynein and kinesins. *Trends in Microbiology*, **13**, 320–327.

104 Sodeik, B. (2002) Unchain my heart, baby let me go: the entry and intracellular transport of HIV. *Cell Biol*, **159**, 393–395.

105 Kulkarni, R.P., Wu, D.D., Davis, M.E. and Fraser, S.E. (2005) Quantitating intracellular transport of polyplexes by spatio-temporal image correlation spectroscopy. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 7523–7528.

106 Suh, J., Wirtz, D. and Hanes, J. (2003) Efficient active transport of gene nanocarriers to the cell nucleus. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 3878–3882.

107 Goel, A., Astumian, R.D. and Herschbach, D. (2003) Tuning and switching a DNA polymerase motor with mechanical tension. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 9699–9704.

108 Donlin, M.J., Patel, S.S. and Johnson, K.A. (1991) Kinetic partitioning between the exonuclease and polymerase sites in DNA error correction. *Biochemistry*, **30**, 538–546.

109 Fersht, A.R., Knill-Jones, J.W. and Tsui, W.C. (1982) Kinetic basis of spontaneous mutation. Misinsertion frequencies, proofreading specificities and cost, of proofreading by DNA polymerases of *Escherichia coli*. *Journal of Molecular Biology*, **156**, 37–51.

110 Hopfield, J.J. (1974) Kinetic proofreading: a new mechanism for reducing errors in biosynthetic processes requiring high specificity. *Proceedings of the National Academy of Sciences of the United States of America*, **71**, 4135–4139.

111 Hopfield, J.J. (1980) The energy relay: a proofreading scheme based on dynamic cooperativity and lacking all characteristic symptoms of kinetic proofreading in DNA replication and protein synthesis. *Proceedings of the National Academy of Sciences of the United States of America*, **77**, 5248.

112 Rodnina, M.V. and Wintermeyer, W. (2001) Ribosome fidelity: tRNA discrimination, proofreading and induced fit. *Trends in Biochemical Sciences*, **26**, 124–130.

113 Wang, D. and Hawley, D.K. (1993) Identification of a $3' \rightarrow 5'$exonuclease activity associated with human RNA polymerase II. *Proceedings of the National Academy of Sciences of the United States of America*, **90**, 843–847.

114 Cavalli, V., Kujala, P., Klumperman, J. and Goldstein, L.S. (2005) Sunday Driver links axonal transport to damage signaling. *The Journal of Cell Biology*, **168**, 775–787.

115 Mandelkow, E.M., Stamer, K., Vogel, R., Thies, E. and Mandelkow, E. (2003) Clogging of axons by tau, inhibition of axonal traffic and starvation of synapses. *Neurobiology of Aging*, **24**, 1079–1085.

116 Goel, A., Ellenberger, T., Frank-Kamenetskii, M.D. and Herschbach, D. (2002) Unifying themes in DNA replication: reconciling single molecule kinetic studies with structural data on DNA polymerases. *Journal of Biomolecular Structure & Dynamics*, **19**, 571–584.

117 Guydosh, N.R. and Block, S.M. (2006) Backsteps induced by nucleotide analogs suggest the front head of kinesin is gated by strain. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 8054–8059.

118 Spudich, J. (2006) Molecular motors take tension in stride. *Cell*, **126**, 242–244.

119 Vale, R.D. and Milligan, R.A. (2000) The way things move: looking under the hood of molecular motor proteins. *Science*, **288**, 88–95.

120 Veigel, C., Schmitz, S., Wang, F. and Sellers, J.R. (2005) Load-dependent kinetics of myosin-V can explain its high

processivity. *Nature Cell Biology*, **7**, 861–869.

121 Maier, B., Bensimon, D. and Croquette, V. (2000) Replication by a single DNA polymerase of a stretched single-stranded DNA. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 12002–12007.

122 Abbondanzieri, E.A., Greenleaf, W.J., Shaevitz, J.W., Landick, R. and Block, S.M. (2005) Direct observation of base-pair stepping by RNA polymerase. *Nature*, **438**, 460–465.

123 Shaevitz, J.W., Abbondanzieri, E.A., Landick, R. and Block, S. (2003) Backtracking by single RNA polymerase molecules observed at near-base pair resolution. *Nature*, **426**, 684–687.

124 Sinha, D.K., Bhalla, U.S. and Shivashankar, G.V. (2004) Kinetic measurement of ribosome motor stalling force. *Applied Physics Letters*, **85**, 4789–4791.

125 Astumian, R.D. (1997) Thermodynamics and kinetics of a Brownian motor. *Science*, **276**, 917–922.

126 Bustamante, C., Keller, D. and Oster, G. (2001) The physics of molecular motors. *Accounts of Chemical Research*, **34**, 412–420.

127 Goel, A., Frank-Kamenetskii, M.D., Ellenberger, T. and Herschbach, D. (2001) Tuning DNA 'strings': modulating the rate of DNA replication with mechanical tension. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 8485–8489.

128 Goel, A. and Herschbach, D.R. (2003) Controlling the speed and direction of molecular motors that replicate DNA. *Proc SPIE*, **5110**, 63–68.

129 Mallik, R., Carter, B.C., Lex, S.A., King, S.J. and Gross, S.P. (2004) Cytoplasmic dynein functions as a gear in response to load. *Nature*, **427**, 649–652.

130 Rosenfeld, S.S., Fordyce, P.M., Jefferson, G.M., King, P.H. and Block, S.M. (2003) Stepping and stretching. How kinesin uses internal strain to walk processively. *The Journal of Biological Chemistry*, **278**, 18550–18556.

131 Andricioaei, I., Goel, A., Herschbach, D. and Karplus, M. (2004) Dependence of DNA polymerase replication rate on external forces: a model based on molecular dynamics simulations. *Biophysical Journal*, **87**, 1478–1497.

132 Liu, H. *et al.* (2002) Control of a biomolecular motor-powered nanodevice with an engineered chemical switch. *Nature Mater*, **1**, 173–177.

133 Konishi, K., Uyeda, T.Q. and Kubo, T. (2006) Genetic engineering of a Ca(2+) dependent chemical switch into the linear biomotor kinesin. *FEBS Letters*, **580**, 3589–3594.

134 Nomura, A., Uyeda, T.Q., Yumoto, N. and Tatsu, Y. (2006) Photo-control of kinesin-microtubule motility using caged peptides derived from the kinesin C-terminus domain. *Chemical Communications*, **1**, 3588–3590.

135 Ionov, L., Stamm, M. and Diez, S. (2006) Reversible switching of microtubule motility using thermoresponsive polymer surfaces. *Nano Letters*, **6**, 1982–1987.

136 van den Heuvel, M.G.L. and Dekker, C. (2007) Motor proteins at work for nanotechnology. *Science*, **317**, 333–336.

137 Browne, W.R. and Feringa, B.L. (2006) Making molecular machines work. *Nature Nanotechnology*, **1**, 25–35.

138 Hess, H. *et al.* (2002) Ratchet patterns sort molecular shuttles. *Applied Physics A*, **75**, 309–313.

Part Four:
Innovative Disease Treatments and Regenerative Medicine

Nanotechnology, Volume 5: Nanomedicine. Edited by Viola Vogel

ISBN: 978-3-527-31736-3

9
Mechanical Forces Matter in Health and Disease: From Cancer to Tissue Engineering

Viola Vogel and Michael P. Sheetz

9.1
Introduction: Mechanical Forces and Medical Indications

One of our earliest experiences showing that mechanical forces matter goes back to when we got our first blisters. Excessive friction causes a tear between the upper layer of the skin – the epidermis – and the layers beneath. When these skin layers – which in healthy skin are held together by cell–cell adhesion complexes – begin to separate, the resultant pocket fills with serum or blood. In some people, who have inherited skin diseases, the blistering occurs much more easily, and studies of point mutations that cause easy blistering have provided considerable insights into the underlying molecular mechanisms. Molecular defects can exist in different intracellular and extracellular proteins that are responsible for weakening the mechanical strength of cell–cell adhesions. The proteins implicated by genetic analysis include keratins, laminins, collagens and integrins [1–3]. Unfortunately, exactly how mutations in these proteins regulate the mechanical stability of the linkages that cells form with their environment remains unknown.

Mechanical forces acting on cells also affect our lives in many other, often unexpected, ways. Regular exercise, for example, not only strengthens our body tone but also offers protection against mortality by delaying the onset of various diseases. It is thought that physical training reduces the chance of chronic heart diseases, atherosclerosis and also type 2 diabetes [4]. But how can exercise have such a profound impact on so many diseases? Chronic low-grade systemic inflammation is a feature of these and many other chronic diseases that have been correlated with elevated levels of several cytokines [5–7]. By yet unknown mechanisms, it is suggested that regular exercise induces anti-inflammatory processes, thus suppressing the production of pro-inflammatory signaling proteins [5, 8].

Many more severe diseases for which we do not have cures also have a mechanical origin, or show abnormalities in cellular mechanoresponses. These range from cancer to cardiovascular disorders, from osteoporosis to other aging-related diseases. In the case of many cancers, the cells grow inappropriately and with the wrong mechanoresponse, which in turn destroys normal tissue mechanics and often also

Nanotechnology, Volume 5: Nanomedicine. Edited by Viola Vogel

ISBN: 978-3-527-31736-3

tissue function [9–11]. While cardiovascular diseases have many forms, cardiac hypertrophy, plaque formation and heart repair are obvious cases where mechanosensory functions are important [12–14]. Abnormal mechanical forces can trigger an aberrant proliferation of endothelial and smooth muscle cells, as observed in the progression of vascular diseases such as atherosclerosis [15]. There is furthermore emerging evidence that immune synapse formation is a mechanically driven process [16]. Finally, damaged tissue is often repaired by new cells that differentiate from pluripotent cells to finally replace and regenerate the damaged regions. Successful healing includes re-establishing the proper mechanical tissue characteristics; even bioscaffolds that are used in reconstructive surgery heal best if they are mechanically exercised [14, 17]. Thus, from molecules to tissue, although the mechanical aspects are recognized as being critical, relatively little has yet been done to correlate mechanical effects with biochemical signal changes, and how these impact clinical outcomes.

There is, therefore, overwhelming evidence that physical and not just biochemical stimuli matter in tissue growth and repair in health and disease. But how do cells sense mechanical forces? A complete answer cannot yet be given as too few techniques have been available in the past to explore this question. However, the broad availability of nanoanalytical and nanomanipulation tools is beginning to have impact. This tool chest provides novel opportunities to decipher how physical and biochemical factors, in combination, can orchestrate the hierarchical control of cell and tissue functions (we will illustrate this point with some concrete examples later in the chapter). The diversity of biological forms in different organisms most likely belies a wide range of mechanosensing mechanisms that are specifically engineered to provide the desired morphology.

To summarize, based on the progress that has been made recently in the field of cell biomechanics, it is now clear that individual cells are dramatically affected in their functions, from growth to differentiation, by the mechanical properties of their environments and by externally applied forces (for reviews, see Refs [18–21]). But the question remains: How do cells sense mechanical forces, and how are mechanical stimuli translated at the nanoscale into biochemical signal changes that ultimately regulate cell function? A few examples are illustrated here of how physical junctions are formed between cells and their environment, how mechanical forces acting on molecules associated with junctions regulate their functional states, and what the downstream implications might be on cell signaling events. Considering the complexity of the puzzle, this chapter cannot provide a comprehensive review; rather, we will focus on describing a few selected molecular players involved in mechanochemical signal conversion, followed by a discussion of the associated signaling pathways and subsequent cellular responses, and then concluding on the role of physical stimuli in various diseases. Once the molecular pathways are identified, and the mechanisms deciphered by which force regulates diverse cell functions, the development of new drugs and therapies will surely emerge. In particular, it is expected that in the future, a number of diseases associated with altered mechanoresponses will be resolved more efficiently by treating the source of the problem, rather than the symptoms.

9.2
Force-Bearing Protein Networks Hold the Tissue Together

The search for proteins that are structurally altered if mechanically stretched, and which could thus serve as force sensors for cells, should start in junctions between cells and their environments. The focus should first be on the junctions that experience the highest tensile forces. The force-bearing elements in tissues are typically the cytoskeleton and extracellular matrix (ECM) fibers, and all the proteins that physically link the cell interior to the exterior. For different tissues, the major force-bearing elements can differ.

9.2.1
Cell–Cell Junctions

Some tissues such as epithelial and endothelial cell layers have barrier functions (Figure 9.1), where the majority of the force is born by the tight cell–cell junctions. These junctions couple the cell–cell adhesion molecules (cadherins) that hold the cells together to the cytoplasmic proteins that ultimately link cadherins to the actin, myosin and intermediate filaments in the cytoskeleton [22].

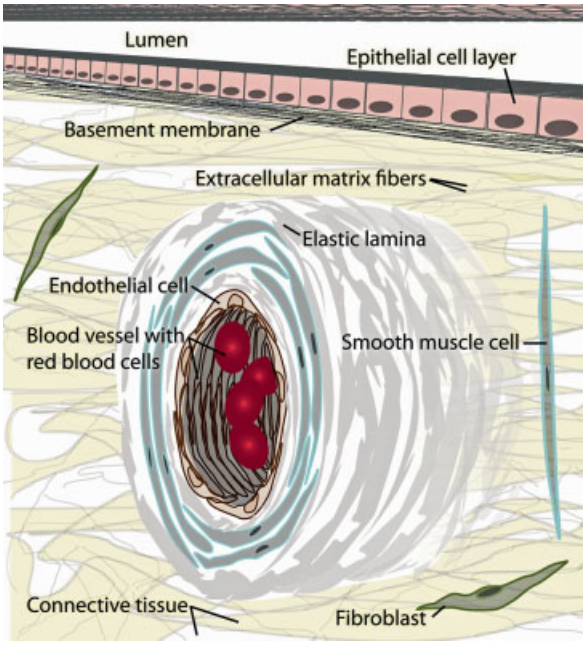


Figure 9.1 Schematic section through tissue, showing how layers of cells form a barrier that separates the connective tissue from the lumen of the lung or of the intestines (epithelial cells), or the blood vessels (endothelial cells). Within endothelial and epithelial cell layers, the cells are tightly connected to each other via cadherin junctions, while integrin junctions anchor cells to the basement membrane as well as to the extracellular matrix (ECM) of connective tissues.

Epithelial tissue lines both, the outside of the body (skin) and the cavities that are connected to the outside, such as the lungs and the gastrointestinal tract. Epithelial cells assume packing geometries in junctional networks that are characterized by different cell shapes, number of neighbor cells and contact areas. The development of specific packing geometries is tightly controlled [23].

Endothelial cells line the tight barrier between the circulating blood and the surrounding vessel wall. A synchronized migration of endothelial cells is required in order to grow blood vessels (the process of *angiogenesis*). When a new blood vessel is forming, for example in response to a lack of oxygen, the endothelial cells must maintain their cell–cell contacts, remain anchored to the basement membrane, and form curved continuous surfaces [24]; otherwise, the walls of the growing vessels would become leaky. Blood vessel formation is thus a tightly regulated process.

9.2.2 Cell–Matrix Junctions

In contrast to the tight cell–cell junctions, cells can also form junctions with surrounding extracellular fibers. The ECM, which is abundant in connective tissue, includes the interstitial matrix and basement membranes. The ECM provides structural support to the cells (Figure 9.1), in addition to performing many other important functions that regulate cell behavior. Cell–matrix contacts are formed by integrins; these molecules can link various ECM proteins, including fibronectin, vitronectin, laminins and collagens, via cytoplasmic adopter proteins to the cytoskeleton. During the formation or regeneration of tissue, major cell movements occur on or through the ECM, such that the cell–matrix junctions enable and facilitate integrin-mediated tissue growth, remodeling and repair processes. Integrins are also required for the assembly of the ECM (for reviews, see Refs [2]). The fibronectin matrix, the assembly of which is upregulated in embryogenesis and wound-healing processes, often serves as an early provisional matrix that is reinforced at later stages, for example, by collagen deposition [27]. Integrins thus mediate the regulatory functions of the ECM on cell migration, growth and differentiation. During wound healing, angiogenesis and tumor invasion, cells often change their expression profiles of fibronectin-binding integrins [28, 29]. Integrin–matrix interactions thus play central roles in regulating cell migration, invasion and extra- and intra-vasation (i.e. moving from the vasculature to the tissue, or *vice versa*), as well as in platelet interaction and wound healing [24, 29–36]. The functional roles of these interactions in health and disease will be discussed in much more detail below.

The forces acting on cell–cell or cell–matrix junctions can either be applied externally, or generated by the contractile cytoskeleton. Shear stresses due to the flow of blood, urine and of other body fluids impart forces on either the endothelial blood vessel linings, the linings of the epithelial urinary tract, as well as on bone cells, respectively, and are known to actively influence cell morphology, function and tissue remodeling (for reviews, see Refs [12, 37–41]). Lung expansion and contraction imposes great strain on lung tissue, and mechanical forces exerted on the lung epithelium are a major regulator of fetal lung development, as well as of the overall

pulmonary physiology [42]. Mechanical exercising of the lung also triggers the release of surfactants onto the epithelial surface [43]. Consequently, the levels of force generated and transmitted through cell–cell and cell–matrix junctions can change drastically with time, and between different organ tissues.

The forces that cells apply to their neighbors and matrices are furthermore dependent on the rigidity of their environment. Cell-generated tractile forces are lowest for soft tissues and increase with the rigidity of the organ. The brain is one of the softest tissues, whereas bone cells find themselves in one of the stiffest microenvironments of the body [18]. Yet, in all of these tissues the cells generate forces that provide the basis of active mechanosensing and mechanochemical signal conversion processes. The formation of force-bearing protein networks that connect the contractile cytoskeleton of cells with their surroundings is essential to prevent cell apoptosis of most normal cells.

What is missing is a mechanistic understanding of how the forces that are applied to cells are locally sensed and finally regulate a collective response of many cells to produce the proper tissue morphology and morphological transformations (Figure 9.2). Although it may be general for all tissues, in endothelia there is a

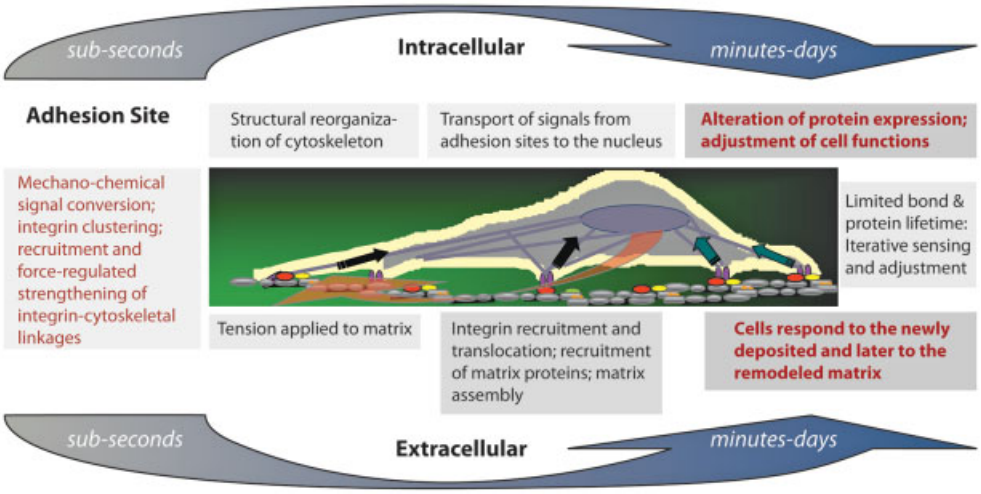


Figure 9.2 Sequential cellular processes of adhesion, mechanosensing and responses with the associated time scales. Initiated by cell adhesion, a cell responds to its environment by subsequent events that involve mechanosensing, reorganization of the cytoskeleton, adjustment of protein expression patterns and, in a secondary feedback loop, remodeling of the extracellular matrix. Initially, cells will sense the mechanical features of their environment, which will cause rapid motility and signaling responses. As the cell pulls on the environment, it will modify the extracellular matrix and create new signals, such as those originating from fibronectin stretching and unfolding. Intracellular signals will alter the expression pattern of the cell and, over time, the cellular forces and cell-generated matrices will change the cell shape. At any stage, extracellular signals, such as hormones or external mechanical stimuli, can cause acute changes that will set off a further round of cell and matrix modifications.

polarity and bending that must be controlled over many cell lengths. Studies of the development of fly wings and convergent extension in frogs have provided some important clues about mechanisms that can establish an axis in a tissue that would then result in axial contractions. In the fly wing, there are gradients of proteins that affect wing organization by influencing the physical properties [23, 44, 45], and some of those proteins are asymmetrically distributed in the hexagonal wing cells. In many tissues, however, the cells can move and change partners while they change tissue morphology in a stereotypical way, indicating that the multicellular coordination does not solely rely upon stationary protein complexes but rather is sensitive to intercellular forces or curvature.

To better understand the cellular nanomachinery by which cells sense mechanical stimuli, and how forces might synchronize cellular responses, it should be noted that the cellular nanomachinery is subjected both to exogenously applied forces and to cell-generated forces that the cells apply locally to the ECM and neighboring cells. As compressive forces on cells are primarily counterbalanced by the hydrostatic pressure of the cell volume that is contained by the plasma membrane, we will focus here entirely on the impact of tensile forces on proteins and protein networks and the subsequent changes in cell signaling.

When cells stretch their proteins, the protein structural changes may represent one important motif by which mechanical factors can be translated into biochemical signal changes in a variety of tissues and cell types (Figure 9.3). Many proteins are involved in force-bearing networks that connect the cell interior with the exterior, and they all are potential candidate proteins for mechanosensors (for reviews, see Refs [21, 26, 46–54]).

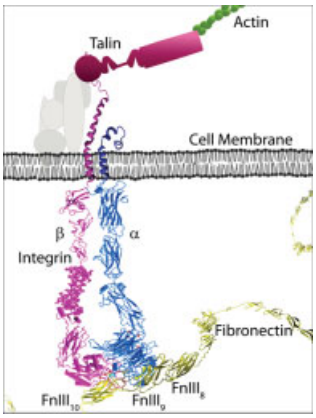


Figure 9.3 The integrin junction connecting the contractile actomyosin cytoskeleton with the extracellular matrix. In the schematic structure shown here, the integrin $\alpha_v\beta_3$ forms a complex with the extracellular matrix protein fibronectin via its cell binding peptide, RGD. In the cell interior, talin couples the cytoplasmic integrin tails to an actin filament. The stretching of talin leads to a reinforcement of the talin–actin linkage through the recruitment of further proteins that are subsequently involved in downstream cell signaling events. Particularly, the recruitment and stretching of p130cas regulates cell signaling events due to its phosphorylation, which is upregulated when stretched.

As cells actively bind, stretch and remodel their surroundings, they use a variety of specialized adhesion structures [25, 56], and their molecular composition will be discussed below (see Section 9.4). Once formed, the first contacts either mature rapidly or break (see Sections 9.5 and 9.7). These structures mechanically link the cell cytoskeleton and force-generating machinery within the cell to the ECM. Intracellular traction can thus generate large forces on the adhesive junctions – forces which are easily visualized as strain applied by cells to stretchable substrates [57–59], as discussed in Section 9.8. In addition, focal contacts are not passively resistant to force, but force actively induces focal contact strengthening through the recruitment of additional focal adhesion proteins, and finally initiates intracellular signaling events [60–64] (Section 9.6). Cell generated forces allow for rigidity sensing (Section 9.9), and causes matrix assembly and remodeling (Section 9.10). The matrix in turn regulates cell motility (Section 9.11). Ultimately, the structure and composition of the adhesions play regulatory roles in tissue formation and remodeling, and also control whether cells derail and evolve into cancer cells or cause other disease conditions (Section 9.12).

9.3
Nanotechnology has Opened a new Era in Protein Research

The advent of nanotech tools, particularly atomic force microscopy (AFM) and optical tweezers [65–67], followed by atomistic simulations of the force-induced unfolding pathways [68], were a major milestone in recognizing the unique mechanical properties of proteins and other biopolymers. The first force measurements on single multimodular proteins were performed on titin, and revealed that the modules cannot be deformed continuously but rather that they ruptured sequentially. But do cells take advantage of switching protein function mechanically? The first functional significance of unfolding proteins upon rapid tissue extension, for example when overstretching a muscle, was seen in them serving as mechanical shock absorbers.

Beyond muscle tissue, protein unfolding might be a much more common theme by which cells sense and transduce a broad range of mechanical forces into distinct sets of biochemical signals that ultimately regulate cellular processes, including adhesion, migration, proliferation, differentiation and apoptosis. The results of recent studies have shown that force-induced protein unfolding does indeed occur in cells and in their surrounding matrices [51–55, 69–71].

9.3.1
Mechanochemical Signal Conversion and Mechanotransduction

How, then, is force translated at the molecular level into biochemical signal changes (mechanochemical signal conversion) that have the potential to alter cellular behavior (mechanotransduction)? Despite all the experimental indications, only limited information is available on how mechanical forces alter the structure–function relationship of proteins and thus coregulate cell-signaling events. After a decade of new insights into single molecule mechanics, a new field is beginning to emerge:

How can the force-induced mechanical unfolding of proteins and other biomolecules switch their functions?

Through careful investigations of the conformational changes of isolated proteins that are mechanically stretched *in vitro*, and through computational simulations that have provided high-resolution structural information of the unfolding pathways of proteins, key design principles are beginning to emerge that describe how intracellular, extracellular and transmembrane proteins might sense mechanical forces and convert them into biochemical signal changes as discussed below (for reviews, see Refs [21, 26, 72, 73]). Stretch has been shown experimentally to expose cryptic phosphorylation sites, resulting in the onset of a major signaling cascade [51], to increase the reactivity of cysteines [52], and also to induce fibronectin fibrillogenesis (for a review, see Ref. [26]). Yeast two-hybrid measurements, crystallographic analyses and high-performance steered molecular dynamics (SMD) calculations all indicate that the exposure of amphipathic helices (e.g. talin, α-actinin) will cause binding to unstrained proteins (vinculin) or to the membrane, as detailed below. Thus, it seems that not a single mechanism can account for all the mechanical activities sensed by cells. Consequently, there is a need to develop a detailed understanding of the mechanical steps in each function of interest, in order to elucidate which of these mechanisms is responsible, or whether a new one must be formulated.

Design principles are also emerging by which such mechanosensory elements are integrated into structural motifs of various proteins, the conformations of which can be switched mechanically (for reviews, see Refs [26, 47, 74–79]). Multidomain proteins that are large and have many interaction sites constitute a major class of potentially force-transducing proteins [26, 80]. Both, matrix and cytoskeletal proteins fall into this class; for example, the cytoskeletal (titin, alpha actinin, filamin, etc.) and membrane skeletal molecules (spectrin, dystrophin, ankyrin) have series of between four and 100 repeat domains that can be stretched over a range of forces. An important feature here is that the repeats are often structurally homologous, but differ in their mechanical stability. Indeed, the differences in the mechanical stability of individual domains determines the time-dependent order in which their structure is altered by force, and consequently the sequence in which the molecular recognition sites are switched by force. Multimodularity thus provides for a mechanism not only for sensing but also for transducing a broad range of strains into a graded alteration of biochemical functionalities. Matrix molecules also have multiple domains and presumably exhibit similar characteristics. In both cases, the stretching of molecules can either reveal sites which can bind to and activate other proteins that could start a signaling cascade, or they can destroy recognition sites that are exposed only under equilibrium conditions [26].

9.3.2
Mechanical Forces and Structure–Function Relationships

As tensile force can stabilize proteins in otherwise short-lived structural intermediates, deciphering how the structure–function relationship of proteins is altered by

mechanical forces may well open totally new avenues in biotechnology, systems biology, pharmaceutics and medicine. In order to summarize our current understanding and future opportunities, we will first identify the critical molecules that are involved in linking the cell outside to the inside, and then discuss current knowledge on the effect of force on protein structure and associated force-regulated changes of protein function, and the downstream consequences. Cellular mechanotransduction systems can then transduce these primary physical signals into biochemical responses. More complex physical factors, such as matrix rigidity and the microscale and nanoscale textures of their environments, can be measured by cells through integrated force- and geometry-dependent transduction processes. Thus, it is important to differentiate between the primary sensory processes, the transduction processes and the downstream mechanoresponsive pathways that integrate multiple biochemical signals from sensing and transduction events over space and time, as shown schematically in Figure 9.2. It has also been postulated that cytoskeletal filaments can directly transmit stresses to distant cytoskeletal transduction sites [81, 82], which would involve additional distant mechanosensory and transductional components. Even in those cases, the forces would be focused on sites where primary transduction would occur.

Beyond the unfolding of stretched proteins, there are also other mechanisms in place by which force can alter many biochemical activities (see Box 9.1). The specific force-induced changes in motor protein velocity can lead to stalling their movement or buckling of their respective filaments [74, 83, 84]. Stretch-sensitive ion channels exist where the membrane pressure can regulate the ion current [20, 85–87]. Finally, even the lifetime of the strongest noncovalent bonds that last days under equilibrium, break down within seconds under the tensile force generated by a single kinesin [88, 89]. Not surprisingly, some adhesive bonds have evolved that are not weakened but are strengthened by force; these are also referred to as 'catch bonds' (as reviewed in Refs [90–93]). However, most of these force-regulated processes do not have an evident link to changes in cellular-level functions, or the links are currently not understood. For example, motor protein velocity is not generally linked to mechanically induced changes in cellular function, and neither are the ion currents that accompany the stretch activation of ion channels. Thus, it is unclear whether observed mechanochemical responses are products of the primary transduction of mechanical

Box 9.1
Activities altered by force-induced structural alterations:

- Motor protein velocity
- Stretch-sensitive ion channels (bacteria, hearing, touch)
- Catch bonds (bacterial and lymphocyte rolling and firm adhesion)
- Outside-in cell signaling through stretch-induced alterations of ECM binding sites
- Cytoskeletal protein stretching – phosphorylation by Tyr kinases of cryptic tyrosine repeat domains.

stimulation, or are just part of secondary downstream signaling cascades. Electrophysiological measurements reveal that distinct types of ion channels are mechanically activated [85, 87, 94, 95]; however, the biochemical consequences of channel opening are currently unclear – as are the relationships to downstream mechanoresponsive signaling pathways. Motor proteins will change the rate of movement and ATP hydrolysis in response to load (for a review, see Ref. [96]); however, the molecular pathway linking myosin mutations and cardiac hypertrophy is very unclear [13, 97]. Although catch bonds have a very clear role in enabling cells to adhere to surfaces under flow conditions, the link to subsequent infection or extravasation has not been determined. In any of the systems that employ specific mechanosensors, further investigations are required to determine whether – and/or how – these are functionally linked with specific steps in the cellular functions that are altered by mechanical force.

In the following section, which relates to mechanosensitive processes, we will discuss a few selected proteins that are part of the physical network through which force is transmitted bidirectionally from the cell exterior to the interior, and vice versa. Attention is focused here on the primary changes that have been shown to produce biochemical changes that lead in turn to general signals, although many more possible mechanisms clearly exist.

9.4
Making the Very First Contacts

9.4.1
Molecular Players of Cell–Extracellular Matrix Junctions

Cell motility is regulated by the polymerization of actin which drives the protrusion of the leading edge of the cell. Cells use lamellipodia and filopodia to 'feel' their environment and to identify locations to which they can adhere. Lamellipodia are flat, thin extensions of the cell edge that are supported by branched actin networks, while filopodia are finger-like extensions of the cell surface supported by parallel bundles of actin filaments [98]. Both are involved in sensing the environment through cycles of extension and retraction, in the attachment of particles for phagocytosis, in the anchorage of cells on a substratum, and in the response to chemoattractants or other guidance cues [99, 100]. When cells encounter a ligand bound to an extracellular surface, the ligand might bind to a transmembrane protein and ultimately induce coupling of the transmembrane protein to the cytoskeleton. *Integrins* are the key transmembrane proteins that mediate cell matrix interactions. Some integrins can recognize the tripeptide RGD, which is found for example in fibronectin, vitronectin and other matrix molecules, while other integrins bind specifically to collagens and laminins. Once a first bond (or set of bonds) is formed, a competition sets in between the time taken for a bond to break again and the cellular processes that can stabilize an early adhesion. The bond lifetime, however, is significantly decreased if a high tensile force is applied to it [101]. For example, without force, fibronectin can bind to $\alpha_5\beta_1$

integrin for minutes before releasing, whereas a force of approximately 40 pN will cause release in milliseconds [102, 103].

To illustrate some of the general concepts, rather than providing a detailed literature review, we will now briefly describe one of the force-bearing junctions that connects an ECM protein, via integrins, to the cytoskeleton (Figure 9.3).

9.4.1.1 Fibronectin

Fibronectin is a dimeric protein of more than 440 kDa (Figure 9.4), which is a pervasive component of the ECM during development and within healing wounds [24–26, 104, 105]. Fibronectin is composed of three types of repeating

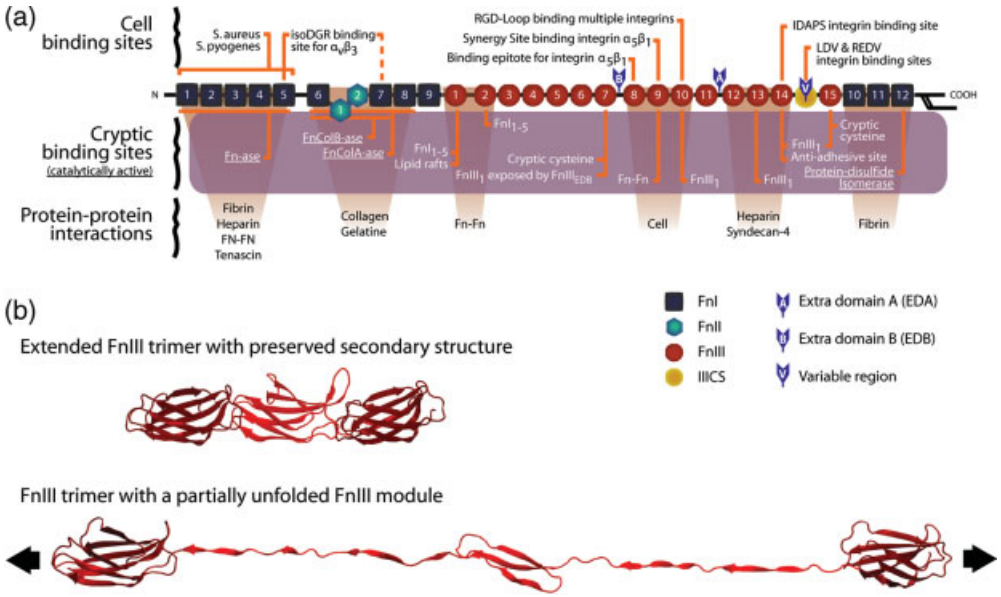


Figure 9.4 Fibronectin's major binding sites and an example of module unfolding under tensile stress [269]. Fibronectins are dimeric molecules composed of over 50 repeats of three different β-sheet modules (FnI, FnII and FnIII). (a) One monomer of as fibronectin found in blood plasma. Fibronectin produced by cells may contain additional alternatively spliced modules, as indicated. Fibronectins contain a large number of molecular recognition and cryptic sites, including the cell-binding site RGD, which is recognized by multiple integrins; the synergy site PHSRN, which is recognized by $\alpha_5\beta_1$ and $\alpha_{IIb}\beta_3$ integrins; the sequence IDAPS at the FnIII13–14 junction in the heparin II binding region of fibronectin, which also supports $\alpha_4\beta_1$-dependent cell adhesion; and the NGR motif in FnI5, which is nonenzymatically converted to isoDGR and can then bind the $\alpha_v\beta_3$ integrin [249]. A similar, highly conserved NGR motif occurs in FnI7, but has not been extensively studied. The cryptic sites include various Fn self-assembly sites, the exposure of which is needed to induce fibronectin fibrillogenesis. Finally, there are two cryptic, nondisulfide-bonded cysteines on each monomer, in modules FnIII7 and FnIII15 which are utilized for site-specific labeling studies by fluorescence resonance energy transfer; (b) Tensile stress applied to Fn fibers causes changes in the quaternary, tertiary and secondary structure of Fn molecules. The figure shows three FnIII modules with intact secondary structure (upper) and with the partial unfolding of one module due to increased tensile stress (lower). (Reproduced with permission from Ref. [269].)

module, each of which has different structural folds, including 12 Fn type I domains, two Fn type II domains, and 15–17 Fn type III domains per Fn monomer. Both, FnI and FnII domains contain two intrachain disulfide bonds, while FnIII domains are not stabilized by disulfides and are hence more susceptible to force-dependent unfolding. Fibronectin displays a number of surface-exposed molecular recognition sites for cells, including integrin binding sites such as the RGD loop, PHSRN synergy site and LDV sequence, as well as binding sites for other ECM components, including collagen, heparin and fibrin. A number of cryptic binding sites and surface-exposed binding sites have been proposed to be exposed or deactivated, respectively, as a result of force-dependent conformational change (as reviewed in Ref. [26]). Interestingly, it is not only fibronectin that contains these modules; in fact, approximately 1% of all mammalian proteins contain FnIII domains that adopt a similar structural fold to the FnIII domains in fibronectin [80].

9.4.1.2 **Integrins**

Integrins – the major cell matrix adhesins – are transmembrane dimers composed of noncovalently bound α and β subunits which associate to form the extracellular, ligand-binding head, followed by two multi-domain 'legs', two single-pass transmembrane helices and two short cytoplasmic tails (Figure 9.5). Although integrins are not constitutively active, their activation is required to form a firm connection with RGD–ligands. Conformational alterations at the ligand binding site of the extracellular integrin head domains propagate all the way to the cytoplasmic integrin

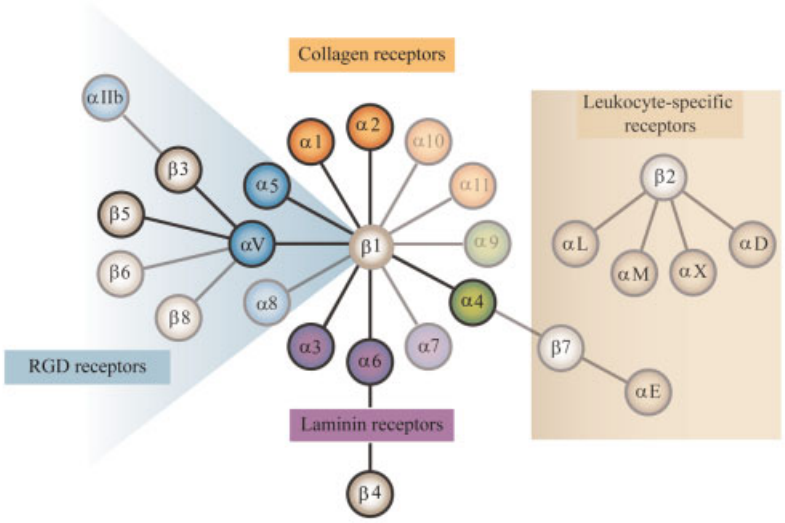


Figure 9.5 The integrin receptor family. Integrins are αβ heterodimers whereby the eight α subunits can assort with 18 β subunits to form 24 distinct integrins. Some integrins recognize the RGD-ligand (blue), while others bind to collagens (orange) or laminins (green), as further discussed in Refs [24, 31, 32]. (Adopted from Refs [31, 32].)

tails, and vice versa, by not-yet understood mechanisms (for reviews, see Refs [31, 32, 35, 106–108]). When a ligand binds to the integrin head, it becomes activated. The activation involves a conformational change that propagates through the extracellular integrin domains, finally forcing the crossed transmembrane helices of the integrin α- and β-subunits to separate, thereby opening up binding sites on their cytoplasmic tails. In contrast, if intracellular events force the crossed integrin tails to separate, then a conformational change will propagate to the extracellular headpiece, thereby priming the integrin head into the high-binding state, even in the absence of an RGD–ligand. This bidirectional conformational coupling between the outside and inside is remarkable, as the integrin molecule is approximately 28 nm long [35, 106–108]. Integrins, however, can also be constitutively activated, for example in the presence of Mn^{2+} ions, by point mutations and via activating monoclonal antibodies [48, 49, 109–111]. Integrin-mediated adhesion often occurs under tensile forces such as fluid flow or myosin-mediated contractions that cells exert to sample the rigidity of their surroundings. In fact, a dynamic mechanism has recently been proposed as to how mechanical forces can accelerate the activation of the RGD–integrin complex [112].

9.4.1.3 **Talin**

Talin is a cytoplasmic protein that can not only activate integrins [113], but also physically links integrins to the contractile cytoskeleton [114, 115], as depicted in Figure 9.6. The talin head has binding sites for integrin β-tails [116], PIP kinase

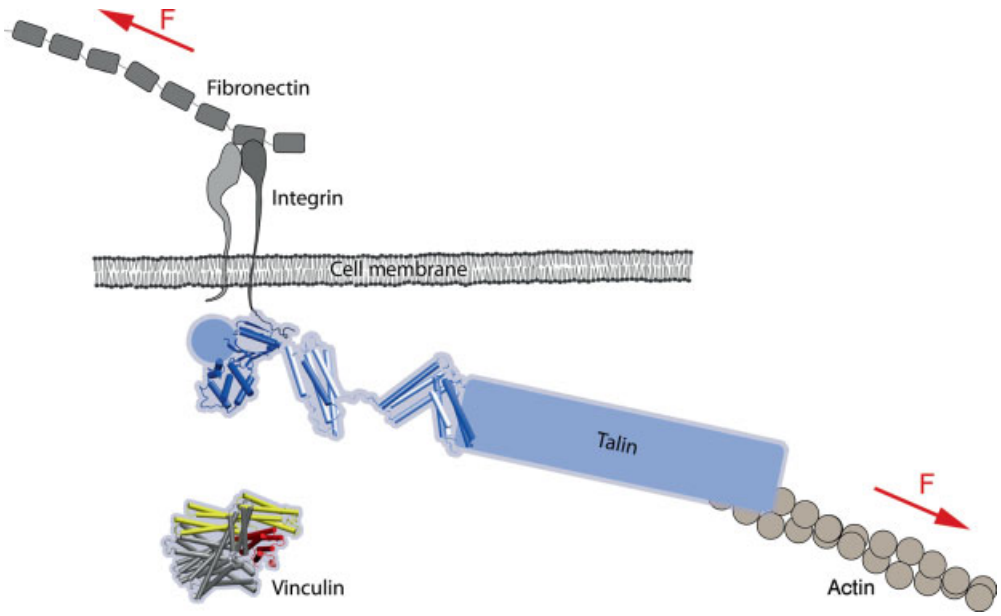


Figure 9.6 Schematic diagram showing how talin anchors integrins to an actin filament. The stretching and partially unfolding of talin (blue) exposes the vinculin binding helices.

γ [117], focal adhesion kinase (FAK) [118], layilin [118] and actin [119] (see also Figure 9.8 below). The 60 nm long talin rod is composed of bundles of amphipathic α-helices [120, 121]. The talin rod contains up to 11 vinculin binding sites (VBSs) [122], including five located within the helices H1–H12, residues 486–889. All of these five binding sites are buried inside helix bundles (native talin shows a considerably lower affinity for vinculin compared to peptide fragments isolated from talin). In addition to the VBSs, the talin rod has binding sites for actin [123] and for integrins [124].

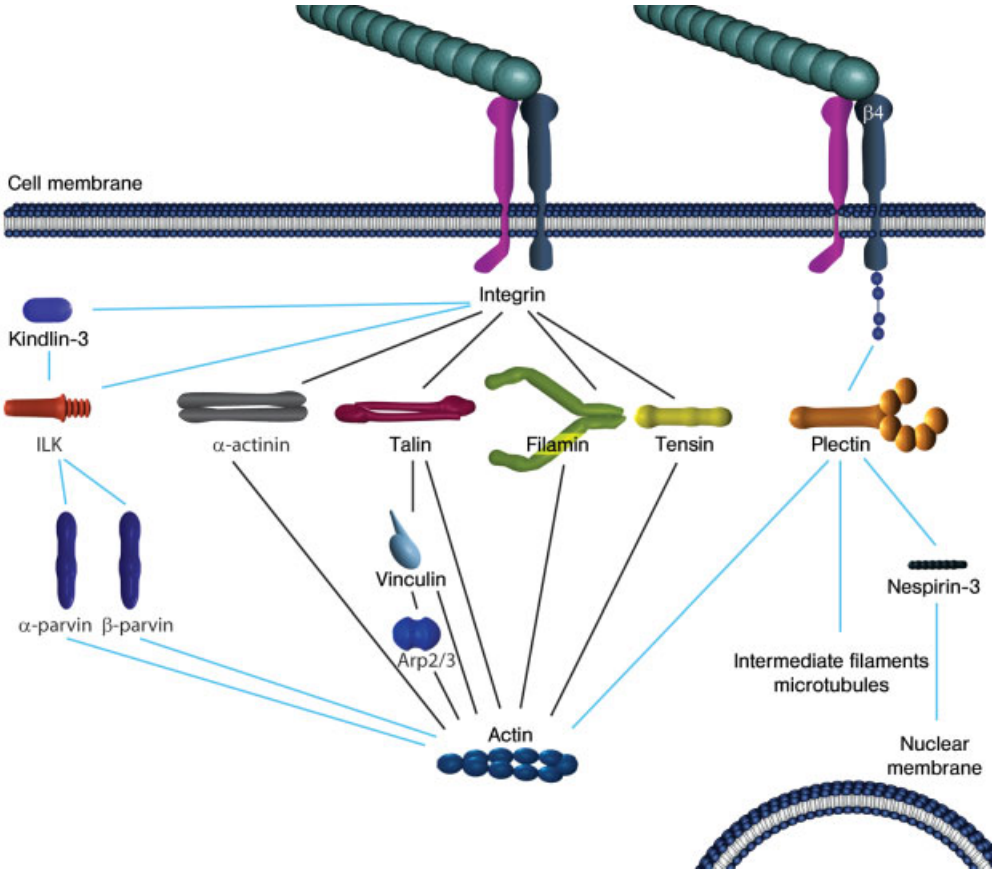


Figure 9.7 Scaffolding proteins that directly link various integrins to actin. Talin, tensin, plectin, filamin and α-actinin were reported to form a single bridge between the various integrins and actin [125–127, 334, 335]. Kindlin-3 was recently added to this list [128], while ILK binds via the formation of a ternary complex with PINCH and parvin [201]. The β_4 integrin has a highly unique intracellular tail which contains four FnIII modules. It can bind via plectin not only to actin, but also intermediate filaments and microtubules, as well as to the nuclear membrane via nesprin-3 [131, 319, 336].

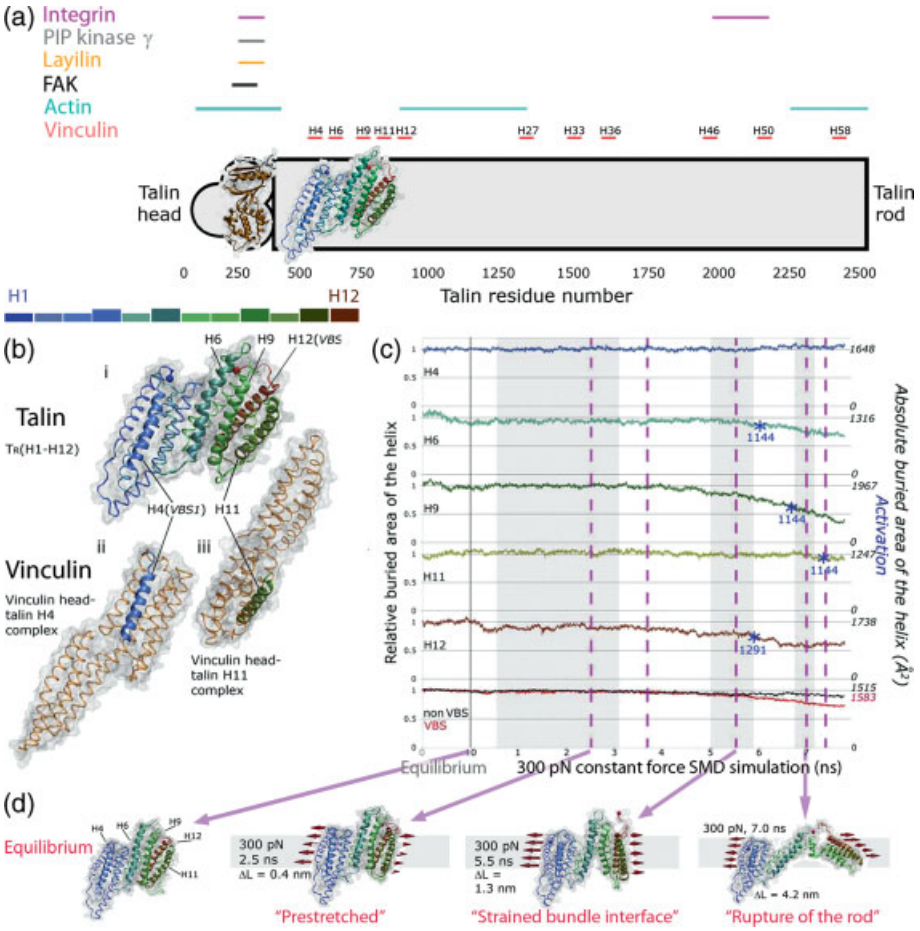


Figure 9.8 Structural mechanism showing how force alters the structure of the N-terminal talin rod comprising helices H1 to H12. (a) Structure of the α-helix bundle of talin, which includes five of the vinculin-binding helices (bold ribbons, namely H4, H6, H9, H11 and H12; (b) The vinculin-binding helices H4 (also referred to as (VBS1)) and H11 in complex with the vinculin head [149] (PDB 1SYQ). The molecular surface is presented in gray; (c) Steered molecular dynamic simulation of the force-induced exposure of the viculin-binding helices to water and the concomitant structural changes shown in (d). Change in the buried surface area of the viculin-binding helices during equilibration and when extended under 300 pN force. The buried areas are shown normalized to the average buried area obtained during equilibration. The respective points of 'activation' – that is, when the buried areas of helices H6, H9, H11 and H12 – in talin equal the experimentally found buried areas of isolated talin helices in complex with the vinculin head, are given as blue asterisks in (c). For H6 and H9, the buried area determined for the H11–vinculin complex is used as a reference because there is no available structure of those helices in complex with vinculin. The buried area of H4 was higher than the buried area of the VH–H4 complex for the whole simulation period. For more detailed information, see Ref. [150].

9.4.1.4 Other Scaffolding Proteins that Provide a Linkage Between Integrins and F-Actin

The physical linkage between integrins and actin can be formed independently by five cytosolic proteins (Figure 9.7). Talin, tensin, plectin, filamin and α-actinin were reported to form a single bridge between the various integrins and actin [125–127], and kindlin-3 was recently added to this list [128]. Talin binds the integrins beta 1, 2, 3 and 5, and weakly to 7 [113]. Tensin and filamin bind to integrins via the same NPxY motif that is recognized by talin [129, 130]. Plectin binds to the laminin-binding integrin β_4 [131], and α-actinin binds to β_1 and β_3 [132, 133]. In addition to integrins, there are ten other membrane-bound adhesion-receptor proteins which bind to either integrins and/or to other adhesion-plaque proteins. Recent data have suggested that certain receptors, for example syndecan [134], can synergize with integrins in adhesion formation [126, 127].

9.4.1.5 Cell Cytoskeleton

Cell–substrate and cell–cell forces are balanced by their interaction partners, except in the case of endothelial cells that experience high fluid flow rates. Thus, the cell cytoskeleton must transmit force across the cell to other sites. This has been observed in the studies of magnetic beads as the propagation of forces to distant substrate sites [135]. There are many ramifications of force propagation in that the cytoskeleton is constantly under tension. Although some of the contractile tension of the cytoskeleton is counterbalanced by the pressure in the cytoplasm, in most cases the intracellular pressure is relatively small (ca. 20 N per m^2) [136]. The majority of the tension is exerted on the actin cytoskeleton, however, we do not yet understand how the spatial distribution of force-bearing adhesions is determined.

9.5 Force-Upregulated Maturation of Early Cell–Matrix Adhesions

9.5.1 Protein Stretching Plays a Central Role

When the integrins latch on to their binding sites in the ECM, the cells apply force to these newly formed adhesion sites, ultimately promoting a rapid bond reinforcement through molecular recruitment. Such recruitment must occur within the lifetime of the initially labile adhesion bond. Key to the reinforcement is *integrin clustering*, followed or paralleled by protein recruitment [125, 137–139]. At least three integrins are needed to form an adhesion [140], and cells show a delayed spreading if the integrins are not sufficiently close [141]. The maturation of adhesion sites seems to involve the stretching and unfolding of proteins, since proteins that are part of such force-bearing linkages might change their structure and, therefore, also their function. One protein which is stretched early in the adhesion process is talin, which links integrins to the cytoskeleton. One of the many proteins that are recruited to newly formed adhesions is vinculin.

9.5.1.1 Vinculin is Recruited to Stretched Talin in a Force-dependent Manner

Upon cell adhesion, talin rapidly accumulates in focal contacts prior to vinculin recruitment [142]. In cases where integrin activation occurs without the application of force, and is thus not part of a force-bearing protein network [139], other adhesion proteins are not recruited. Indeed, the recruitment of vinculin to cell adhesion sites has been shown to be force-dependent [62–64] and to correlate with adhesion strengthening [143] and reduced focal adhesion turnover [144]. Even if not directly shown, this suggested that vinculin recruitment to focal adhesions is upregulated by force [62–64, 145, 146].

Since talin's vinculin-binding helices are buried in its native structure (Figure 9.8b), how might tensile mechanical forces activate them? Some key experimental observations [147–149], together with computational simulations [150] that provided high-resolution structural insights into the force-induced unfolding process of the N-terminal helix bundle of the talin rod which contains five of the vinculin binding sites, suggest the following model of activation.

As the vinculin head consisting of helix bundle is thermodynamically stabilized if it can recruit one additional helix, the vinculin head forms an auto-inhibited complex with its tail domain under equilibrium conditions [151] [PDB: 1TR2]. Instead of binding to itself, the head domain of vinculin can also be stabilized by recruiting other amphipathic helices. For example, isolated vinculin-binding helices of talin can activate vinculin by binding to the vinculin head if added to solution [147, 148, 152, 153]. The release of auto-inhibition is also needed to increase its affinity for actin [154].

Important for the force-activated mechanism is the fact that a larger hydrophobic surface area of talin's vinculin-binding helices can be shielded if they bury themselves in the talin rod rather than in complex with the vinculin head (Figure 9.8c). When mechanically strained, the tightly packed helix bundle of the talin rod breaks into fragments (Figure 9.8d), thereby gradually exposing the buried surface area of the vinculin-binding helices [150]. Once the buried surface area of the vinculin-binding helices in strained talin falls below that shielded if in complex with the vinculin head, the vinculin-binding helices can spontaneously switch their association, breaking off from the strained talin and associating with vinculin; this process is referred to as the *helix swap mechanism* [150]. It was suggested that a vinculin-binding helix would become 'activated' if the buried surface area in mechanically strained talin were to fall below the buried surface area if in complex with vinculin (Figure 9.8c). Vinculin recruitment to talin thus initially increases if talin is incorporated into a force-bearing network formed when a cell adheres to a surface or matrix fibrils [150]. However, as each of the vinculin binding helices is exposed to water at a different time point in the unfolding pathway of the talin (Figure 9.8c), talin can recruit vinculin in a graded response that is upregulated by force. As vinculin can bridge talin and actin, it may reinforce the talin–actin linkage that has been shown previously to be a rather weak bond, breaking at a force of 2 pN [115].

The mechanism described here might not be unique to the talin–vinculin bond, but may be more widespread among other intracellular proteins composed of α-helical bundles. First, when a force breaks away an amphipathic helix from a

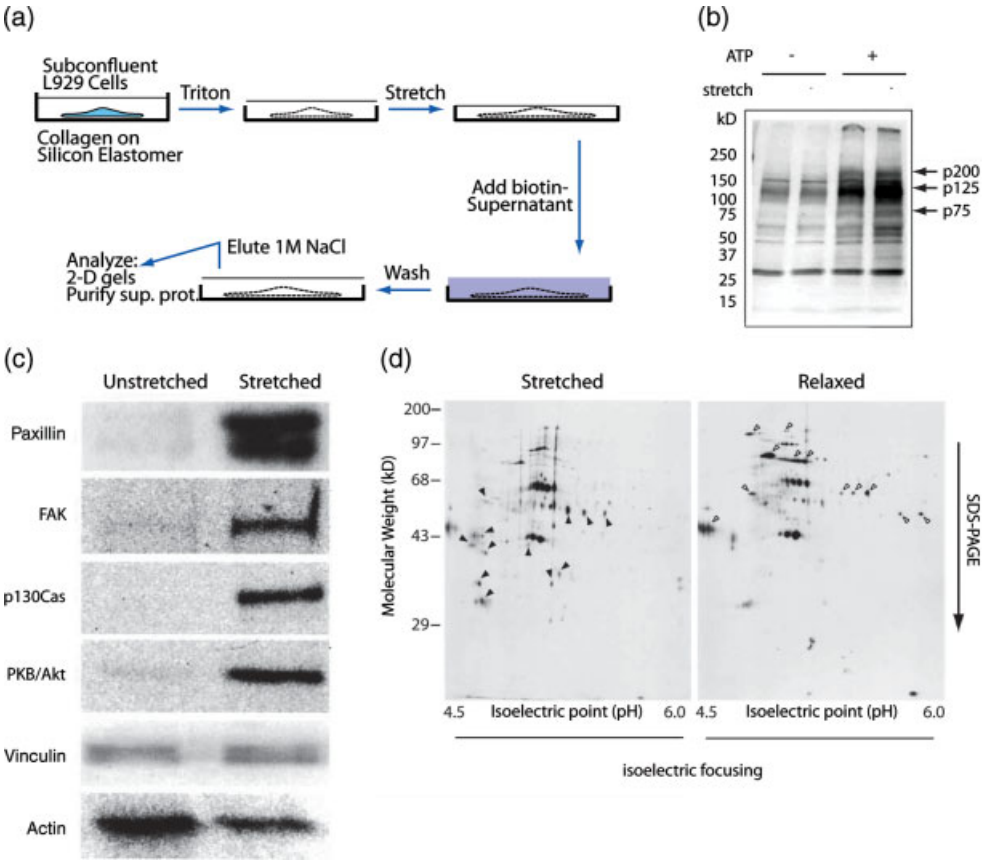
larger bundle, it might be stabilized by insertion into either the hydrophilic pockets of other proteins or even into the lipid bilayer [148]. Alternatively, other proteins that form helix bundles might also bind vinculin in a force-regulated manner. For example, α-actinin also has a vinculin-binding helix that can form a similar structural complex with vinculin [152, 153, 155, 156]. Similarly to talin, the VBS in α-actinin is buried in the native structure. Identifying the repertoire of mechanisms by which forces can upregulate adhesive interactions has led to the recent discovery of catch bonds, where a receptor–ligand interaction is enhanced when tensile mechanical force is applied between a receptor and its ligand (for reviews, see Refs [90, 93]). In contrast, the force-activated helix-swapping mechanism proposed here requires that the force is applied to just one of the binding partners, thereby activating bond formation with a free ligand. Also in contrast to catch bonds, the ligand need not necessarily form part of the force-bearing protein network at the time the switch is initiated. Thus, while force-induced helix swapping primarily upregulates the bond-formation rate, the catch bond mechanism primarily extends the lifetime of an already existing complex under tension.

9.6
Cell Signaling by Force-Upregulated Phosphorylation of Stretched Proteins

9.6.1
Phosphorylation is Central to Regulating Cell Phenotypes

While bond reinforcement is crucial for the cell to develop a stable adhesion site, the subsequent transformation of mechanical stimuli into biochemical signals is needed to alter cell behavior. But, which molecules act as the major mechanochemical signal converters? Although any experimental demonstration of the stretch-dependence of binding to the cytoskeleton had long been missing, there has always been some concern that the opening of stretch-activated ion channels was the cause of mechanosensation. By using matrix-attached, detergent-extracted cell cytoskeletons, it could

Figure 9.9 Stretch of cytoskeletons activates adhesion protein binding and tyrosine phosphorylation. (a) Diagram of protocol for stretch-dependent binding of cytoplasmic proteins to Triton X-100-insoluble cytoskeletons. L-929 cells were cultured on a collagen-coated silicone substrate, and cytoskeletons prepared by treating with 0.25% Triton X-100/ISO buffer for 2 min. Triton X-100-insoluble cytoskeletons were either left unstretched or stretched (or relaxed from prestretch) with ISO after washing three times. The ISO buffer was replaced with the cytoplasmic lysate solution, incubated for 2 min at room temperature, and washed four times with ISO (b) Tyrosine phosphorylation of many proteins increases upon cytoskeleton stretch. Detergent-extracted cell cytoskeletons showed dramatic increases in phosphotyrosine levels in many different proteins upon stretch. Because soluble kinases have been extracted, it is believed that much of the increased phosphorylation is due to stretch of substrate proteins, such as p130Cas. Thus, it appears that there are many additional proteins that could be involved in sensing stretch of cytoskeletally attached components. (Reproduced from Ref. [160]); (c) Focal contact proteins bind preferentially to stretched cytoskeletons. Western blots of focal contact proteins bound



to unstretched and stretched cytoskeletons. L-929 cytoplasmic proteins tagged with a photocleavable biotin (NHS-PC-LC-biotin) were added to Triton X-100-insoluble cytoskeletons of L-929 cells on a stretchable silicone dish [158], and cytoskeletons were stretched or left unstretched (see Figure 9.1). After washing, the bound cytoplasmic proteins were eluted with 1 ml 1 *M* NaCl in HYPO buffer, precipitated with avidin beads (immobilized neutravidin; Pierce Chemical Co.) after sevenfold dilution with HYPO buffer, and released from the bead complex by irradiation with 302 nm UV light (10 min). After photocleavage, proteins were eluted with 120 µl HYPO buffer, and 40 µl of the sample was subjected to 10% SDS–PAGE followed by immunoblotting with antibodies to paxillin, FAK, p130Cas, PKB/Akt (Transduction Laboratories), vinculin (Upstate Biotechnology) or actin (Santa Cruz Biotechnology). Scale bar = 10 µm; (d) 2-D gels of biotinylated proteins that were bound to stretched or relaxed cytoskeletons. The complex of the cytoskeleton with the biotinylated cytoplasmic proteins was solubilized with 1 ml of rehydration buffer (8 *M* urea, 2% CHAPS, 20 m*M* DTT, 0.5% IPG buffer) for isoelectric focusing (the first dimension of 2-D gel electrophoresis). Immobiline dry strip (pH 4–7; Amersham Pharmacia Biotech) was rehydrated with 350 µl of each sample and subjected to isoelectric focusing followed by SDS–PAGE. Biotinylated cytoplasmic proteins in 2-D gels were visualized with affinity blotting using horseradish peroxidase-conjugated streptavidin. Arrowheads mark the spots that were found specifically in Stretched or Relaxed samples. (Reproduced from Ref. [71].)

be shown that different sets of cytoplasmic proteins would bind to cytoskeletons, depending on the extension status (relaxed or stretched) of the cytoskeletons (Figure 9.9), and that binding of the focal adhesion proteins, paxillin, FAK and p130Cas to the cytoskeletons was increased by cytoskeleton stretching [71]. Any increased binding of the cytoplasmic proteins to stretched cytoskeletons would most likely result from the exposure of cryptic binding sites in the cytoskeleton. Whilst it was shown that the binding of another focal adhesion protein – vinculin – remained unchanged in L-929 cells on collagen [71], the force-dependent assembly of vinculin at fibronectin adhesion sites has been reported in other cells [58, 63]. Any specificity derived from the cell type and the substrate to which the cells adhere (including the ECM) appears to account for this discrepancy. In particular, there were no changes in the binding of vinculin to collagen adhesions in intact L-929 cells upon stretching. Subsequent analyses of the range of proteins that were bound to stretched cytoskeletons indicated that both heat-shock proteins and normal focal adhesion proteins would bind to cytoskeletons upon stretching. Both, heat-shock and adhesion stress signals could result from stretch, although the primary signal in the cellular environment is not clear.

9.6.1.1 Stretch-Dependent Binding of Some Cytoplasmic Proteins to Cytoskeletons

Stretch-specific binding studies (Figure 9.9d) indicate that some cytoplasmic proteins will be released from cytoskeletons upon stretching. For example, the binding of actin in a cytoplasmic extract to cytoskeletons was decreased upon cytoskeleton stretching [71]. It is likely that the increase in actin binding to triton (Triton-X-treated) cytoskeletons upon relaxation from a prestressed state is the result of an increase in actin filament assembly (Figure 9.9c), since in intact cells there is an increase in assembly upon relaxation – in that the cell edge becomes very active when a prestretched substrate is relaxed [157]. This indicates that some of the cellular enzyme pathways can be mechanically activated by relaxation, and that some of the binding to the cytoskeletons could result from the activation of enzyme pathways. In addition, cell relaxation-dependent signal activation was observed for Ras [158]. Alternatively, relaxation could enable the refolding of cytoskeletal proteins so that new binding sites would be formed. The binding of cytoplasmic proteins to cytoskeletons could then occur through the relaxation and refolding of cytoskeletal elements. In any case, the cyclic stretching and relaxation of cytoskeletons could play a significant role in controlling the local binding and release of cytoplasmic proteins to the cytoskeletons (Figure 9.9d). The cytoskeleton could thus act as a reservoir such that the mechanical strain would regulate the relative local concentrations of free proteins. This should have an additional impact on biochemical mechanotransduction processes.

Another class of proteins are the *scaffolding proteins* (p130Cas and other candidates that increase in phosphorylation upon cell stretching). These are able to associate with multiple cytoskeletal or signaling complexes through their N- and C-terminal ends, and have multiple phosphorylation sites in a central region (Figure 9.10). The scaffolding proteins appear to have more complex signaling roles, since both of their binding partners and the degree of stretching can change in response to hormone or

other signaling pathways. The phosphorylation of Cas requires both an active Src or Abl family kinase, as well as mechanical unfolding of Cas. The phosphotyrosine sites recruit other signaling molecules such as Crk that initiate signaling cascades. The primary transduction event is complicated, however, because the kinase activation step may occur through a force-activated pathway such as a receptor-like protein tyrosine phosphatase or through a hormone receptor. Consequently, primary and secondary force-sensing distinctions can become blurred and involve extrapolation from only a couple of incomplete examples (these are mentioned here only to stimulate further thought on these important mechanotransduction pathways).

9.6.1.2 **Tyrosine Phosphorylation as a General Mechanism of Force Sensing**

The reversible phosphorylation of intracellular proteins catalyzed by a multitude of protein kinases and phosphatases is central to cell signaling. The recently described phenomenon of substrate priming or stretch-activation of a tyrosine kinase substrate appears to be a major mechanism of force transduction [51]. Anti-phosphotyrosine immunostaining of individual fibroblasts has revealed that tyrosine-phosphorylated proteins are predominantly located at focal adhesions [159–161], where cell-generated forces are concentrated. Furthermore, an adhesion- or stretch-dependent enhancement of tyrosine phosphorylation was observed in many tyrosine phosphorylation sites in T cells [162], fibroblasts (Figure 9.9b) [160, 161] and epithelial cells (Y. Sawada, unpublished observations). In addition, receptor tyrosine kinases have been reported to be tyrosine phosphorylated (i.e. activated) by mechanical stimulation in a ligand-independent manner [163, 164]. These findings indicate that tyrosine phosphorylation plays a general role in adhesion and force-sensing [126, 127]. Due to their hydrophobic character, phosphorylatable tyrosines are typically 'buried' by intramolecular interactions under equilibrium conditions. When such proteins are subjected to stress, however, the buried tyrosines may be exposed, thus enabling them to become phosphorylated. Tyrosine-phosphorylatable proteins also very often carry more than one tyrosine that can be phosphorylated, as does Cas. Multiple repeats of structurally homologous domains are characteristic of many proteins with mechanical functions [26]. Progressive stretching of the molecule can then affect one domain after the other, thus gradually upregulating the response [51]. These observations indicate that substrate priming is a common mechanism for the regulation of tyrosine phosphorylation. As tyrosine phosphorylation appears to be generally involved in force-response (as mentioned above), substrate priming is most likely a universal mechanism of force sensing.

With regards to the force available for stretching molecules in the adhesion complex, the force exerted on one integrin molecule in the adhesion site is estimated to be on the order of 1 pN [58]; this is lower than the forces that allow refolding of many proteins in atomic force microscopy (AFM) experiments [165]. Consistently, the stretching of CasSD (p130Cas substrate domain, the central portion that contains 15 YXXP motifs and is phosphorylated upon stretch) by using AFM gave the appearance of stretching a random coil without any distinct barriers to unfolding (Y. Sawada, J.M. Fernandez and M.P. Sheetz, unpublished observations), suggesting that the Cas substrate domain could be extended by forces below

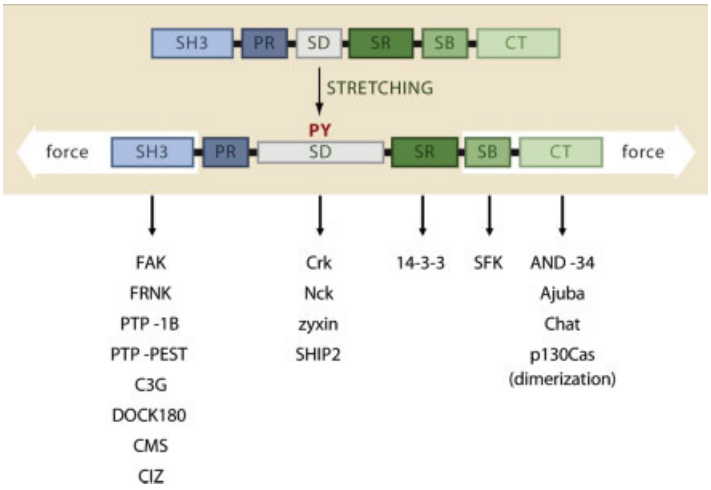


Figure 9.10 Mechanotransduction at focal adhesions through the stretch-dependent phosphorylation of p130cas. The domain structure of the p130CAS molecule is shown (top), before and after stretching. The domains include (from left to right): Src homology 3 (SH3) domain; the proline-rich region (PR); the substrate domain (SD); the serine-rich region (SR), the Src-binding domain (SB); and the C-terminal region. The extension of the substrate domain following stretching and subsequent tyrosine phosphorylation (PY) are indicated. p130Cas and its molecular binding partners are depicted (bottom), including adaptor ('scaffolding') molecules, tyrosine kinases, a serine/threonine kinase, GEFs and tyrosine phosphatases. (Reproduced with permission from Ref. [337].)

the detection limit of AFM (~10 pN). Further, it was observed that CasSD was significantly phosphorylated by Src-kinase with longer incubation times *in vitro*, and that the stretch-dependent enhancement of *in vitro* CasSD-phosphorylation (i.e. fold phosphorylation of stretched/unstretched) is attenuated in longer incubations with Src-kinase (Y. Sawada and M.P. Sheetz, unpublished observations). This indicated that thermal fluctuations of the Cas substrate domain were sufficient to expose tyrosine phosphorylation sites buried in the domain, and raises the possibility that proteins that bind to the native substrate domain like zyxin could stabilize it and inhibit phosphorylation. Thus, the unfolding of p130Cas and its phosphorylation appear to occur at very low applied forces, although the phosphatases that bind to p130Cas could rapidly remove the phosphates.

Finally, tyrosine phosphorylation of the adaptor protein, paxillin, functions as a major switch, regulating the adhesive phenotype of cells [126, 127]. Paxillin, which has binding sites for vinculin and p130cas, can be phosphorylated by tyrosine kinases (including FAK and ABL) and dephosphorylated by the phosphatase Shp2 [126, 127]. Whilst phosphorylated paxillin enhanced lamellipodial protrusions, nonphosphorylated paxillin is essential for fibrillar adhesion formation and for fibronectin fibrillogenesis. The modulation of tyrosine phosphorylation of paxillin thus regulates both the assembly and turnover of adhesion sites. Whilst the method by which force regulates paxillin recruitment and phosphorylation remains unknown,

enzymatic activities appeared to be unnecessary for the reversible, stretch-dependent binding of paxillin as the removal of ATP and inhibition of phosphatases did not block paxillin (Figure 9.9c) binding and release from stretched and relaxed cytoskeletons, respectively [71].

Other intracellular proteins have also been shown to be structurally altered by tensile forces. Some proteins (including nonmuscle myosin IIA, vimentin and spectrin), when mechanically stretched within a living cell [52], expose free cysteines, the functional significance of which is not yet known. Yet, key to all of these mechanisms – from force-induced exposure of otherwise buried residues to helix swapping – is that force induces a structural alteration that allows another protein to bind only if the binding partner is mechanically strained.

9.7
Dynamic Interplay between the Assembly and Disassembly of Adhesion Sites

9.7.1
Molecular Players of the Adhesome

It is essential that cells are able constantly to sense their environment and respond to alterations in mechanical parameters. Thus, while one set of cues and molecular players is required that will promote and drive the assembly of an adhesion complex, another set is responsible for regulating their disassembly. Cell adhesion to the ECM triggers the formation of integrin-mediated contacts and ultimately the reorganization of the actin cytoskeleton. The formation of matrix adhesions is a hierarchical process, consisting of several sequential molecular events during maturation (for reviews, see Refs [24, 125, 166–170]). The very first contacts are formed by matrix-specific integrins, and this leads to the immediate recruitment of talin and of phosphorylated paxilin [171]. This event of building the first integrin connection with actin filaments is followed by FAK activation and the force-sensitive recruitment of vinculin [58, 172–174] and the recruitment of α-actinin (α-actinin crosslinks actin filaments). Vinculin has binding sites for vasodilator-stimulated phosphoprotein (VASP) [175] and FAK [176], both of which coregulate actin assembly (via recruitment of profilin/G-actin complex to talin as well as Arp2/3, respectively). pp125FAK also functions as a key regulator of fibronectin receptor-stimulated cell migration events through the recruitment of both SH2 and SH3 domain-containing signaling proteins to sites of integrin receptor clustering [177, 178]. While the adhesions mature further, zyxin and tensin are recruited [167], and zyxin upregulates actin polymerization [179, 180]. The transition from paxillin-rich focal complexes to definitive, zyxin-containing focal adhesions, takes place only after the leading edge stops advancing or retracts [181]. A decrease in cellular traction forces on focal adhesions then leads to an increased off-rate for zyxin [182].

Tensin plays a central role in fibronectin fibrillogenesis which is upregulated by enhanced cell contractility [183]. As with talin, tensin binds to the NPxY

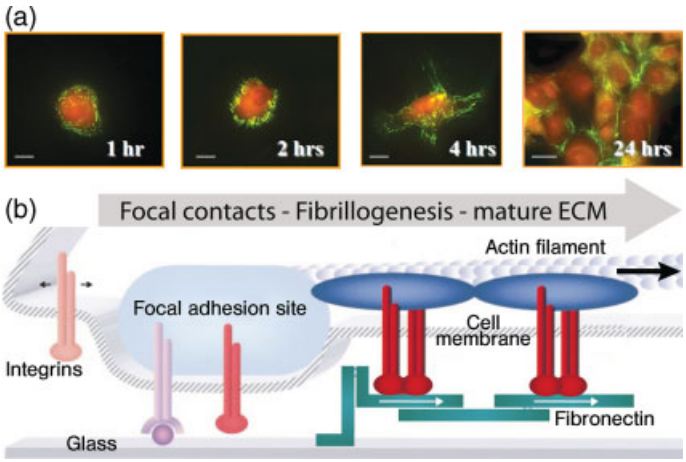


Figure 9.11 Sequential steps in the formation of fibrillar adhesions and fibronectin fibrillogenesis. (a) Time sequence of fibronectin fibril assembly for fibroblasts seeded on fibronectin-coated glass surfaces. (From Ref. [69].) The cells harvested photolabeled plasma fibronectin from solution and incorporated it into newly formed fibers (green). The autofluorescence of the cell bodies is shown in red; (b) Proposed mechanism by which fibronectin-bound integrins translocate out of the focal adhesion sites to form elongated fibrillar adhesions. This process is thought to initiate fibronectin fibrillogenesis. (Adopted from Ref. [167]).

motif of the cytoplasmic β-integrin tails. Fibronectin fibrillogenesis sets in when fibronectin-bound α5β1 integrins coupled via tensin to actin filaments are pulled out of the focal adhesions to form fibrillar adhesions [184] (Figure 9.11). $\alpha_5\beta_1$ integrins translocate along actin fibers, while the other integrins stay back in the adhesion contacts. Fibrillar adhesions translocate centripetally at a mean rate of 18 μm per hour in an actomyosin-dependent manner [56], and evidence is mounting that the stretching of fibronectin induces its polymerization into fibers [69, 70, 185–187]. While phosphorylation of the NPxY tyrosine disrupts talin binding, it has a negligible effect on tensin binding. This suggests that the tyrosine phosphorylation of integrins, which occurs during the maturation of focal adhesions, could act as a switch to promote the formation of fibrillar adhesions [130]. Tyrosine phosphorylation of paxillin regulates both the assembly and turnover of adhesion sites. Moreover, phosphorylated paxillin enhanced lamellipodial protrusions, whereas nonphosphorylated paxillin was essential for fibrillar adhesion formation and for fibronectin fibrillogenesis [126, 127].

The question is, therefore, how are all these processes linked to cell contractility? Rho family GTPases, the major substrates of which are myosin light chain and myosin phosphatase, upregulate myosin activity [15, 188–191], and thus play central roles in integrin signaling [192]. RhoA in particular has been linked with upregulated fibronectin fibrillogenesis [193, 194]. In this context, it is important to note that it is not the intracellular but the extracellular domain of integrin β_1 that controls RhoA activity [194, 195], potentially via its colocalization with syndecan-4 [134].

The formation of such adhesions depends on actomyosin contractility, matrix rigidity [58, 62, 196–198], and the spacing of the integrin ligands on the surface to which the cell adheres [140, 141]. Rigidity sensing is mediated through a mechanism which is further discussed below, where the receptor-like tyrosine phosphatase alpha (RPTPalpha) colocalizes with α_v-integrins at the leading edge of the cell and regulates the activation of Src family kinases [173, 174]. Src family kinases, particularly Fyn, phosphorylate p130Cas in a force-dependent manner [197]; thus, actomyosin contractility enables both fibronectin fibrillogenesis and rigidity sensing. Fibronectin fibrillogenesis, however, occurs via $\alpha_5\beta_1$ integrins, while rigidity sensing is mediated by α_v-integrins.

Finally, microfilament and microtubule networks are significantly reorganized by cyclic stretching, and the cytoskeletal reorganization plays an important role in stretch-induced gene transfer and expression [199]. Integrin-linked kinase (ILK) activity thereby plays an important role in Rac- and Cdc42-mediated actin cytoskeleton reorganization and gene transcription [200, 201]. ILK is a component in focal adhesions that interacts with the cytoplasmic domains of integrins, recruits adaptor proteins that link integrins to the actin cytoskeleton, and phosphorylates the serine/threonine kinases PKB/Akt and GSK-3beta [202]. Finally, the disassembly of adhesion sites is critical, especially at the rear of the cell to enable its forward locomotion [166, 203–205].

The fine-tuning of cell adhesion and detachment, however, requires far more proteins than the few discussed so far. The 'integrin adhesome' consists of a complex network of 156 so-far identified components that are linked in modular complexes and are modified by 690 interactions that have been identified to date [126, 127]. Three major protein families comprise the physical framework of the adhesome, the membrane-anchored adhesion receptors, adaptor or scaffolding proteins, and actin regulators (Figure 9.12). The remaining families consist of mostly enzymes that have roles in regulating the assembly and turnover of the adhesion sites, as well as signaling from the adhesion site into the cell [126, 127]. There are two proteolytic systems associated with the adhesome (ubiquitin E3 ligase protein (Cbl) and calpain), each acting on multiple substrates. Cbl is regulated by tyrosine kinases, and calpain by serine/threonine kinases [126, 127]. Calpain also degrades two tyrosine phosphatases – one of these, shp1, is a regulator of Cbl. Cbl is activated by tyrosine phosphorylation, whilst through its E3-ligase activity it downregulates tyrosine kinase signaling and promotes the proteasomal degradation of integrins [126, 127]. This regulation of binding interactions is very important. Anchoring of adhesion components through multiple links might suggest a robust scaffold structure. In contrast, it must be a highly dynamic, regulated structure that needs to respond to external stimuli and to support morphogenesis and cell migration [126, 127].

Several functional 'subnets' are involved in switching on or off many of the molecular interactions within the network, consequently affecting cell adhesion, migration and cytoskeletal organization. An examination of the adhesome network motifs reveals a relatively small number of key motifs, dominated by three-component complexes in which a scaffolding molecule recruits both a signaling molecule and its downstream target [126, 127]. The authors estimate that more than half of

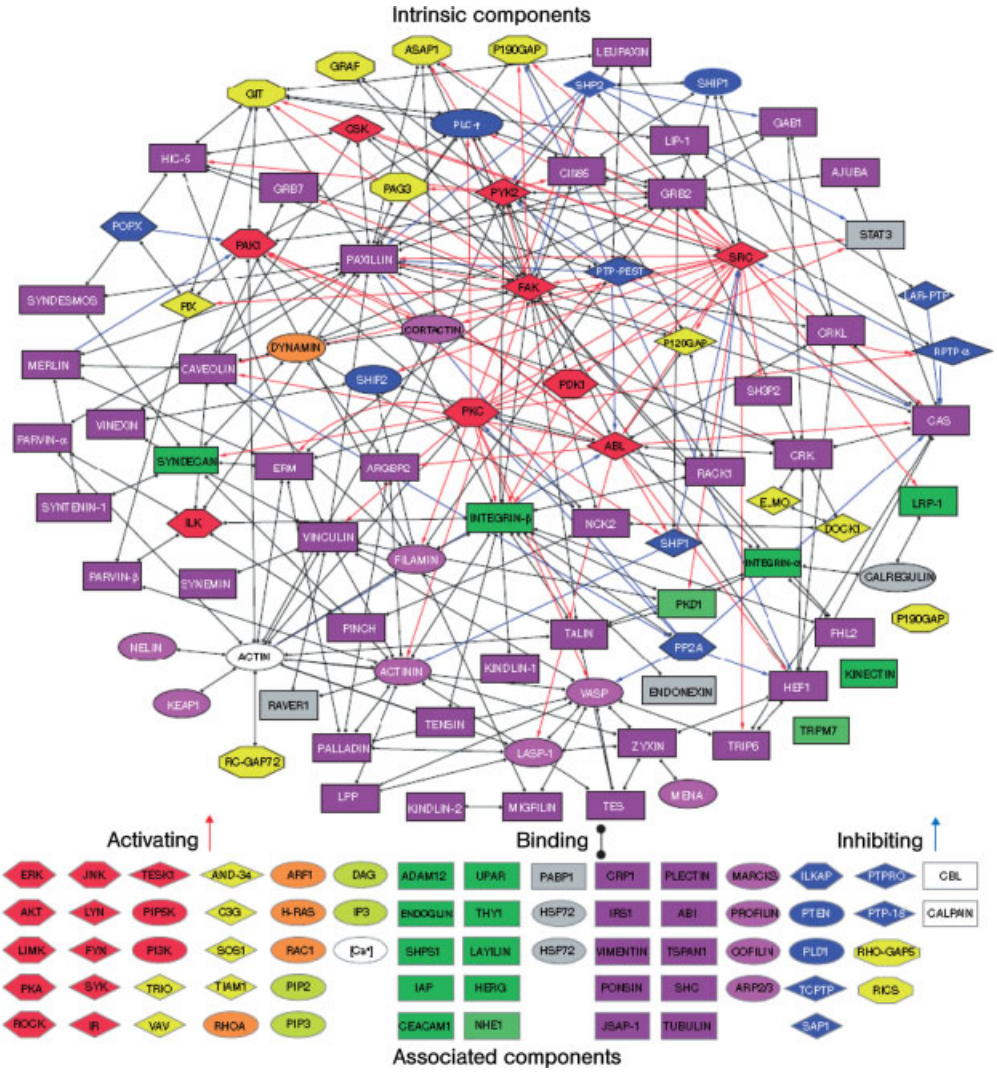


Figure 9.12 Interactions between all intrinsic components of the adhesome and a grouped list of the associated components [126, 127]. Black lines with full circles at their ends denote nondirectional binding interactions; blue arrows represent directional inhibition (e.g. dephosphorylation, G-protein inactivation, proteolysis); red arrows represent directional activation (e.g. phosphorylation, G-protein activation) interactions. The nodes are shape- and color-coded according to the function of the proteins. Intrinsic components are surrounded by a black frame and associated components by a gray frame. (Reproduced with permission from Refs [126, 127].)

the links interconnecting different adhesome components can be switched on or off by signaling elements. There are several types of regulated interaction switches, including conformational switches, GTPase switches, lipid switches, proteolytic switches and PY–SH2 switches [126, 127].

9.8
Forces that Cells Apply to Mature Cell–Matrix and Cell–Cell Junctions

9.8.1
Insights Obtained from Micro- and Nanofabricated Tools

Major experimental tools were developed to probe, with high spatial and temporal resolution, the forces that cells apply to two-dimensional (2-D) surfaces [58, 59, 206–212]. For example, the deflection of microfabricated pillars makes it possible to observe the complete spatial pattern of actin–myosin-driven traction forces applied to the substrate, as shown in Figure 9.13 [59]. This and other tools enable research groups to determine how the linkage between the ECM and the cytoskeleton is stabilized by mechanical force [62, 115, 173, 174, 213–221].

If the force that cells apply to a newly formed junction is too high, the junction will break instantaneously. Thus, the generation and sensing of force is critical for the correct formation of the organism and functions of its tissues. First, however, we should define what is meant by the basic physical parameters of stress and strain in the cellular context, since the anchoring of a cell to an environment is critical for its survival. If firmly anchored to the ECM, the integrins couple the matrix to the contractile machinery of the cell; in this way the major cellular forces are generated by myosin II filaments pulling on actin in early spreading cells [222]. While the level of force orthogonal to the membrane plane that can be supported by the fluid lipid

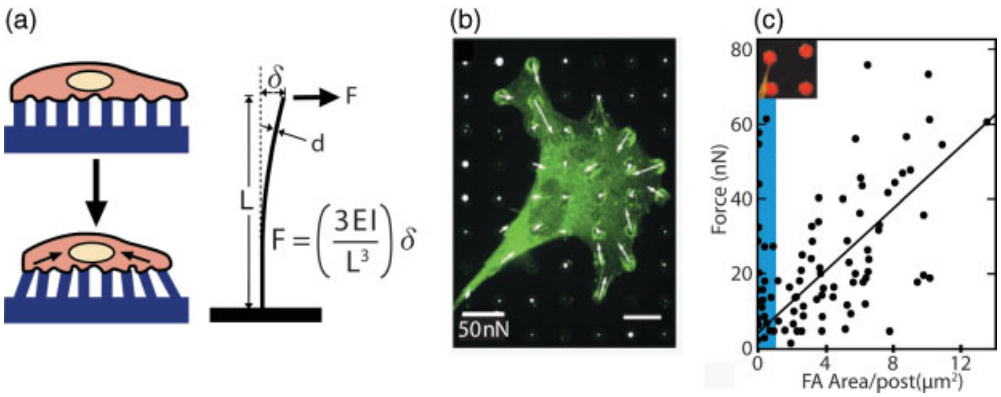


Figure 9.13 Measurement of contractile forces that cells apply to substrates [59]. (a) Cell culture on arrays of polydimethylsiloxane (PDMS) posts covered with fibronectin (as produced by microcontact printing); (b) Confocal images of immunofluorescence staining of a smooth muscle cell on posts. Cells deflected posts maximally during the 1–2 h period after plating, and were fully spread after 2 h. Scale bars indicate 10 μm. The positions of the tips of the posts were used to calculate the force exerted by cells (white arrows). The lengths of arrows indicate the magnitude of the calculated force (top right arrow indicates 50 nN); (c) Plot of the force generated on each post as a function of total area of focal adhesion staining per post. Each point represents the force and area of vinculin staining associated with each post; focal adhesions from five cells were analyzed. The shaded region (blue) indicates the adhesions smaller than 1 μm^2 (inset). (Reproduced with permission from Ref. [59].)

membrane alone is quite small (typically 10–20 pN for a circular area of 100 nm diameter; i.e. a membrane tether) for mammalian cells [223], the plasma membrane is supported by internal and external filamentous proteins that have links to the cytoskeleton. As discussed above, mechanical reinforcement of the very first contacts that a cell forms with the ECM is thus essential. When the adhesions have matured, they can typically withstand forces of a few nN per square micrometer (see Figure 9.13). Tensile forces are then transmitted to the cytoskeleton network in the cell, which can disseminate the force to many or a few sites, even on the opposite side of the cell [224]. Major cellular forces are generated by myosin II filaments pulling on actin throughout the cytoskeleton in early spreading cells, and in the periphery of epithelial cells, particularly around damaged areas in the tissue [160]. In mature epithelial cells, networks of intermediate filaments are generated that bear much of the force when the tissues are stretched.

Far more is currently known about the mechanical characteristics of cell–ECM contacts than about the mechanical properties of cell–cell contacts [21, 126, 127, 225]. The formation of tight cell–cell junctions, however, is critical for many developmental processes such as formation of the gut, kidney, breast and many other epithelial tissues, and is mediated by homologous cadherin–cadherin bonds [22]. These bonds must be dynamic because there are movements of cells relative to each other in epithelial monolayers [23, 226]. Further, when a cell dies, its neighbors rapidly close the gap by first forming an actomyosin collar around the hole; the collar then contracts to cover the hole [160]. Similarly, in the process of convergent extension during embryogenesis, cells converge along one axis while being able to move relative to one another. This is the major morphological movement responsible for organizing the spinal cord axis [227]. Many of the important morphological changes in development involve the contraction of epithelial cells, from the early formation of the gut to the later formation of kidney tubules. In all cases, there is evidence that although the individual cells can move independently, the whole tissue still acts as a unit to undergo a coordinated morphological change. The molecular basis of the mechanical coordination in epithelial or endothelial cell monolayers is not known, but the precision of movement implies that a rapid feedback mechanism is present. It is thus very interesting to note that when force-sensing micropillars were coated with cadherins rather than with fibronectin, the mechanical stresses transduced through cadherin-adhesions were of the same order of magnitude as those previously characterized for focal adhesions on fibronectin [225]. So, the question is, what is the relative importance of cell–cell and cell–matrix contacts in different tissues on transducing mechanical stimuli into altering the downstream behavior? In both tissue types, cell–cell interactions predominate. Endothelial cells require cell–cell contacts, while vascular endothelial cells utilize cadherin engagement to transduce stretch into proliferative signals [15]. Hence, stretch stimulated Rac1 activity in endothelial cells, whereas RhoA was activated by stretch in smooth muscle cells.

Finally, tissue remodeling often reflects alterations in local mechanical conditions that result in an integrated response among the different cell types that share – and thus cooperatively manage – the surrounding ECM and its remodeling. The question therefore is whether mechanical stresses can be communicated between different

cell types to synergize a matrix remodeling response. When normal stresses were imposed on bronchial epithelial cells in the presence of unstimulated lung fibroblasts, it could be shown that mechanical stresses can be communicated from stressed to unstressed cells to elicit a remodeling response. Thus, the integrated response of two cocultured cell types to mechanical stress mimics the key features of airways remodeling as seen in asthma: namely, an increase in production of fibronectin, collagen types III and V and matrix metalloproteinase type 9 (MMP-9) [228].

9.9
Sensing Matrix Rigidity

9.9.1
Reciprocity of the Physical Aspects of the Extracellular Matrix and Intracellular Events

As long as the cell–matrix and cell–cell linkages hold tight, intracellular motile activity will interrogate the matrix, and the subsequent cellular activity will depend on the physical properties of the extracellular fibrillar network, and vice versa. The interrogation involves a mechanical testing of the rigidity as well as the geometry of the environment through the normal cell motility processes (Box 9.2). When the rigidity is determined, the cell will respond appropriately. For example, the extracellular network structure is remodeled if it is of the same or a softer compliance than the intracellular network [18]. Rigidity is an important part of the environment of a cell, and different tissues have different rigidities as well as different levels of activity (this point will be discussed later). There is considerable interest in learning how this is achieved at a molecular level.

A number of reports have indicated that matrix rigidity is a critical factor regulating fundamental cell processes, including differentiation and growth. Discher's group recently showed that the differentiation of mesenchymal stem cells is heavily dependent on the rigidity of the matrix to which cells adhere [19]. The group reported that mesenchymal stem cells preferentially become neurogenic on soft substrates, while they preferentially commit to myogenic and osteogenic differentiation on intermediate and rigid substrates, respectively. The cellular response to rigidity has been seen at the time scale of seconds for submicron latex beads [213] and during cell spreading [229]. Fibroblast migration toward rigid substrates indicates further that the process has important ramifications for *in vivo* motile activities [230]. From the signal transduction point of view, these observations indicate that the sensing of different rigidities can be very rapid and may have profound effects on cell function at a variety of levels. Rigidity is a rather complicated parameter for the cell to sense because measurements of both force and displacement must be combined. In order for a cell to sense matrix rigidity, the cell must actively pull on the matrix; thus, the cell must actively test the rigidity of its environment. As a corollary, the cell in an inactive state will not be able to develop a rigidity response. Although these statements apply for single cells *in vitro*, the situation is more complicated in a tissue environment where neighboring cells and

Box 9.2
Cell Forces

- *Tensile Forces:* As the cytoskeleton of the cell is generally contractile, the transmembrane forces on cells are primarily tensile. Large forces in the movement of organisms or tissues are typically generated by linkages from the cytoskeleton to the ECM or neighboring cells through integrins or cadherins, respectively. In contrast, the lipid bilayer of the plasma membrane is fluid and can be distorted with relative ease. Forces are exerted typically on noncovalent protein–protein bonds (one exception is the transglutaminase linkage of lysines to glutamines on collagen or fibrin). Typical protein–protein bonds can sustain about 1 pN per bond (e.g. 1 pN per integrin–fibronectin bond in living tissue, where the bond holds for a matter of seconds). Although those forces seem low, they need to be maintained for long periods, and even the very high-affinity avidin–biotin bond only has a lifetime of about 5 s under a force of 5 pN. Typical forces that fibroblasts can exert on fibronectin-coated surfaces are on the order of 1–5 nNper μm cell edge.
- *Compressive Forces:* Compressive forces are primarily resisted by the hydrostatic pressure of the cytoplasm of the cell (particularly true in plant systems). The other type of compressive force is that generated as cells extend lamellipodia, filopodia or pseudopodia, and those processes push on neighboring cells or the environment.

Tissue Rigidity

- *Resting Rigidity:* A tissue at rest has a rigidity that is defined by the Young's modulus, E, which can be determined from the slope of the tensile stress versus the tensile strain curve:

$$E = \text{tensile stress}/\text{tensile strain} = \delta\sigma/\delta\varepsilon = \delta(F/A_0)/\delta(L/L_0)$$

where: E is the Young's modulus (modulus of elasticity) measured in Pascals; F is the force applied to the object; A_0 is the original cross-sectional area through which the force is applied; δL is the amount by which the length of the object changes; and L_0 is the original length of the object.

- *Rigidity Sensing:* To measure rigidity, cells develop force on matrix over time. This means that the mechanism of sensing rigidity must compare force and displacement during a given time period. If a cell pulls on an object, the total displacement is the sum of the matrix and of the intracellular displacements. Since the cell directly probes the relative molecular displacements in the adhesion site, it should not matter whether the force is applied from the outside or the inside. This statement, however, is true for soft surfaces only if the external force is applied rapidly before the cell has displaced the substrate. The time window is important since the cell does not pull with a constant force but gradually increases the force after making a first contact. The physiologically relevant time window in this context typically lasts for just 1–2 seconds. The real

question is whether the cell ramps up the force until a max force applied to the adhesion site is reached, or alternatively a maximal stress, or until a certain relative displacement of the intracellular rigidity sensors within adhesion sites is reached which might be in the order of 130 nm. If the latter is the case, the cells pull with larger forces on rigid objects attempting to reach the critical displacement of the intracellular rigidity sensors.

tissue modulation can produce stresses and strains on cellular components in a sustained manner. The complexity of the rigidity measurement can potentially be an important part of tissue assembly.

In terms of the mechanism by which cells can sense rigidity, they must measure the force needed for a given displacement (rigidity) or the displacement for a given force (compliance). The rigidity of tissues can be based upon the rigidity of either the matrix or of the cells themselves. Most of the focus of *in vitro* studies has been on the effect of matrix rigidity rather than cell rigidity, because that is easier to manipulate. By using elastic pillars of different rigidity as substrates, Saez *et al.* showed that the forces exerted by the cells increased linearly with rigidity of the pillars. Thus, the cells deform the pillars by the same amount (on the order of 130 nm) over an almost 100-fold change in rigidity (Figure 9.14) [231]. This observation suggests that cells can sense the displacement of the cell–substrate anchor sites and continue to develop higher forces on rigid substrates until the proper displacement is reached. Because displacement must be measured from the cell edge or the initial site of pulling, the sensor molecules must be anchored both at the edge (initial site of pulling) and at the integrin that is being pulled. In terms of a possible physical mechanism (see Figure 9.14), the movement of a component relative to a stationary component could produce a signal to stop further recruitment or further movement. As many molecules can easily span 130 nm, the movement of the actin (anchored to the integrin) past myosin or some other relative molecular displacement could be linked to a signaling process. If the displacements were less than 130 nm, then a signal for myosin filament contraction and/or assembly could be generated. If the displacements were greater than 130 nm, then further myosin activation would be blocked because the sensor would be physically separated from the enzyme that modifies it or the sensor could be fully stretched. Because the distance is molecular and yet micrometer-sized contacts were produced on rigid pillars, the measurement of rigidity must continually be made over time by activating new regions to contract and allowing old regions to relax. This can be compatible with a previously described model of rigidity sensing through p130Cas by assuming that force must increase until p130Cas is displaced from the kinase in (Figure 9.14) [21, 197]. However, other models, which are more closely linked to the movement of actin relative to myosin, may be used for longer-term rigidity sensing. Multiple mechanisms probably exist to enable different types of cells to properly sense rigidity in different environments. In many cases, there is a strong need for a local feedback between the level of actin and myosin recruited and the rigidity in the epithelial cells. Hence, additional experiments are required to identify possible molecules that could be the rulers in such a system. A consequence of the displacement sensing is that higher

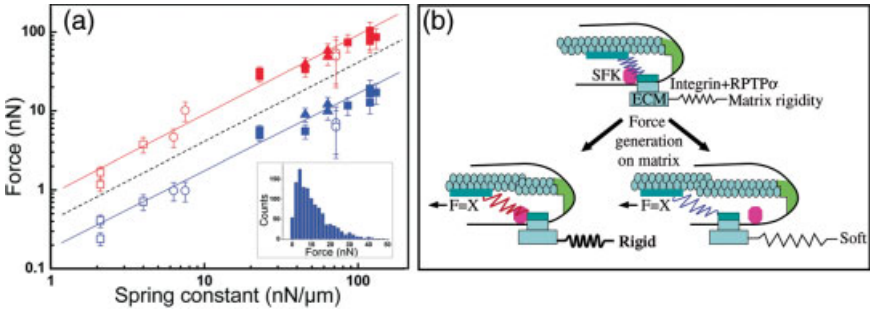


Figure 9.14 Rigidity sensing; evidence for a displacement sensor and one plausible molecular mechanism. (a) Linear relationship between rigidity and average force generated indicates that cells sense displacement [231]. Log-log plot of the force as a function of substrate rigidity. F (blue) and F_{max} (red) within an island of cells are represented for different surface densities: (ratio of the post surface over the total surface) 10% (circles); 22% (squares); 40% (triangles). Open and solid symbols correspond to pillars of 1 and 2 μm in diameter, respectively. The slope of the dashed line is 1. Inset: Typical histogram of force distribution (spring constant 64 nN μm^{-1}); (b) Relative displacement model for rigidity sensing by substrate priming. In this scheme, the displacement of the cytoskeleton–integrin–matrix complex from the active Src-family kinase (Fyn) is the signal to stop further phosphorylation. An initial force signal activates RPTPα at the leading edge that then recruits Fyn to a stationary lipid domain through its palmitoylation. Contraction of the cytoskeleton stretches Fyn substrates such as p130Cas, priming it for phosphorylation by Fyn. In rigid substrates, additional force-generating links are recruited by continued phosphorylation to develop the higher forces needed to displace the p130Cas from the kinase. (Adapted from Ref. [197], where the force needed to stretch p130Cas was emphasized; however, the results of Saez *et al.* [232] and other studies indicated that displacement was measured.)

rigidities will cause higher forces that will in turn enhance the intensity at upstream signals in mechanotransduction (e.g. tyrosine phosphorylation) [51] and may result in a greater activation of downstream signals on rigid substrates.

Another, possibly important, factor which affects the rigidity response is the distance between the cytoskeleton and cell–substrate anchor sites. As elasticity is defined by the magnitude of deformation (e.g. change of dimension) per unit force, cell–substrate anchor sites will be displaced by a larger distance on soft substrates for the same force (see Figure 9.14). If that is the case, the distance between cytoskeleton and cell–substrate anchor sites (the length of the matrix plus the integrin and the actin-binding molecules) would be larger on rigid substrates than on soft substrates [233, 234]. However, larger displacement of cell–substrate anchor sites has not yet been demonstrated, and the uniform cellular deformation of elastic substrates of vastly different rigidities [231] implies that the displacement of anchor sites is not necessarily larger on soft substrate.

9.9.1.1 Time Dependence and Rigidity Responses

When cells pull on the substrate to sense rigidity, they use the rearward movement of actin to generate the force. Actin moves rearward at a rate of 30–120 nm s^{-1}, which means that displacements of 130 nm will take more than 1 s. In the models of rigidity

sensing that have been discussed, a rapid rise in force that is sustained could elicit a rigid substrate response. *In vivo*, if the cell experiences tensile forces from the neighboring cells or the matrix during the period where it is pulling on the matrix, then the matrix can appear rigid because the force will rise rapidly. In experiments where the force was increased rapidly on fibronectin beads with a soft laser tweezers, the bead appeared to be in a rigid laser tweezers, as was predicted. Similarly, *in vivo* many tissues experience external forces on roughly a second time-scale that could develop a rigidity response. Thus, rigidity-dependent growth could be stimulated *in vivo* by tissue contractions.

The time dependence of the assembly of components in the integrin–cytoskeleton complex might affect the rigidity response, since different components bind and detach during the life cycle of an adhesion site [184, 235]. Primary connections between integrins and the cytoskeleton, and their reinforcement, depend on talin (which is probably one of the first proteins in adhesion sites) [64, 115], on α-actinin (which crosslinks actin filaments) and on zyxin (which enters the adhesion site during its maturation) [181, 236]. The transition from mature focal adhesion to fibrillar adhesion is characterized by the segregation of tensin and specific integrins [56]. Because the ECM–integrin–cytoskeleton connection is a viscoelastic material (i.e. it is not purely elastic) [237], the time required to reach the threshold force for rigidity responses probably differs depending on the stiffness of the ECM. Accordingly, a soft optical trap could mimic the effects of a rigid trap on the stabilization of the integrin–cytoskeleton linkages if externally applied forces rise rapidly [233]. In lamellipodia, the cytoskeletal-dependent radial transport of a contractile signal directs the timing of contraction and, probably, adhesion site initiation to stabilize protrusive events [229] (Figure 9.15). Consequently, the formation of cell contacts with the ECM is not a continuous process but rather involves cycles of contraction and relaxation.

9.9.1.2 **Position and Spacing Dependence of the Rigidity Responses**

The position dependence of rigidity responses is exemplified by the fact that structural and signaling proteins that are necessary for rigidity responses are placed at strategic locations – for example, at the cell edge during protrusive events and at early adhesion sites. Many proteins involved in rigidity responses and/or phosphotyrosine signaling, including talin [64, 115], integrins ($\alpha_v\beta_3$) [115, 238], paxillin [173, 174], α-actinin [173, 174, 229], RPTPα [173, 174], Rap1 [239] and p130Cas [240], are localized at the leading edge of the cell, ready to respond to any contraction generated by the cell or by the ECM. There is a position-dependent binding-and-release cycle of fibronectin–integrin–cytoskeleton interactions, with preferential binding occurring at the active edges of motile fibroblasts and release at 0.5–3 μm back from the edge [241]. Interestingly, reinforcement occurs preferentially at the edge in rigid tweezers [233], whereas weak connections that break easily are favored by nonrigid tweezers and at sites >1 μm back from the leading edge [115, 233]. At the molecular level, the reinforcement of integrin–cytoskeleton interactions are limited to linkages that have experienced force, and not those nearby (<1 μm) [213].

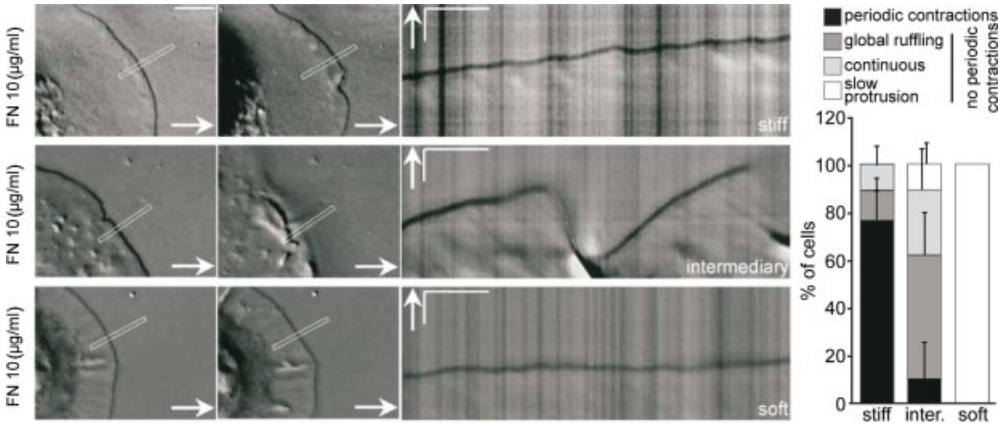


Figure 9.15 Contraction of spreading cells results in periodic contractions and further spreading on rigid surfaces, but no further spreading on soft surfaces. On intermediary stiff surfaces coated with the same concentration of covalently attached fibronectin, cells attempt to spread but lose adhesion. (Reproduced with permission from Ref. [229].)

Finally, many tissues experience periodic stretch *in vivo* during normal activity. When the tissue is inactive, it often experiences atrophy; this is obviously true for muscle, bone and connective tissue. The greatest problem for space travelers is that astronauts typically lose 1–2% of their bone mass for each month in space, even though they may exercise regularly [242]. Similarly, the skin on the feet or hands thickens with use or labor, and thins with disuse. Thus, force from activity or rigidity appears to be a global regulator of tissue function, and an understanding of the mechanisms whereby force is transduced into biochemical signals is an important area of future research.

At the subcellular level, there are many forces that must be regulated to produce normal cell morphology and the proper distribution of organelles. Although the protein composition of many genomes and even individual cell types is known, relatively little knowledge exists of how those proteins are assembled into functional complexes. Individual proteins, typically 5–20 nm in diameter, are assembled into larger functional complexes that can be considered as subcellular machines, controlling and regulating complex cellular functions, from reading and translating the genetic blueprint to the synthesis and transport of proteins, from cell migration to cell division, from cell differentiation to cell death. Those subcellular complexes range in size from ribosomes to the lamellipodial machines that drive ECM assembly and remodeling [243], including collagen fiber rearrangements [244]. It is important to be able to dissect the steps into these subcellular processes to enable greater understanding of the sequence of coherent molecular events [245] (Figure 9.16).

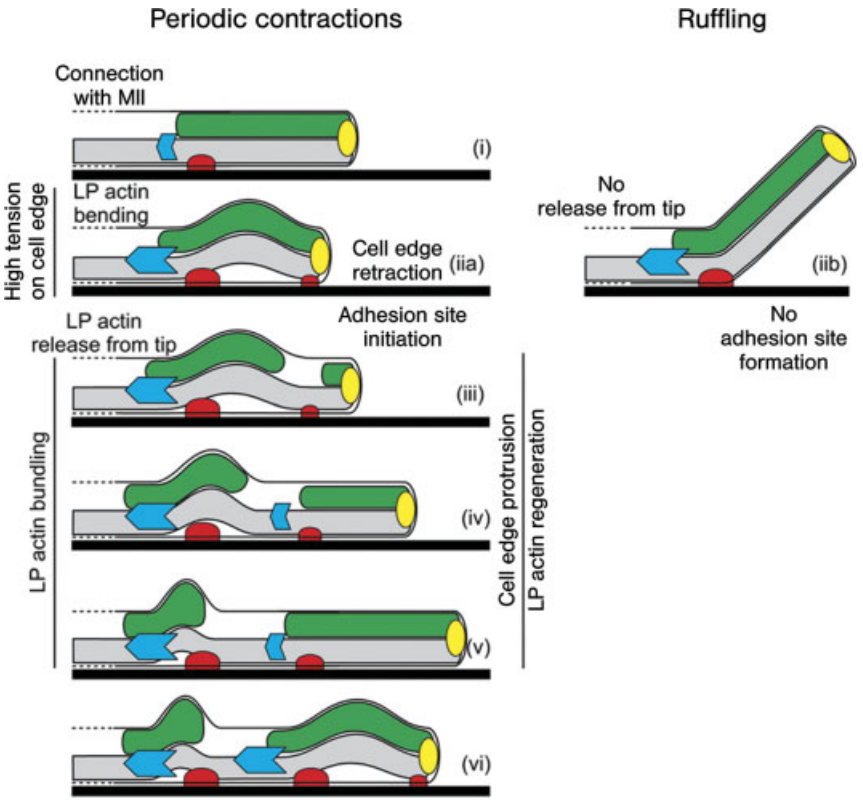


Figure 9.16 Schematic representation of lamellipodial (LP)–actin periodic regeneration. The LP actin (green) is above the LM (gray). Polymerization at the front of the LP actin network causes the back of the network to grow towards the back of the LP contractile module until it reaches an adhesion site (i) where a MII cluster (blue) forms. MII pulls the LP actin, generating high tension on the cell front, causing LP bending, edge retraction and initiation of new adhesion sites (red) on the extracellular matrix (ECM; black rectangle) (iia). The LP actin continues to be pulled until it is released from the tip (iii) and edge protrusion restarts. A new LP actin network immediately resumes growth, which suggests that the actin polymerization machinery (yellow) is still present at the cell tip (iv). The released LP actin, still pulled by MII, further condenses into a bundle at the back of the previous adhesion site (v), while the newly growing LP actin reaches the next adhesion site and the cycle begins anew (vi). LP ruffling (iib) occurred in the case when the total bond energy connecting LP actin to the edge was greater than the bond energy of nascent adhesion sites to the ECM. (Reproduced with permission from Ref. [245]).

9.10
Cellular Response to Initial Matrix Conditions

9.10.1
Assembly, Stretching and Remodeling of the Extracellular Matrix

Once the first contacts have been made by the cell with its surrounding, it will often soon begin to assemble its own matrix. Initially, cells build a provisional matrix which

is rich in fibronectin, which plays a particularly important role in early embryogenesis and in wound healing [104]. In order that the assembly process is started shortly after attachment – and even before the genetic machinery upregulates the expression of matrix proteins – the cells harvest fibronectin from body fluids or the cell medium. Only at later stages do the cells start to produce and secrete their own fibronectin, which has some structural and functional differences compared to plasma fibronectin [246]. The complex fibrillogenesis process begins when the cells apply force to fibronectin molecules (see Figure 9.11). Crucial to the assembly process is integrin $\alpha_5\beta_1$, which specifically recognizes only fibronectin among all other matrix proteins, due to fibronectin's unique synergy site that is sitting on the $FIII_9$ module adjacent to the RGD site (see Figure 9.4). Fibronectin fibrils then emerge as the fibronectin-bound $\alpha_5\beta_1$ integrin complexes are translocated along with actin fibers away from the cell periphery [25, 56, 126, 127, 184, 194]. The tensin-mediated $\alpha_5\beta_1$ integrin translocation thus initiates fibronectin fibrillogenesis on the cell surface. However, an integrin-mediated activation step is not the only way by which fibronectin fibrillogenesis can be initiated. While integrins serve as handles by which cells can apply force to fibronectin molecules, fibronectin fibrillogenesis can also be induced in the absence of integrins, as long as fibronectin molecules are stretched, for example by shear [247] or physical entrapment [248]. The RGD-sequence is thus not necessary for fibronectin fibrillogenesis [249]. It was shown recently that the fibronectin conformation in artificially pulled fibronectin fibers is similar to that found in cell-generated matrix fibers [53–55, 250, 251]. Artificially pulled fibronectin fibers can thus serve as physiologically significant model systems. Finally, src-kinase activity not only regulates rigidity sensing but is also essential in fibronectin fibrillogenesis. Src-induced phosphorylation of paxillin at Y118 is required for assembly of the FN matrix by fibroblasts, as well as for maintaining the attachment of FN matrix fibrils to the cell surface [252].

9.10.1.1 Switching the Biochemistry Displayed by the Matrix by Stretching and Unfolding of Matrix Proteins

When the cells have assembled the extracellular matrix fibers, a number of questions remain unanswered. First, do the cells respond only to the rigidity of the matrix fibers, or can the cell-generated tension alter the biochemistry displayed by the matrix proteins? Can molecular recognition sites that confer biochemical specificity to proteins be altered by stretching proteins? Is it possible that cell-generated tension is sufficient not just to strain but to mechanically unfold those proteins that form part of the force-bearing protein ECM networks in living tissue? Conclusive evidence that cell-generated forces are sufficient to unfold ECM proteins, and that the unfolding imparts new functional switches, rely most importantly on visualization techniques to probe protein conformations *ex vivo* and in cell culture.

Fluorescence resonance energy transfer (FRET) between multiple energy donors and acceptors (Figure 9.17) was indeed used to show that fibronectin in cell culture is exposed to cell-induced dynamic levels of stress which lead to partial fibronectin unfolding [53–55, 69, 70]. Fibronectin unfolding is hypothesized to mediate a variety of functions, ranging from altered mechanisms for cell binding

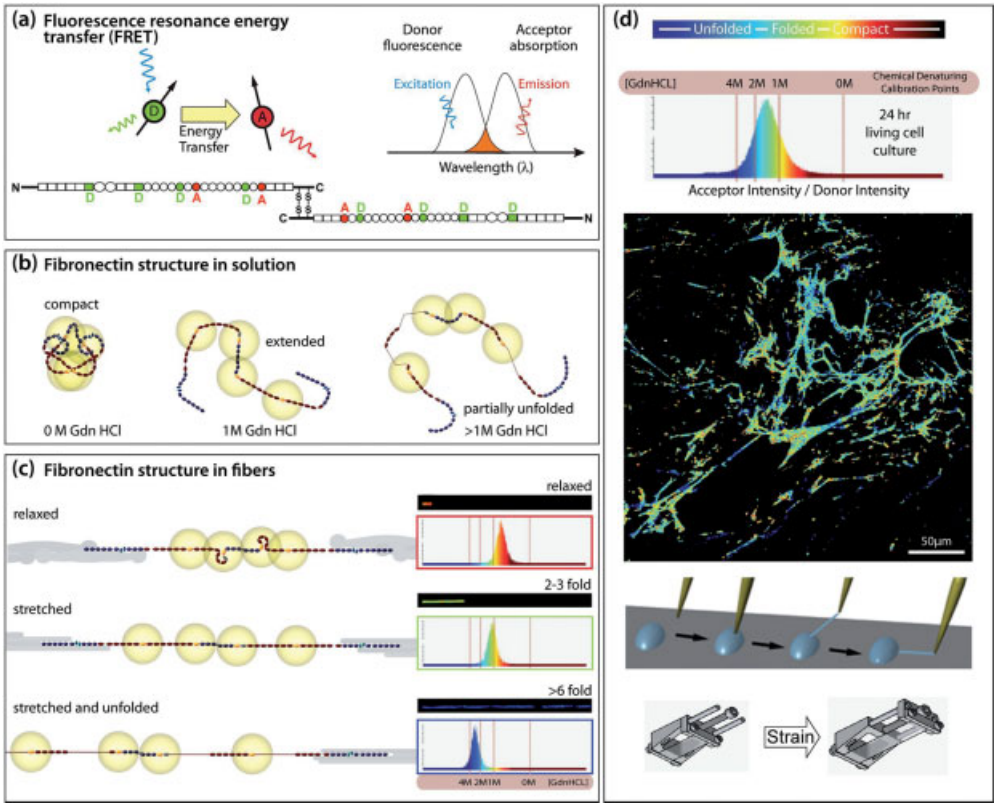


Figure 9.17 The use of fluorescence resonance energy transfer (FRET) to probe matrix stretching and unfolding in cell culture and artificially prepared fibronectin fibers. (a) The two free cysteines per fibronectin monomer, located on modules $FnIII_7$ and $FnIII_{15}$ (see Figure 9.4) are site-specifically labeled with the acceptors, while the donors are randomly distributed along the protein [69, 70]. The distance over which the acceptors can couple with potential energy donors are shown as yellow circles; (b) Schematic drawing of the fibronectin quaternary structure in solution and under denaturing conditions; (c) While the ultrastructure of fibronectin fibers remains unknown, FRET indicates that some quaternary structure is present when the fibers are fully relaxed. When the fibronectin fibers are stretched 200–300% of their equilibrium length, a first loss of secondary structure is observed [53–55, 250]; (d) Image of cell-made fibronectin matrix where trace amounts of FRET-labeled fibronectin were added to the cell culture medium. A broad distribution of different average fibronectin conformations can be seen. False colors have been used to visualize altered FRET ratios. Correlating FRET with mechanical strain was made possible by depositing fibronectin fibers onto stretchable PDMS sheets [250]. The fibers were therefore manually pulled out of a droplet of concentrated fibronectin solution. All of the conformations that can be seen in a broad range of strained artificial fibers coexist in every single field of view of a living cell culture.

to fibronectin to exposure of sites with enzymatic functions [21, 26, 53–55, 69, 70]. These cellular studies are important as they demonstrate that the nonequilibrium conformations of proteins can be stabilized by force and are thus physiologically relevant.

With regards to the translation of mechanical forces into biochemical signal changes, it should be noted that a wide variety of proteins (which are part of force-bearing protein networks linking the intracellular cytoskeleton and ECM) have multimodular structures where each individual module often carries one or more unique binding sites. However, what are the advantages of linking numerous modules equipped with different functionalities into macromolecules? The answer might be found in the following consideration. If domains of multimodular proteins were to possess similar mechanical stabilities, these domains would rupture in a stochastic sequence. But, if the domains have significantly different mechanical stabilities, as observed for fibronectin's FnIII modules [253–257], then sequential domain unfolding would lead to a well-defined graded sequence in which the various molecular recognition sites displayed by those domains would be altered as a function of mechanical strain. The significance of having a hierarchy of mechanical stabilities is thus central to mechanochemical signal conversion: as the mechanical hierarchy defines the sequence in which bonds or modules break, a sequence of stretching of different domains could be translated into a sequence of biochemical signal changes. This is also a likely explanation for why so many proteins contain buried cryptic sites in their hydrophobic interior that are only exposed when the protein unfolds [51, 52, 254, 256, 258–265]. Multidomain proteins are thus ideally suited to serve as mechanochemical signal converters to translate a large range of forces into sequential strain-specific functional changes [26].

Studies of the mechanical characteristics of proteins have thus revealed new insights into protein engineering principles. While fragments or a few domains of these molecules have often been employed in biotechnology or tissue engineering, mimicking only partial aspects of the whole molecule, engineers must ask what functional aspects might be missed if materials and surfaces were to be functionalized with peptides that contained only single molecular recognition motifs, as has been achieved for example with the cell-adhesive tripeptide RGD from the 450 kDa fibronectin molecule, instead of using the full-length protein. Such engineered systems would show a rather different mechanoresponse, and would not make it possible to translate a range of forces into a range of graded biochemical signal changes.

While the biochemistry of the quasi-equilibrium states of proteins is well understood, it is experimentally challenging to investigate how the structure–function relationship of proteins is altered when they are stretched. Deciphering the physiological significance of force-induced unfolding of fibronectin, and whether any of its versatile functions are up- or downregulated in a mechanoresponsive manner, however, has been hampered due to a lack of appropriate assays. In response to the growing need for quantitative biochemical and cellular assays that address whether the ECM acts as a mechanochemical signal converter to coregulate cellular mechanotransduction processes, we have developed a new assay where plasma fibronectin fibers are manually deposited onto elastic sheets, and force-induced changes in protein conformation are monitored using FRET (Figure 9.17). Our aim had been to develop a mechanical strain assays where the conformation of fibronectin can be adjusted externally on demand, while the force-induced protein extension is

monitored optically [250, 251]. To probe how the alterations of the structure of stretched Fn impact biochemical interactions and cellular behavior, such a strain assay needs to be amenable to cell culture environments. To tune the conformation of fibronectin, fibers were drawn manually from a concentrated fibronectin solution [266–268]. When adding trace amounts of FRET-labeled fibronectin into the solution, followed by a step where the freshly drawn fibers are deposited onto polymeric sheets that are mounted into stretch devices, fibers can be generated that have a far more narrow conformational distribution, as found in native matrix [250, 251]. Furthermore, the mechanical strain can be externally adjusted, which enabled protein-binding studies to be conducted as a function of the strain of fibrillar fibronectin [250, 251]. An image of conformationally tuned fibronectin fibers can be seen in Figure 9.17. These assays further open the doors to the question of whether – and how – cell phenotypes are regulated by force-induced alterations of fibronectin conformation.

9.10.1.2 Cell Responses to Initial Biomaterial Properties and Later to Self-Made Extracellular Matrix

While the differentiation of mesenchymal stem cells has been correlated with the rigidity of the substrate on which they were initially plated [19], does the rigidity response of the cell change as it assembles its own matrix hours or days after the cells have been seeded on a substrate with defined rigidity? For example, the initial rigidity of a substrate has been proposed to determine whether or not mammary epithelial cells upregulate integrin expression and differentiate into a malignant phenotype [145], and also to dictate whether mesenchymal stem cells differentiate into bone, muscle or neuronal tissue [19]. In those experiments, the macroscopic materials properties of the substrate were typically correlated with outcome after four to ten days of cell culturing. While the mechanical properties of a substrate or engineered scaffold have indeed been correlated with various aspects of cell behavior, the underlying mechanisms how substrate rigidity ultimately regulates the long-term responses have not yet been defined.

Once cells have been seeded onto synthetic matrices they rapidly begin to assemble their own matrix, and will ultimately touch, feel and respond to their self-made ECM. A few hours after cells have attached to surfaces and begun to assemble their matrix, the physical characteristics of the newly assembled matrix do indeed depend on the rigidity of the substrate. After 4 h of cell culture, the FRET data showed that the fibronectin matrix was indeed more unfolded on a rigid (33 kPa) compared to a soft substrate (7 kPa) (Figure 9.18a) [269]. Surprisingly, however, after only one day the fibroblasts that were initially seeded onto glass had produced sufficient matrix so as to sit on a much softer biopolymer cushion. The cells then assembled a matrix that was far less unfolded than the matrix they made during the first 4 h on glass, as probed by adding FRET-labeled fibronectin only during the last 23–24 h after seeding. This newly made matrix is comparable to the cells feeling a 7 kPa substrate [269]. Most interestingly, the aging matrix changes its physical properties. When the cells were seeded onto glass and allowed to assemble matrix for three days, the matrix deposited during the first 24 h was highly unfolded, while the younger matrix was far less

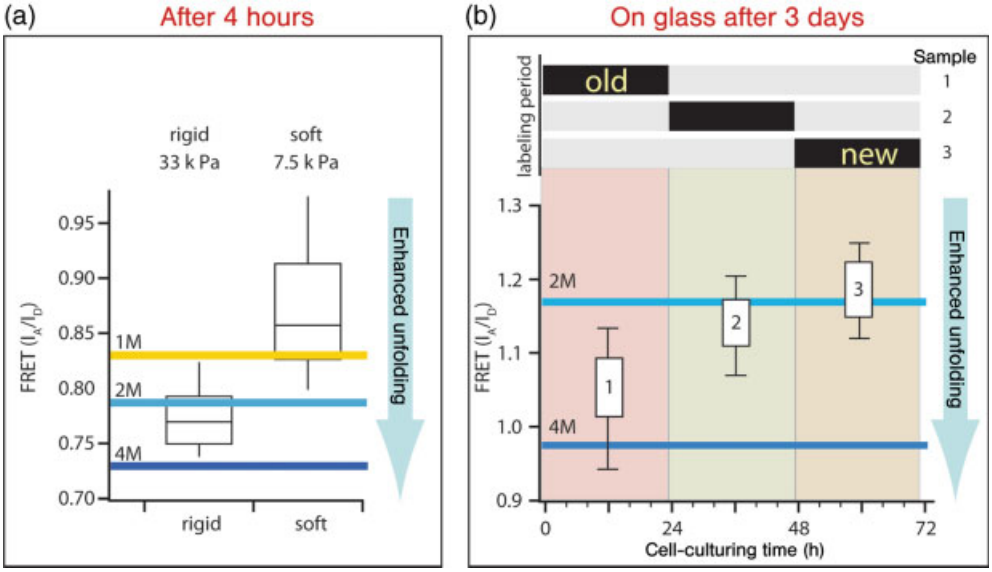


Figure 9.18 The tension of the extracellular matrix (and thus the mechanical strain of fibronectin) is upregulated with the rigidity of the substrate surface and changes as the matrix ages. (a) Fibronectin matrix assembly and unfolding on rigid and soft polyacrylamide surfaces 4 h after seeding the fibroblasts. The probabilities of finding certain FRET ratios are shown as boxes, where the maximum is given as the center line in the box, and the upper and lower ends of the boxes represent the 25th to 75th percentiles, and the 'whiskers' show the 2nd and 98th percentiles. FRET from fibrils on the rigid surface falls far below the FRET signature observed for fibronectin in solution at mild denaturing conditions (1 *M* GnHCl), thereby indicating that Fn is partially unfolded on rigid surfaces [53–55, 70, 250, 251]. FRET values on the soft surface indicate that the secondary structure of fibronectin is intact; (b) Three-day cell culture where FRET-labeled fibronectin was added for only limited time periods, as indicated in the upper bar graph. When the cells are seeded on glass and allowed to assemble matrix for three days, the matrix deposited during the first 24 h was highly unfolded, while the younger matrix was far less unfolded. Thus, the physical properties of matrix change as the matrix ages. (Adopted from Ref. [269].)

unfolded (Figure 9.18b). Thus, the matrix was progressively more unfolded as it aged, while the newly deposited matrix showed little unfolding. These data provided the first evidence that matrix maturation occurs, and that aging is associated with an increased stretching of fibronectin fibrils. Matrix assembly and remodeling involves at least partial unfolding of the secondary structure of fibronectin modules. Consequently, matured and aged matrix may display different physical and biochemical characteristics, and is structurally distinguishable from newly deposited matrix. A comparison of the conformation of Fn in these three-dimensional (3-D) matrices with those constructed by cells on rigid and flexible polyacrylamide surfaces suggests that cells in maturing matrices experience a microenvironment of gradually increased rigidity [269]. A future goal must be to understand the physiological consequences of matrix unfolding on cell function, including cancer and stem cell differentiation.

9.11
Cell Motility in Response to Force Generation and Matrix Properties

The relationship between force generation and motility is not simple. Fibroblasts that develop high forces on substrates do not move rapidly [222], whereas neutrophils that move at the highest rates reported for cells (about $40\,\mu m\,min^{-1}$ or $700\,nm\,s^{-1}$) generate very low forces on substrates [53–55]. There is considerable interest in the extravasation of cancer cells moving out of the bloodstream into tissues in the process of metastasis, although many of the steps in that process involve proteolysis of the matrix and deformation of the cell to pass through small gaps in the endothelium [270–272]. Although force generation is needed for motility, it is not the most important factor – indeed, cell polarization, matrix proteases, directional signals and the deformation of the cytoplasm are often as important.

Cell motility depends on substrate density and rigidity, and therefore also on the processes that respond to rigidity [273]. Many of the proteins involved in rigidity response have been linked to motility disorders, including cancer as well as malformations in development and neuronal connectivity. Src family kinases (SFKs) [274, 275], FAK [173, 174, 276]), the SH2 domain-containing phosphatase SHP-2 [173, 174] and RPTPs) [173, 174] are important components of the force-dependent signal transduction pathways that lead to the assembly of adhesion sites. The force-dependent initiation of adhesion sites and their rapid reinforcement occurs in protruding portions of cells, where adhesion sites can transmit cell propulsive forces [179, 277]. In extending regions of the cell, forces are generated on integrins by actin rearward flow rather than stress fibers. At the trailing end of the cell, mature focal adhesions create passive resistance during cell migration. To overcome this resistance, high forces must be generated by nascent adhesion sites [179]. However, in some static cells, higher forces are correlated with mature focal adhesions [58]. At the cell rear, traction stresses induce the disassembly instead of the reinforcement of focal adhesions and linked stress fibers [189, 278–280]; this is dependent on mechanosensitive ion channels and calcium signaling in keratocytes and astrocytoma cells. SFKs, FAK and PEST domain-enriched tyrosine phosphatase (PTP-PEST) are also crucial factors in adhesion site disassembly [229, 274, 281]. Recent studies have shown that the rigidity of 3-D matrices affects the migration rate differently from 2-D matrices, in that the less-rigid matrices cause an increase in migration rate [271]. At a biochemical level, the actin depolymerizing protein cofilin has been implicated in the motility of fibroblasts *in vitro*, and it is downstream in the biochemical signaling pathway of the integrins that have been activated at the leading edge [282–285]. These varying results indicate that different modalities of force generation and rigidity response at the cell front versus rear of the cell or in 3-D versus 2-D matrices can correlate with position-dependent regulation of phosphotyrosine signaling, and that different mechanisms of rigidity responses based on phosphotyrosine signaling can independently direct cell morphology as well as motility. Many observations point to tight links between morphology, migration, rigidity responses and tyrosine kinase activity.

9.12
Mechanical Forces and Force Sensing in Development and Disease

During the development and regeneration of tissues, forces act on and are propagated throughout most tissues. Such forces provide a local and global mechanism to shape cells and tissues, and to maintain homeostasis. Forces play a critical role by which cells interact with their environment and gain environmental feedback that regulates cell behavior. The signal for wound healing is often the loss of tissue integrity and the concomitant loss of force. Further, use of tissue and the periodic generation of force are often tied to the growth of that tissue, whereas inactivity is tied to the atrophy of the tissue. Bedridden patients suffer from a loss of muscle tissue and other aspects of atrophy. Similarly, with aging, osteoporosis, as well as many other cardiovascular diseases, mechanical changes and inappropriate responses of cells to mechanical changes, are critical and give rise to many symptoms. The size of the organism and its form are also set, at least in part, through a physical feedback between individual cells and their neighbors that is dynamic. As the cells grow and divide in development, they are constantly moving and even changing neighbors on occasion. The force-bearing cytoskeleton is actively remodeling and must clearly be responsive to changes in the level of force, or else the tissue would relax or contract too much. Contractile activity in individual cells can change the turgor of the tissue, and that parameter is under control of the signaling pathways that activate myosin contraction.

Consequently, forces play a critical role in health and disease in controlling the outcome of many biological processes (Figure 9.19). One obvious case is in cancer, where the cells ignore normal environmental cues and grow aberrantly, although there are many other examples (including problems with angiogenesis and tissue repair). Thus, it is speculated that in the future there will be an increasing emphasis

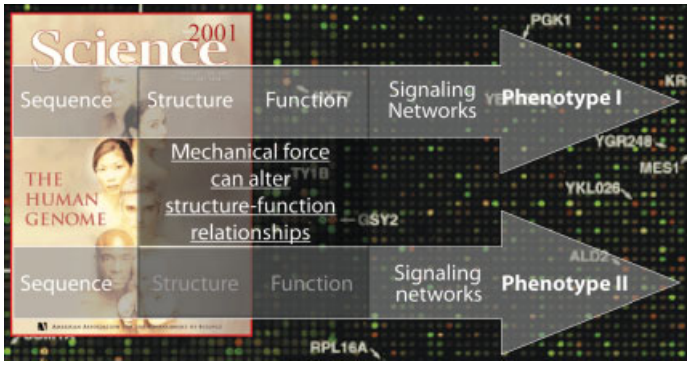


Figure 9.19 From the human genome to quantitative biology. The path is more complicated than originally thought, as mechanical force provides Nature with an additional dimension of regulating protein function. A switch in protein function might then alter cell signaling and thus the cell phenotype. The background shows the cover of *Science* announcing that details of the human genome had been resolved.

on the mechanical treatment of clinical problems and the targeting of therapeutics to mechanical response pathways. For the more effective treatment of diseases, there needs to be a greater recognition of the role that mechanical factors play in the development of the organism, as well as in the onset and progression of diseases. In other words, the genes code for a set of proteins that have developed the proper mechanical responses to shape the organism reproducibly. Similarly, disease-related alterations can result in alterations in mechanoresponsive pathways that are a major part of the disease. A much better understanding of those pathways is needed for proper treatment and therapy.

9.12.1
Cancer and Cell Transformation

Many cancer biologists have realized that cancer is inherently associated with a diseased mechanosensing and mechanoresponsive system. Many cancer cells ignore the normal environmental signals that regulate growth, and many of those cues are mechanical. For example, one of the early hallmarks of cancer cells is that they are often transformed, which was defined as the ability of those cells to grow on soft agar whereas normal cells required a rigid substrate [9–11]. Early observations linked transformation to uncontrolled cellular growth and to profound alterations in cell shape, as well as to the deregulation of tyrosine kinase and phosphatase activity. The first defined oncogene, *v-Src*, encodes an altered form of an important cellular tyrosine kinase, c-Src [286, 287]. In most studies on tumor cells, changes in morphology – but not cytoskeletal dynamics – have been reported.

The progression of cells from normal to a cancerous or even metastatic state is reflected in an increased softening of the cells, as probed by laser traps on suspended single cells [288]. Malignant fibroblasts, for example, have 30–35% less actin than normal cells. Transformed cells in culture often are rounder in morphology than primary cells. Tumor cells are also generally less adhesive than normal cells, and deposit less ECM [289], and the resulting loosened matrix adhesions may contribute to the ability of tumor cells to leave their original position in the tissue. In transformed cells, many aspects of nuclear and cell morphology as well as migration are altered. Focal adhesions can be replaced by podosomes and in addition, stress fibers can be absent [290]. Some transformed cells acquire anchorage independence – that is, they can grow without attachment to a substrate, suggesting a rigidity response deregulation. For example, transformed cells generate weak, poorly coordinated traction forces [291] but increased contractility. Thus, the one generality is that transformed cells can grow inappropriately, ignoring the mechanical cues of the environment that neighboring normal cells will follow to maintain appropriate tissue morphology. Although other factors, such as hormonal signals, form part of many cancers, the inappropriate mechanical responsiveness of cancer cells must also be considered as an important part of the process.

Cancer progression leads to a loss of tissue differentiation due to abnormal cell proliferation rates. Even if isolated malignant cells are associated with an increased softness of their overall cytoskeleton, it is equaly significant that tumors have a stiffer

ECM [145, 292, 293]. Malignant transformations of the breast, for example, are associated with dramatic changes in gland tension that include an increased ECM stiffness of the surrounding stroma [293]. Chronically increased mammary gland tension may influence tumor growth, perturb tissue morphogenesis, facilitate tumor invasion, and alter tumor survival and treatment responsiveness. However, changes in environmental factors (i.e. changes in ECM rigidity) and internal force generation (i.e. inappropriate rigidity responses) might be key factors in determining a transformed cell morphology and malignant phenotype [145]. For example, tumors are stiffer than normal tissue because they have a stiff stroma and elevated Rho-dependent cytoskeletal tension that drives focal adhesions, disrupts adherens junctions, perturbs tissue polarity, enhances growth, and hinders lumen formation [145]. Matrix stiffness thereby perturbs epithelial morphogenesis by clustering integrins to enhance ERK activation and increase ROCK-generated contractility and focal adhesions, thereby promoting malignant behavior [145].

Metastatic cells escape the tumor by invading the surrounding tissue, entering the circulation and finally attaching to previously unaffected tissues in often remote locations. In 1889, Stephen Paget published an article in *The Lancet* that described the propensity of various types of cancer to form metastases in specific organs, and proposed that these patterns were due to the "... dependence of the seed (the cancer cell) on the soil (the secondary organ)" [294]. This has often been linked to the chemical environment of the secondary organ, although recent results have indicated that it could also be a result of the mechanical environment in the secondary organ [295]. It was found that lung metastases from human breast cancer cells would grow better on soft fibronectin substrates than on hard, whereas bone metastases would grow better on hard than on soft (A. Kostic and M.P. Sheetz, unpublished results). Metastasis is an inefficient process, and many cancer cells are shed but few actually grow into a tumor at a new site. One reason for this is that the new site might not have the appropriate mechanical properties. At another level, tumor cells are generally less adhesive than normal cells and deposit less ECM [289]. The resulting loosened matrix adhesions, combined with the softened cytoskeleton, may contribute to the ability of tumor cells to leave their original position in the tissue and squeeze through tiny holes.

Many of the molecules discussed above play key roles in cancer progression, and also metastasis. Integrin-mediated cell adhesion leads to the activation of FAK and c-Src, after which the FAK–Src complex binds to and can phosphorylate various adaptor proteins such as p130Cas and paxillin. The results of recent studies have shown that the FAK–Src complex is activated in many tumor cells, and generates signals leading to tumor growth and metastasis (as reviewed in Ref. [296]). Tyrosine phosphorylation of paxillin regulates both the assembly and turnover of adhesion sites. Phosphorylated paxillin enhanced lamellipodial protrusions and thus promoted cell migration [126, 127]; the migration of tumor cells in 3-D matrices is then governed by matrix stiffness, along with cell–matrix adhesion and proteolysis [271]. As discussed above, the phosphorylation of p130Cas is upregulated when cells are located on a more rigid substrate.

The overall survival of breast cancer patients is inversely correlated with the levels of p130Cas (BCAR1) in the tumors [297] indicating that increased levels of p130Cas in

tumor cells contributed to patient death. It was subsequently found that cell migration was activated by p130Cas and the associated GEF (AND-34 or BCAR3), which indicated that metastasis was favored by elevated p130Cas [298]. Both, p130Cas and a p130Cas binding protein, AND34 (BCAR3), will increase the epithelial to mesenchymal transition when overexpressed [298]. p130Cas appears to have a central role in cell growth and motility; in many cases, it is dramatically altered in its phosphotyrosine levels in correlation with transformation [299–302] and metastasis [303]. In the specific case of lung tumors, metastasis was increased following removal of the primary tumor, and required p130Cas expression. Further, the substrate domain YxxP tyrosines were needed for both invasive and metastatic properties of the cells [304]. Even the invasion of Matrigel and the formation of large podosome structures required the YxxP tyrosines. Thus, it is suggested that the inappropriate growth of cancer cells may be partly due to changes in the normal force and rigidity-sensing pathways that can alter the cellular program. This means, in turn, that the protein mechanisms involved in controlling mechanical responses can be good targets for therapies in cancer. In addition, mechanical treatments can possibly alter the course of cancers. Several levels can be identified in the process of mechanosensing, transduction and response where alterations in cancer cells could result in abnormal growth control. For example, c-Src, Fyn and Yes knockout cells are each missing three important Src family kinases, and will grow on soft agar while not sensing any difference between soft and hard agar. However, the restoration of Fyn activity will enable the cells to sense rigidity, and they will no longer grow on soft agar [198]. Thus, an understanding of the mechanisms of force and rigidity sensing can provide an important perspective on cancer.

9.12.2 Angiogenesis

The growth of new blood vessels – that is, angiogenesis – is crucial not only in tissue growth and remodeling but also in wound healing and cancer. Vascular development requires correct interactions among endothelial cells, pericytes and surrounding cells [24]. Thus, the formation of new blood vessels might be compromised if any of these interactions – including cell–matrix interactions, both with basement membranes and with surrounding ECMs – are perturbed. Equally important, the injury-mediated degradation of the ECM can lead to changes in matrix–integrin interactions, causing an impaired reactivity of the endothelial cells that will lead to vascular wall remodeling. Consequently, alterations in integrin signaling, growth factor signaling, and even of the architecture and composition of the ECM, might all affect vascular development. As in other motility processes, angiogenesis involves a very stereotypical set of movements of the endothelial cells that result in the formation of capillary tubes.

The role and mechanisms by which mechanical forces promote angiogenesis remain unclear. It is notable, however, that angiogenesis is regulated by integrin signaling [305–308]. Angiogenesis is furthermore promoted by vascular endothelial growth factor (VEGF). As tumor neovascularization plays critical roles for the development, progression and metastasis of cancers, new therapeutic approaches to treat malignancies have been aimed at controlling angiogenesis by monoclonal

antibodies targeting VEGF, as well as with several tyrosine kinase inhibitors targeting VEGF-related pathways (for a review, see Ref. [309]).

VEGF binds to its transmembrane receptor by stimulated complex formation between VEGF receptor-2 and β_3 integrin. Prior studies have suggested, for example, that α_v-integrins ($\alpha_v\beta_3$ and $\alpha_v\beta_5$) could act as negative regulators of angiogenesis (as discussed in Refs [31, 32]). Neovascularization is impaired in mutant mice where the β_3 integrins were unable to undergo tyrosine phosphorylation [310]. The lack of integrin phosphorylation suppressed the complex formation with VEGF. Furthermore, the phosphorylation of VEGF receptor-2 was significantly reduced in cells expressing mutant β_3 compared to wild-type, leading to an impaired integrin activation in these cells. With its binding locations at both the N and C termini, VEGF also binds to fibronectin fibers [311, 312] and, when bound, has been shown to increase cell migration, proliferation and differentiation [311, 313, 314]. A reduced extracellular pH is one of the key signals that can induce angiogenesis. By demonstrating that VEGF binding to fibronectin is dependent on pH, and that released VEGF sustained biological activity, Goerges *et al.* [315] suggested that cells may use a lowered pH as a localized mechanism of controlled VEGF release [316]. Goerges and colleagues also suggested that VEGF might be stored in the ECM via interactions with fibronectin and heparan sulfate in tissues that are in need of vascularization, so that it can aid in directing the dynamic process of growth and migration of new blood vessels. If – and how – VEGF signaling is regulated by mechanical force, however, remains unclear.

Tumor blood vessels have an altered integrin expression profile, and both blood and lymphatic vessels have pathological lesions [28]. In contrast to healthy tissue, integrin β_4 signaling in tumor blood vessels promotes the onset of the invasive phase of pathological angiogenesis [317], while loss of the β_4 integrin subunit reduces tumorigenicity [318]. Integrin β_4 binds to laminin (Figure 9.5), which is enriched in basement membranes, but not to the RGD-ligand as exposed in fibronectin. Another difference from the RGD-binding integrins is that integrin β_4 connects to the cytoskeleton via plectin (not talin, as illustrated in Figure 9.7) [319], and little is known about the mechanoresponsivity of that linkage.

Another open question is whether degradation of the ECM is regulated by force. Exploring this question is of particular relevance since, in addition to serving as an anchoring scaffold and storage for growth factors, a group of angiogenesis regulators are derived from fragments of ECM or blood proteins. Endostatin, antithrombin and anastellin are members of this group of substances. Some of these compounds are currently undergoing clinical trials as inhibitors of tumor angiogenesis [320], as well as synthetic peptides modeled after these anti-angiogenic proteins, such as Anginex [321]. RGD-containing breakdown products of the ECM may also cause sustained vasodilation [87].

9.12.3
Tissue Engineering

Deciphering the mechanisms by which ECMs might sense and transduce mechanical stimulation into functional alterations of cell behavior and fate is also a critical

issue in advancing tissue engineering and regenerative medicine, as our abilities to interface synthetic materials with living soft tissue to promote angiogenesis and healing are still rather limited. Beyond artificial skin, few advances have been made when using synthetic scaffolds to grow functional soft organs or organ patches that can at least support some crucial physiological organ functions. The conventional thinking was that once surfaces are coated with the correct set of biomolecules, the cells might recognize them as 'biological'. If it will indeed be confirmed that cells have the ability to dynamically regulate the biochemical display of the surrounding matrix on demand by applying force, the currently pursued more 'static' approaches for designing tissue engineering scaffolds do neglect force as major regulatory factor of ECM function [322]. If the cells are in contact with synthetic surfaces, their ability to dynamically regulate the conformations of matrix proteins by stretching them might be compromised. A full appreciation of the engineering principles of adhesion molecules, and of the complexity by which ECM can respond to cell contractility [26], might thus lead to new approaches for how to better engineer the interface between cells and synthetic materials.

The engineering of scaffolds that can regenerate soft organs or support some soft organ functions remains a daunting task. Scaffolds derived from natural tissues or matrix proteins have so far shown significantly better clinical performance than their synthetic counterparts [323, 324]. The creation of a bioartificial heart, for example, requires the engineering of cardiac architecture, as well as of appropriate cellular constituents and pump function. A major breakthrough in bioartificial heart engineering has recently been reported [14]. While many approaches have been attempted in the past, some success was achieved by using decellularized organ-specific matrices as scaffolds. For this, hearts were first decellularized by coronary perfusion with detergents; this preserved the underlying ECM, the vascular architecture, competent acellular valves and an intact chamber geometry (Figure 9.20). After decellularization, collagens I and III, laminin and fibronectin remained within the thinned heart matrix. The fiber composition and orientation of the myocardial ECM were preserved, the ventricular ECM was retained, and the vascular basal membranes remained intact. In order to mimic cardiac cell composition, these decellularized heart matrices were reseeded first with cardiac and later endothelial cells [14], with macroscopic contractions being observed at day 4. By day 8, under physiological load and electrical stimulation, the constructs could generate a pump function (equivalent to about 2% of adult or 25% of 16-week fetal heart function). Notably however, perfusion and physiological stimulation were absolutely needed for tissue formation and to regain tissue function. The authors speculated that such organs, if matured even further, could become transplantable either in part (e.g. as a ventricle for congenital heart disease such as hypoplastic left heart syndrome) or as an entire donor heart in end-stage heart failure. The technique was subsequently applied to a variety of mammalian organs, including lung, liver, kidney and muscle. Ongoing studies are directed towards optimizing reseeding strategies to promote the dispersion of cells throughout the construct, *in vitro* conditions required for organ maturation, and the choice of stem or progenitor cells necessary to generate either autologous or off-the-shelf bioartificial solid organs for transplantation. In contrast,

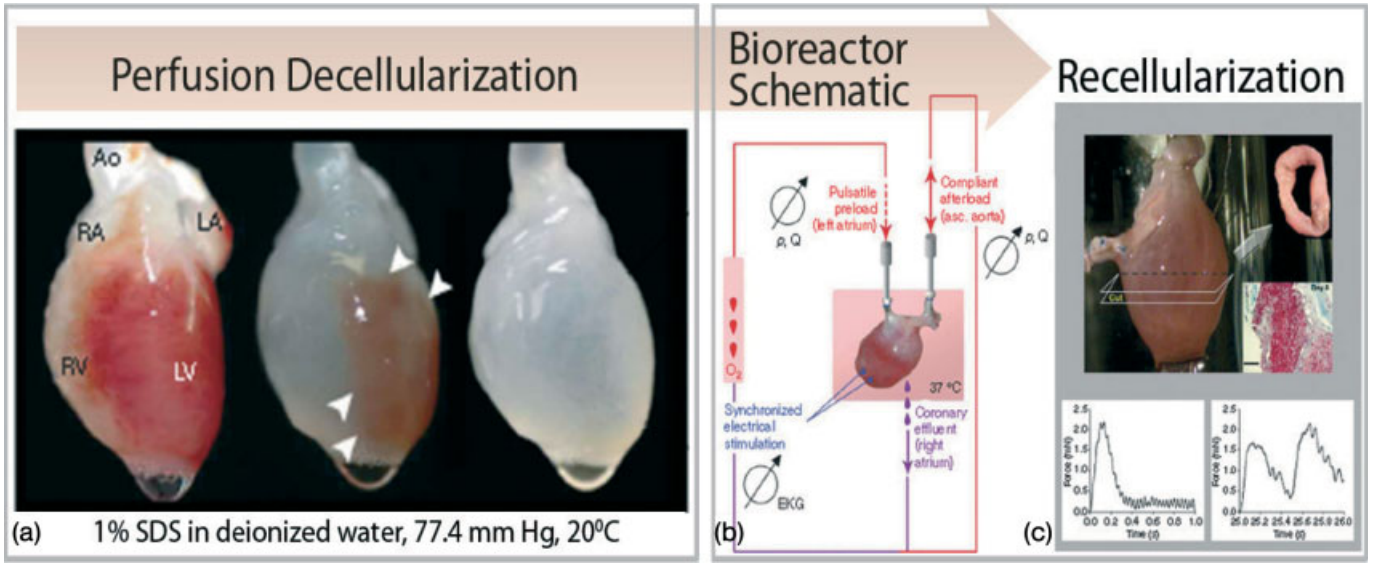


Figure 9.20 Decellularization and recellularization of a working heart-like construct. (a) Perfusion decellularization of whole rat hearts. Photographs of cadaveric rat hearts mounted on a Langendorff apparatus. Ao = aorta; LA = left atrium; LV = left ventricle; RA = right atrium; RV = right ventricle. Retrograde perfusion of cadaveric rat heart using sodium dodecyl sulfate (SDS) over 12 h. The heart becomes more translucent as cellular material is washed out from the right ventricle, followed by the atria and finally the left ventricle; (b) Schematic of working heart bioreactor showing cannulation of the left atrium and ascending (asc.) aorta. The heart is exposed to physiological preload, afterload and intraventricular pressure, and electrically stimulated at 5–20 V. Oxygenated medium containing serum and antibiotics enters through the left atrium and exits through the aortic valve. Pulsatile distention of the left ventricle and a compliance loop attached to the ascending aorta provide physiological coronary perfusion and afterload. Coronary perfusate (effluent) exits through the right atrium; (c) Formation of a working perfused bioartificial heart-like construct by recellularization of decellularized cardiac extracellular matrix. Top, recellularized whole rat heart at day 4 of perfusion culture in a working heart bioreactor. Upper insert: cross-sectional ring harvested for functional analysis (day 8). Lower insert: Masson's trichrome staining of a ring thin section, showing cells throughout the thickness of the wall. Scale bar = 100 mm. Bottom: force generation in left ventricular rings after 1 Hz (left) and 2 Hz (right) electrical stimulation.

stem cells directly injected into scarred tissue, for example into an infarcted heart, are not properly directed by the matrix to form new myocardium [325]. Rather than injecting cells into rigidified scar tissue, tissue regeneration was far more effective when transplanting an entire monolayer sheet of mesenchymal stem cells [326]. The engrafted cell sheet gradually grew to form a thick stratum that included newly formed vessels, undifferentiated cells and few cardiomyocytes which might be promoted by the more favorable microenvironment that the cells would find if transplanted as sheets rather than being injected individually.

Early studies have already shown that matrix crosslinking would compromise the ability of tissue-derived matrices for use in functional reconstruction [327]. One of several reasons for this might be that crosslinking alters the rigidity of the matrix, and an upregulated rigidity response of the reseeded cells might interfere with regaining tissue function. Another – not necessarily exclusive – possibility is that crosslinking would inhibit the protein conformation changes caused by stretching the matrix fibers. It has been found that fibronectin fibers, for example, can be stretched on average more than five times their equilibrium length before they break [250, 251], whereas crosslinked fibers show a markedly decreased extensibility. Crosslinked cell-derived matrices cause an upregulated cellular rigidity response and alter the biophysical properties of the matrix that the newly seeded cells are generating [328]. Thus, force-induced protein unfolding in a newly deposited matrix is, at least in part, upregulated by crosslinking, with all the functional implications as discussed above. Another aspect of native matrix, that might be compromised by crosslinking, is its ability to serve as a scaffold for storing cytokines and growth factors and to release them upon demand. Integrins were also shown recently to play a central role in activating the matrix-bound cytokine transforming growth factor-beta 1 (TGF-β1) by cell-generated tension acting on the matrix [329]. TGF-β1 controls tissue homeostasis in embryonic and normal adult tissues, and also contributes to the development of fibrosis, cancer, autoimmune and vascular diseases. In most of these conditions, active TGF-β1 is generated by dissociation from a large latent protein complex that sequesters latent TGF-β1 in the fibronectin-containing ECM [330]. The studies of Wipff and colleagues might suggest that matrix stiffness could regulate the equilibrium between storage and release of a host of matrix-bound growth factors [331].

Finally, the fact that not only the intact ECM but also its breakdown products have regulatory functions, can be actively exploited in tissue engineering. Low-molecular-weight peptides derived from the ECM, for example, can act as chemo-attractants for primary endothelial cells [138]. ECM extracts were found to have antimicrobial activity [332], and fragments of ECM or blood proteins, including endostatin, antithrombin and anastellin, may serves as inhibitors of angiogenesis [321]. Moreover, these angiostatic peptides use plasma fibronectin to home to the angiogenic vasculature [321]. Finally, uncharacterized digestive products of the ECM seem to act as strong inflammatory mediators [333]. Extensive future investigations are required in order to provide a full comprehension of the multifaceted regulatory roles of the ECM and its constituents, and how forces might coregulate many of these functions.

Consequently, learning how to switch the structure–function relationship of proteins by force has far-reaching potential not only in tissue engineering but also

in biotechnology, and for the development of new drugs that might target proteins stretched into nonequilibrium states.

Acknowledgments

We gratefully acknowledge the many discussions with colleagues and our students, and thank in particular Sheila Luna for the graphics. Financial support was provided by the Nanotechnology Center for Mechanics in Regenerative Medicine (an NIH Roadmap Nanomedicine Development Center), the Volkswagen Stiftung, and various grants from NIH and ETH Zurich.

References

1 Korge, B.P. and Krieg, T. (1996) The molecular basis for inherited bullous diseases. *Journal of Molecular Medicine (Berlin, Germany)*, **74** (2), 59–70.

2 McGrath, J.A. (1999) Hereditary diseases of desmosomes. *Journal of Dermatological Science*, **20** (2), 85–91.

3 Jonkman, M.F., Pas, H.H., Nijenhuis, M., Kloosterhuis, G. and Steege, G. (2002) Deletion of a cytoplasmic domain of integrin beta4 causes epidermolysis bullosa simplex. *The Journal of Investigative Dermatology*, **119** (6), 1275–1281.

4 Pedersen, B.K. (2006) The anti-inflammatory effect of exercise: its role in diabetes and cardiovascular disease control. *Essays in Biochemistry*, **42**, 105–117.

5 Petersen, A.M. and Pedersen, B.K. (2005) The anti-inflammatory effect of exercise. *Journal of Applied Physiology (Bethesda, Md: 1985)*, **98** (4), 1154–1162.

6 Fries, R.S., Mahboubi, P., Mahapatra, N.R., Mahata, S.K., Schork, N.J., Schmid-Schoenbein, G.W. and O'Connor, D.T. (2004) Neuroendocrine transcriptome in genetic hypertension: multiple changes in diverse adrenal physiological systems. *Hypertension*, **43** (6), 1301–1311.

7 Harrison, D.G., Widder, J., Grumbach, I., Chen, W., Weber, M. and Searles, C. (2006) Endothelial mechanotransduction, nitric oxide and vascular inflammation. *Journal of Internal Medicine*, **259** (4), 351–363.

8 McGarry, J.D. (2002) Banting lecture 2001: dysregulation of fatty acid metabolism in the etiology of type 2 diabetes. *Diabetes*, **51** (1), 7–18.

9 Quigley, J.P. (1979) Phorbol ester-induced morphological changes in transformed chick fibroblasts: evidence for direct catalytic involvement of plasminogen activator. *Cell*, **17** (1), 131–141.

10 Giguere, L. and Gospodarowicz, D. (1983) Effect of Rous sarcoma virus transformation of rat-1 fibroblasts upon their growth factor and anchorage requirements in serum-free medium. *Cancer Research*, **43** (5), 2121–2130.

11 McClure, D.B. (1983) Anchorage-independent colony formation of SV40 transformed BALB/c-3T3 cells in serum-free medium: role of cell- and serum-derived factors. *Cell*, **32** (3), 999–1006.

12 Riha, G.M., Lin, P.H., Lumsden, A.B., Yao, Q. and Chen, C. (2005) Roles of hemodynamic forces in vascular cell differentiation. *Annals of Biomedical Engineering*, **33** (6), 772–779.

13 Jacques, A.M., Briceno, N., Messer, A.E., Gallon, C.E., Jalilzadeh, S., Garcia, E., Kikonda-Kanda, G., Goddard, J., Harding, S.E., Watkins, H., Esteban, M.T., Tsang,

V.T., McKenna, W.J. and Marston, S.B. (2008) The molecular phenotype of human cardiac myosin associated with hypertrophic obstructive cardiomyopathy. *Cardiovascular Research*, **79**, 481–491.

14 Ott, H.C., Matthiesen, T.S., Goh, S.K., Black, L.D., Kren, S.M., Netoff, T.I. and Taylor, D.A. (2008) Perfusion-decellularized matrix: using nature's platform to engineer a bioartificial heart. *Nature Medicine*, **14** (2), 213–221.

15 Liu, W.F., Nelson, C.M., Tan, J.L. and Chen, C.S. (2007) Cadherins, RhoA, and Rac1 are differentially required for stretch-mediated proliferation in endothelial versus smooth muscle cells. *Circulation Research*, **101** (5), e44–e52.

16 Sims, T.N., Soos, T.J., Xenias, H.S., Dubin-Thaler, B., Hofman, J.M., Waite, J.C., Cameron, T.O., Thomas, V.K., Varma, R., Wiggins, C.H., Sheetz, M.P., Littman, D.R. and Dustin, M.L. (2007) Opposing effects of PKCtheta and WASp on symmetry breaking and relocation of the immunological synapse. *Cell*, **129** (4), 773–785.

17 Badylak, S.F. (2007) The extracellular matrix as a biologic scaffold material. *Biomaterials*, **28** (25), 3587–3593.

18 Discher, D.E., Janmey, P. and Wang, Y.L. (2005) Tissue cells feel and respond to the stiffness of their substrate. *Science*, **310** (5751), 1139–1143.

19 Engler, A.J., Sen, S., Sweeney, H.L. and Discher, D.E. (2006) Matrix elasticity directs stem cell lineage specification. *Cell*, **126** (4), 677–689.

20 Ingber, D.E. (2006) Cellular mechanotransduction: putting all the pieces together again. *The FASEB Journal*, **20** (7), 811–827.

21 Vogel, V. and Sheetz, M. (2006) Local force and geometry sensing regulate cell functions. *Nature Reviews - Molecular Cell Biology*, **7** (4), 265–275.

22 Hartsock, A. and Nelson, W.J. (2008) Adherens and tight junctions: structure, function and connections to the actin cytoskeleton. *Biochimica et Biophysica Acta*, **1778** (3), 660–669.

23 Farhadifar, R., Roper, J.C., Aigouy, B., Eaton, S. and Julicher, F. (2007) The influence of cell mechanics, cell-cell interactions, and proliferation on epithelial packing. *Current Biology*, **17** (24), 2095–2104.

24 Hynes, R.O. (2007) Cell-matrix adhesion in vascular development. *Journal of Thrombosis and Haemostasis*, **5** (Suppl. 1), 32–40.

25 Mao, Y. and Schwarzbauer, J.E. (2005) Fibronectin fibrillogenesis, a cell-mediated matrix assembly process. *Matrix Biology*, **24**, 389–399.

26 Vogel, V. (2006) Mechanotransduction involving multimodular proteins: converting force into biochemical signals. *Annual Review of Biophysics and Biomolecular Structure*, **35**, 459–488.

27 Kadler, K.E., Hill, A. and Canty-Laird, E.G. (2008) Collagen fibrillogenesis: fibronectin, integrins, and minor collagens as organizers and nucleators. *Current Opinion in Cell Biology*, **20**, 495–501.

28 Ruoslahti, E. (2004) Vascular zip codes in angiogenesis and metastasis. *Biochemical Society Transactions*, **32** (Pt 3), 397–402.

29 Zaman, M.H. (2007) Understanding the molecular basis for differential binding of integrins to collagen and gelatine. *Biophysical Journal*, **92** (2), L17–L19.

30 Mizejewski, G.J. (1999) Role of integrins in cancer: survey of expression patterns. *Proceedings of the Society for Experimental Biology and Medicine*, **222** (2), 124–138.

31 Hynes, R.O. (2002) Integrins: bidirectional, allosteric signaling machines. *Cell*, **110** (6), 673–687.

32 Hynes, R.O. (2002) A reevaluation of integrins as regulators of angiogenesis. *Nature Medicine*, **8** (9), 918–921.

33 Ginsberg, M.H., Partridge, A. and Shattil, S.J. (2005) Integrin regulation. *Current Opinion in Cell Biology*, **17** (5), 509–516.

34 Arnaout, M.A., Goodman, S.L. and Xiong, J.P. (2007) Structure and mechanics of integrin-based cell adhesion. *Current Opinion in Cell Biology*, **19** (5), 495–507.

35 Luo, B.H., Carman, C.V. and Springer, T.A. (2007) Structural basis of integrin regulation and signalling. *Annual Review of Immunology*, **25**, 619–647.

36 Petrich, B.G., Marchese, P., Ruggeri, Z.M., Spiess, S., Weichert, R.A., Ye, F., Tiedt, R., Skoda, R.C., Monkley, S.J., Critchley, D.R. and Ginsberg, M.H. (2007) Talin is required for integrin-mediated platelet function in hemostasis and thrombosis. *The Journal of Experimental Medicine*, **204** (13), 3103–3111.

37 Han, Y., Cowin, S.C., Schaffler, M.B. and Weinbaum, S. (2004) Mechanotransduction and strain amplification in osteocyte cell processes. *Proceedings of the National Academy of Sciences of the United States of America*, **101** (47), 16689–16694.

38 Davies, P.F., Spaan, J.A. and Krams, R. (2005) Shear stress biology of the endothelium. *Annals of Biomedical Engineering*, **33** (12), 1714–1718.

39 Chien, S. (2008) Effects of disturbed flow on endothelial cells. *Annals of Biomedical Engineering*, **36** (4), 554–562.

40 Robling, A.G., Castillo, A.B. and Turner, C.H. (2006) Biomechanical and molecular regulation of bone remodelling. *Annual Review of Biomedical Engineering*, **8**, 455–498.

41 Wang, Y., McNamara, L.M., Schaffler, M.B. and Weinbaum, S. (2007) A model for the role of integrins in flow induced mechanotransduction in osteocytes. *Proceedings of the National Academy of Sciences of the United States of America*, **104** (40), 15941–15946.

42 Hooper, S.B. and Wallace, M.J. (2006) Role of the physicochemical environment in lung development. *Clinical and Experimental Pharmacology and Physiology*, **33** (3), 273–279.

43 Vlahakis, N.E. and Hubmayr, R.D. (2003) Response of alveolar cells to mechanical stress. *Current Opinion in Critical Care*, **9** (1), 2–8.

44 Classen, A.K., Anderson, K.I., Marois, E. and Eaton, S. (2005) Hexagonal packing of Drosophila wing epithelial cells by the planar cell polarity pathway. *Developmental Cell*, **9** (6), 805–817.

45 Marois, E. and Eaton, S. (2007) RNAi in the Hedgehog signaling pathway: pFRiPE, a vector for temporally and spatially controlled RNAi in *Drosophila*. *Methods in Molecular Biology (Clifton, NJ)*, **397**, 115–128.

46 Bao, G. and Suresh, S. (2003) Cell and molecular mechanics of biological materials. *Nature Materials*, **2** (11), 715–725.

47 Bustamante, C., Chemla, Y.R., Forde, N.R. and Izhaky, D. (2004) Mechanical processes in biochemistry. *Annual Review of Biochemistry*, **73**, 705–748.

48 Chen, C.S., Tan, J. and Tien, J. (2004) Mechanotransduction at cell-matrix and cell-cell contacts. *Annual Review of Biomedical Engineering*, **6**, 275–302.

49 Chen, J., Takagi, J., Xie, C., Xiao, T., Luo, B.H. and Springer, T.A. (2004) The relative influence of metal ion binding sites in the I-like domain and the interface with the hybrid domain on rolling and firm adhesion by integrin alpha4beta7. *The Journal of Biological Chemistry*, **279** (53), 55556–55561.

50 Orr, A.W., Helmke, B.P., Blackman, B.R. and Schwartz, M.A. (2006) Mechanisms of mechanotransduction. *Developmental Cell*, **10** (1), 11–20.

51 Sawada, Y., Tamada, M., Dubin-Thaler, B.J., Cherniavskaya, O., Sakai, R., Tanaka, S. and Sheetz, M.P. (2006) Force sensing by mechanical extension of the Src family kinase substrate p130Cas. *Cell*, **127** (5), 1015–1026.

52 Johnson, C.P., Tang, H.Y., Carag, C., Speicher, D.W. and Discher, D.E. (2007) Forced unfolding of proteins within cells. *Science*, **317** (5838), 663–666.

53 Smith, L.A., Aranda-Espinoza, H., Haun, J.B., Dembo, M. and Hammer, D.A. (2007) Neutrophil traction stresses are concentrated in the uropod during migration. *Biophysical Journal*, **92** (7), L58–L60.

54 Smith, M.L., Gourdon, D., Little, W.C., Kubow, K.E., Eguiluz, R.A., Luna-Morris, S. and Vogel, V. (2007) Force-induced unfolding of fibronectin in the extracellular matrix of living cells. *PLoS Biology*, **5** (10), e268.

55 Smith, M.L., Gourdon, D., Little, W.C., Kubow, K.E., Eguiluz, R.A., Luna-Morris, S. and Vogel, V. (2007) Force-induced unfolding of fibronectin in the extracellular matrix of living cells. *Public Library of Science Biology*, **5** (10).

56 Vogel, V., Sheetz, M.P. (2009) Cell Fate Regulation by Coupling Mechanical Cycles to Biochemical Signaling Pathways. *Current Opinion Cell Biol.* **21** (1):in press.

57 Harris, A.K., Wild, P. and Stopak, D. (1980) Silicone rubber substrata: a new wrinkle in the study of cell locomotion. *Science*, **208** (4440), 177–179.

58 Balaban, N.Q., Schwarz, U.S., Riveline, D., Goichberg, P., Tzur, G., Sabanay, I., Mahalu, D., Safran, S., Bershadsky, A., Addadi, L. and Geiger, B. (2001) Force and focal adhesion assembly: a close relationship studied using elastic micropatterned substrates. *Nature Cell Biology*, **3** (5), 466–472.

59 Tan, J.L., Tien, J., Pirone, D.M., Gray, D.S., Bhadriraju, K. and Chen, C.S. (2003) Cells lying on a bed of microneedles: an approach to isolate mechanical force. *Proceedings of the National Academy of Sciences of the United States of America*, **100** (4), 1484–1489.

60 Chrzanowska-Wodnicka, M. and Burridge, K. (1996) Rho-stimulated contractility drives the formation of stress fibers and focal adhesions. *The Journal of Cell Biology*, **133** (6), 1403–1415.

61 Helfman, D.M., Levy, E.T., Berthier, C., Shtutman, M., Riveline, D., Grosheva, I., Lachish-Zalait, A., Elbaum, M. and Bershadsky, A.D. (1999) Caldesmon inhibits nonmuscle cell contractility and interferes with the formation of focal adhesions. *Molecular Biology of the Cell*, **10** (10), 3097–3112.

62 Riveline, D., Zamir, E., Balaban, N.Q., Schwarz, U.S., Ishizaki, T., Narumiya, S., Kam, Z., Geiger, B. and Bershadsky, A.D. (2001) Focal contacts as mechanosensors: externally applied local mechanical force induces growth of focal contacts by an mDia1-dependent and ROCK-independent mechanism. *The Journal of Cell Biology*, **153** (6), 1175–1186.

63 Galbraith, C.G., Yamada, K.M. and Sheetz, M.P. (2002) The relationship between force and focal complex development. *The Journal of Cell Biology*, **159** (4), 695–705.

64 Giannone, G., Jiang, G., Sutton, D.H., Critchley, D.R. and Sheetz, M.P. (2003) Talin1 is critical for force-dependent reinforcement of initial integrin-cytoskeleton bonds but not tyrosine kinase activation. *The Journal of Cell Biology*, **163** (2), 409–419.

65 Kellermayer, M.S., Smith, S.B., Granzier, H.L. and Bustamante, C. (1997) Folding-unfolding transitions in single titin molecules characterized with laser tweezers. *Science*, **276** (5315), 1112–1116.

66 Rief, M., Gautel, M., Oesterhelt, F., Fernandez, J.M. and Gaub, H.E. (1997) Reversible unfolding of individual titin immunoglobulin domains by AFM. *Science*, **276** (5315), 1109–1112.

67 Tskhovrebova, L., Trinick, J., Sleep, J.A. and Simmons, R.M. (1997) Elasticity and unfolding of single molecules of the giant muscle protein titin. *Nature*, **387** (6630), 308–312.

68 Lu, H., Isralewitz, B., Krammer, A., Vogel, V. and Schulten, K. (1998) Unfolding of titin immunoglobulin domains by steered molecular dynamics simulation. *Biophysical Journal*, **75**, 662–671.

69 Baneyx, G., Baugh, L. and Vogel, V. (2001) Coexisting conformations of fibronectin in cell culture imaged using fluorescence resonance energy transfer. *Proceedings of the National Academy of Sciences of the United States of America*, **98** (25), 14464–14468.

70 Baneyx, G., Baugh, L. and Vogel, V. (2002) Fibronectin extension and unfolding within cell matrix fibrils controlled by cytoskeletal tension. *Proceedings of the National Academy of Sciences of the United States of America*, **99** (8), 5139–5143.

71 Sawada, Y. and Sheetz, M.P. (2002) Force transduction by Triton cytoskeletons. *The Journal of Cell Biology*, **156** (4), 609–615.

72 Clausen-Schaumann, H., Seitz, M., Krautbauer, R. and Gaub, H.E. (2000) Force spectroscopy with single bio-molecules. *Current Opinion in Chemical Biology*, **4** (5), 524–530.

73 Fredberg, J.J. and Kamm, R.D. (2006) Stress transmission in the lung: Pathways from organ to molecule. *Annual Review of Physiology*, **68**, 507–541.

74 Khan, S. and Sheetz, M.P. (1997) Force effects on biochemical kinetics. *Annual Review of Biochemistry*, **66**, 785–805.

75 Bershadsky, A.D., Balaban, N.Q. and Geiger, B. (2003) Adhesion-dependent cell mechanosensitivity. *Annual Review of Cell and Developmental Biology*, **19**, 677–695.

76 Silver, F.H. and Siperko, L.M. (2003) Mechanosensing and mechanochemical transduction: how is mechanical energy sensed and converted into chemical energy in an extracellular matrix? *Critical Reviews in Biomedical Engineering*, **31** (4), 255–331.

77 Martinac, B. (2004) Mechanosensitive ion channels: molecules of mechanotransduction. *Journal of Cell Science*, **117** (Pt 12), 2449–2460.

78 Kung, C. (2005) A possible unifying principle for mechanosensation. *Nature*, **436** (7051), 647–654.

79 Shemesh, T., Geiger, B., Bershadsky, A.D. and Kozlov, M.M. (2005) Focal adhesions as mechanosensors: A physical mechanism. *Proceedings of the National Academy of Sciences of the United States of America*, **102** (35), 12383–12388.

80 Hytönen, V.P., Smith, M.L. and Vogel, V. (2009) Translating mechanical force into discrete biochemical signal changes: multimodularity imposes unique properties to mechanotransductive proteins. *Mechanotransduction* (eds R. Kamm and M.R.K. Mofrad), Cambridge University Press, in press.

81 Ingber, D.E. (2005) Mechanical control of tissue growth: function follows form. *Proceedings of the National Academy of Sciences of the United States of America*, **102** (33), 11571–11572.

82 Wang, N. and Suo, Z. (2005) Long-distance propagation of forces in a cell. *Biochemical and Biophysical Research Communications*, **328** (4), 1133–1138.

83 Gittes, F., Meyhofer, E., Baek, S. and Howard, J. (1996) Directional loading of the kinesin motor molecule as it buckles a microtubule. *Biophysical Journal*, **70** (1), 418–429.

84 Block, S.M., Asbury, C.L., Shaevitz, J.W. and Lang, M.J. (2003) Probing the kinesin reaction cycle with a 2D optical force clamp. *Proceedings of the National Academy of Sciences of the United States of America*, **100** (5), 2351–2356.

85 Sukharev, S.I., Blount, P., Martinac, B., Blattner, F.R. and Kung, C. (1994) A large-conductance mechanosensitive channel in *E. coli* encoded by mscL alone. *Nature*, **368** (6468), 265–268.

86 Gullingsrud, J. and Schulten, K. (2004) Lipid bilayer pressure profiles and mechanosensitive channel gating. *Biophysical Journal*, **86** (6), 3496–3509.

87 Wu, X., Yang, Y., Gui, P., Sohma, Y., Meininger, G.A., Davis, G.E., Braun, A.P. and Davis, M.J. (2008) Potentiation of large conductance, Ca^{2+}-activated K^+ (BK) channels by alpha5beta1 integrin activation in arteriolar smooth muscle. *The Journal of Physiology*, **586** (6), 1699–1713.

88 Merkel, R., Nassoy, P., Leung, A., Ritchie, K. and Evans, E. (1999) Energy landscapes of receptor-ligand bonds explored with dynamic force spectroscopy. *Nature*, **397** (6714), 50–53.

89 Hess, H., Howard, J. and Vogel, V. (2002) A piconewton forcemeter assembled

from microtubules and kinesins. *Nano Letters*, **2** (10), 1113–1115.

90 Zhu, C. and McEver, R.P. (2005) Catch bonds: physical models and biological functions. *Molecular and Cellular Biomechanics*, **2** (3), 91–104.

91 Evans, E.A. and Calderwood, D.A. (2007) Forces and bond dynamics in cell adhesion. *Science*, **316** (5828), 1148–1153.

92 Sokurenko, E., Vogel, V. and Thomas, W.E. (2008) Catch bond mechanism of force-enhanced adhesion: counter-intuitive, elusive but . . . widespread? *Cell Host and Microbe*, **4**, 314–323.

93 Thomas, W.E., Vogel, V. and Sokurenko, E. (2008) Biophysics of catch bonds. *Annual Review of Biophysics*, **37**, 399–416.

94 Bianchi, L. (2007) Mechanotransduction: touch and feel at the molecular level as modeled in *Caenorhabditis elegans*. *Molecular Neurobiology*, **36**, 254–271.

95 Christensen, A.P. and Corey, D.P. (2007) TRP channels in mechanosensation: direct or indirect activation? *Nature Reviews in Neuroscience*, **8**, 510–521.

96 Kolomeisky, A.B. and Fisher, M.E. (2007) Molecular motors: a theorist's perspective. *Annual Review of Physical Chemistry*, **58**, 675–695.

97 Bartman, T., Walsh, E.C., Wen, K.K., McKane, M., Ren, J., Alexander, J., Rubenstein, P.A. and Stainier, D.Y. (2004) Early myocardial function affects endocardial cushion development in zebrafish. *PLoS Biology*, **2**, E129.

98 Vignjevic, D. and Montagnac, G. (2008) Reorganisation of the dendritic actin network during cancer cell migration and invasion. *Seminars in Cancer Biology*, **18** (1), 12–22.

99 Gupton, S.L. and Gertler, F.B. (2007) Filopodia: the fingers that do the walking. *Science's STKE: Signal Transduction Knowledge Environment*, **2007** (400), re5.

100 Medalia, O., Beck, M., Ecke, M., Weber, I., Neujahr, R., Baumeister, W. and Gerisch, G. (2007) Organization of actin networks in intact filopodia. *Current Biology*, **17** (1), 79–84.

101 Evans, E. (2001) Probing the relation between force--lifetime--and chemistry in single molecular bonds. *Annual Review of Biophysics and Biomolecular Structure*, **30**, 105–128.

102 Lehenkari, P.P. and Horton, M.A. (1999) Single integrin molecule adhesion forces in intact cells measured by atomic force microscopy. *Biochemical and Biophysical Research Communications*, **259** (3), 645–650.

103 Sun, Z., Martinez-Lemus, L.A., Trache, A., Trzeciakowski, J.P., Davis, G.E., Pohl, U. and Meininger, G.A. (2005) Mechanical properties of the interaction between fibronectin and {alpha}5{beta}1 integrins on vascular smooth muscle cells studied using atomic force microscopy. *American Journal of Physiology. Heart and Circulatory Physiology*, **289**, 2526–2535.

104 Hynes, R.O. (1990) *Fibronectins*, Springer-Verlag, New York, NY.

105 Pankov, R. and Yamada, K.M. (2002) Fibronectin at a glance. *Journal of Cell Science*, **115** (Pt 20), 3861–3863.

106 Mould, A.P. and Humphries, M.J. (2004) Regulation of integrin function through conformational complexity: not simply a knee-jerk reaction? *Current Opinion in Cell Biology*, **16** (5), 544–551.

107 Arnaout, M.A., Mahalingam, B. and Xiong, J.P. (2005) Integrin structure, allostery, and bidirectional signaling. *Annual Review of Cell and Developmental Biology*, **21**, 381–410.

108 Wegener, K.L., Partridge, A.W., Han, J., Pickford, A.R., Liddington, R.C., Ginsberg, M.H. and Campbell, I.D. (2007) Structural basis of integrin activation by talin. *Cell*, **128** (1), 171–182.

109 Mould, A.P., Barton, S.J., Askari, J.A., Craig, S.E. and Humphries, M.J. (2003) Role of ADMIDAS cation-binding site in ligand recognition by integrin alpha 5 beta 1. *The Journal of Biological Chemistry*, **278** (51), 51622–51629.

110 Adair, B.D., Xiong, J.P., Maddock, C., Goodman, S.L., Arnaout, M.A. and Yeager, M. (2005) Three-dimensional EM

structure of the ectodomain of integrin {alpha}V{beta}3 in a complex with fibronectin. *The Journal of Cell Biology*, **168** (7), 1109–1118.

111 Mould, A.P., Travis, M.A., Barton, S.J., Hamilton, J.A., Askari, J.A., Craig, S.E., Macdonald, P.R., Kammerer, R.A., Buckley, P.A. and Humphries, M.J. (2005) Evidence that monoclonal antibodies directed against the integrin beta subunit plexin/semaphorin/integrin domain stimulate function by inducing receptor extension. *The Journal of Biological Chemistry*, **280** (6), 4238–4246.

112 Puklin-Faucher, E., Gao, M., Schulten, K. and Vogel, V. (2006) How the headpiece hinge angle is opened: New insights into the dynamics of integrin activation. *The Journal of Cell Biology*, **175** (2), 349–360.

113 Calderwood, D.A. (2004) Talin controls integrin activation. *Biochemical Society Transactions*, **32** (Pt 3), 434–437.

114 Calderwood, D.A. and Ginsberg, M.H. (2003) Talin forges the links between integrins and actin. *Nature Cell Biology*, **5** (8), 694–697.

115 Jiang, G., Giannone, G., Critchley, D.R., Fukumoto, E. and Sheetz, M.P. (2003) Two-piconewton slip bond between fibronectin and the cytoskeleton depends on talin. *Nature*, **424** (6946), 334–337.

116 Calderwood, D.A., Zent, R., Grant, R., Rees, D.J., Hynes, R.O. and Ginsberg, M.H. (1999) The talin head domain binds to integrin beta subunit cytoplasmic tails and regulates integrin activation. *The Journal of Biological Chemistry*, **274** (40), 28071–28074.

117 Barsukov, I.L., Prescot, A., Bate, N., Patel, B., Floyd, D.N., Bhanji, N., Bagshaw, C.R., Letinic, K., Di Paolo, G., De Camilli, P., Roberts, G.C. and Critchley, D.R. (2003) Phosphatidylinositol phosphate kinase type 1gamma and beta1-integrin cytoplasmic domain bind to the same region in the talin FERM domain. *The Journal of Biological Chemistry*, **278** (33), 31202–31209.

118 Borowsky, M.L. and Hynes, R.O. (1998) Layilin, a novel talin-binding transmembrane protein homologous with C-type lectins, is localized in membrane ruffles. *The Journal of Cell Biology*, **143** (2), 429–442.

119 Lee, H.S., Bellin, R.M., Walker, D.L., Patel, B., Powers, P., Liu, H., Garcia-Alvarez, B., de Pereda, J.M., Liddington, R.C., Volkmann, N., Hanein, D., Critchley, D.R. and Robson, R.M. (2004) Characterization of an actin-binding site within the talin FERM domain. *Journal of Molecular Biology*, **343** (3), 771–784.

120 Molony, L., McCaslin, D., Abernethy, J., Paschal, B. and Burridge, K. (1987) Properties of talin from chicken gizzard smooth muscle. *The Journal of Biological Chemistry*, **262** (16), 7790–7795.

121 McLachlan, A.D., Stewart, M., Hynes, R.O. and Rees, D.J. (1994) Analysis of repeated motifs in the talin rod. *Journal of Molecular Biology*, **235** (4), 1278–1290.

122 Gingras, A.R., Ziegler, W.H., Frank, R., Barsukov, I.L., Roberts, G.C., Critchley, D.R. and Emsley, J. (2005) Mapping and consensus sequence identification for multiple vinculin binding sites within the talin rod. *The Journal of Biological Chemistry*, **280** (44), 37217–37224.

123 Hemmings, L., Rees, D.J., Ohanian, V., Bolton, S.J., Gilmore, A.P., Patel, B., Priddle, H., Trevithick, J.E., Hynes, R.O. and Critchley, D.R. (1996) Talin contains three actin-binding sites each of which is adjacent to a vinculin-binding site. *Journal of Cell Science*, **109**, 2715–2726.

124 Xing, B., Jedsadayanmata, A. and Lam, S.C. (2001) Localization of an integrin binding site to the C terminus of talin. *The Journal of Biological Chemistry*, **276** (48), 44373–44378.

125 Wiesner, S., Lange, A. and Fassler, R. (2006) Local call: from integrins to actin assembly. *Trends in Cell Biology*, **16** (7), 327–329.

126 Zaidel-Bar, R., Itzkovitz, S., Ma'ayan, A., Iyengar, R. and Geiger, B. (2007)

Functional atlas of the integrin adhesome. *Nature Cell Biology*, **9** (8), 858–867.

127 Zaidel-Bar, R., Milo, R., Kam, Z. and Geiger, B. (2007) A paxillin tyrosine phosphorylation switch regulates the assembly and form of cell-matrix adhesions. *Journal of Cell Science*, **120** (Pt 1), 137–148.

128 Moser, M., Nieswandt, B., Ussar, S., Pozgajova, M. and Fassler, R. (2008) Kindlin-3 is essential for integrin activation and platelet aggregation. *Nature Medicine*, **14** (3), 325–330.

129 Kiema, T., Lad, Y., Jiang, P., Oxley, C.L., Baldassarre, M., Wegener, K.L., Campbell, I.D., Ylanne, J. and Calderwood, D.A. (2006) The molecular basis of filamin binding to integrins and competition with talin. *Molecular Cell*, **21** (3), 337–347.

130 McCleverty, C.J., Lin, D.C. and Liddington, R.C. (2007) Structure of the PTB domain of tensin1 and a model for its recruitment to fibrillar adhesions. *Protein Science: A Publication of the Protein Society*, **16** (6), 1223–1229.

131 Litjens, S.H. and de Pereda, J.M. and Sonnenberg, A. (2006) Current insights into the formation and breakdown of hemidesmosomes. *Trends in Cell Biology*, **16** (7), 376–383.

132 Otey, C.A., Vasquez, G.B., Burridge, K. and Erickson, B.W. (1993) Mapping of the alpha-actinin binding site within the beta 1 integrin cytoplasmic domain. *The Journal of Biological Chemistry*, **268** (28), 21193–21197.

133 Greenwood, J.A., Theibert, A.B., Prestwich, G.D. and Murphy-Ullrich, J.E. (2000) Restructuring of focal adhesion plaques by PI 3-kinase. Regulation by PtdIns (3,4,5)-p(3) binding to alpha-actinin. *The Journal of Cell Biology*, **150** (3), 627–642.

134 Mostafavi-Pour, Z., Askari, J.A., Parkinson, S.J., Parker, P.J., Ng, T.T. and Humphries, M.J. (2003) Integrin-specific signaling pathways controlling focal adhesion formation and cell migration. *The Journal of Cell Biology*, **161** (1), 155–167.

135 Hu, S. and Wang, N. (2006) Control of stress propagation in the cytoplasm by prestress and loading frequency. *Molecular & Cellular Biomechanics*, **3** (2), 49–60.

136 Dai, J. and Sheetz, M.P. (1999) Membrane tether formation from blebbing cells. *Biophysical Journal*, **77** (6), 3363–3370.

137 Katz, B.Z., Miyamoto, S., Teramoto, H., Zohar, M., Krylov, D., Vinson, C., Gutkind, J.S. and Yamada, K.M. (2002) Direct transmembrane clustering and cytoplasmic dimerization of focal adhesion kinase initiates its tyrosine phosphorylation. *Biochimica et Biophysica Acta*, **1592** (2), 141–152.

138 Li, R., Bennett, J.S. and Degrado, W.F. (2004) Structural basis for integrin alphaII beta3 clustering. *Biochemical Society Transactions*, **32** (Pt 3), 412–415.

139 Cluzel, C., Saltel, F., Lussi, J., Paulhe, F., Imhof, B.A. and Wehrle-Haller, B. (2005) The mechanisms and dynamics of (alpha) v(beta)3 integrin clustering in living cells. *The Journal of Cell Biology*, **171** (2), 383–392.

140 Coussen, F., Choquet, D., Sheetz, M.P. and Erickson, H.P. (2002) Trimers of the fibronectin cell adhesion domain localize to actin filament bundles and undergo rearward translocation. *Journal of Cell Science*, **115** (Pt 12), 2581–2590.

141 Cavalcanti-Adam, E.A., Volberg, T., Micoulet, A., Kessler, H., Geiger, B. and Spatz, J.P. (2007) Cell spreading and focal adhesion dynamics are regulated by spacing of integrin ligands. *Biophysical Journal*, **92** (8), 2964–2974.

142 De Pasquale, J.A. and Izzard, C.S. (1991) Accumulation of talin in nodes at the edge of the lamellipodium and separate incorporation into adhesion plaques at focal contacts in fibroblasts. *The Journal of Cell Biology*, **113** (6), 1351–1359.

143 Gallant, N.D., Michael, K.E. and Garcia, A.J. (2005) Cell adhesion strengthening: contributions of adhesive area, integrin

binding, and focal adhesion assembly. *Molecular Biology of the Cell*, **16** (9), 4329–4340.

144 Chandrasekar, I., Stradal, T.E., Holt, M.R., Entschladen, F., Jockusch, B.M. and Ziegler, W.H. (2005) Vinculin acts as a sensor in lipid regulation of adhesion-site turnover. *Journal of Cell Science*, **118** (Pt 7), 1461–1472.

145 Paszek, M.J., Zahir, N., Johnson, K.R., Lakins, J.N., Rozenberg, G.I., Gefen, A., Reinhart-King, C.A., Margulies, S.S., Dembo, M., Boettiger, D., Hammer, D.A. and Weaver, V.M. (2005) Tensional homeostasis and the malignant phenotype. *Cancer Cell*, **8** (3), 241–254.

146 Ziegler, W.H., Liddington, R.C. and Critchley, D.R. (2006) The structure and regulation of vinculin. *Trends in Cell Biology*, **16** (9), 453–460.

147 Izard, T., Evans, G., Borgon, R.A., Rush, C.L., Bricogne, G. and Bois, P.R. (2004) Vinculin activation by talin through helical bundle conversion. *Nature*, **427** (6970), 171–175.

148 Papagrigoriou, E., Gingras, A.R., Barsukov, I.L., Bate, N., Fillingham, I.J., Patel, B., Frank, R., Ziegler, W.H., Roberts, G.C., Critchley, D.R. and Emsley, J. (2004) Activation of a vinculin-binding site in the talin rod involves rearrangement of a five-helix bundle. *The EMBO Journal*, **23** (15), 2942–2951.

149 Fillingham, I., Gingras, A.R., Papagrigoriou, E., Patel, B., Emsley, J., Critchley, D.R., Roberts, G.C. and Barsukov, I.L. (2005) A vinculin binding domain from the talin rod unfolds to form a complex with the vinculin head. *Structure (Camb.)*, **13** (1), 65–74.

150 Hytonen, V.P. and Vogel, V. (2008) How force might activate talin's vinculin binding sites: SMD reveals a structural mechanism. *PLoS Computational Biology*, **4** (2), e24.

151 Chen, H., Cohen, D.M., Choudhury, D.M., Kioka, N. and Craig, S.W. (2005) Spatial distribution and functional significance of activated vinculin in living cells. *The Journal of Cell Biology*, **169** (3), 459–470.

152 Bois, P.R., O'Hara, B.P., Nietlispach, D., Kirkpatrick, J. and Izard, T. (2006) The vinculin binding sites of talin and alpha-actinin are sufficient to activate vinculin. *The Journal of Biological Chemistry*, **281**, 7228–7236.

153 Bois, P.R., O'Hara, B.P., Nietlispach, D., Kirkpatrick, J. and Izard, T. (2006) The vinculin binding sites of talin and alpha-actinin are sufficient to activate vinculin. *The Journal of Biological Chemistry*, **281** (11), 7228–7236.

154 Johnson, R.P. and Craig, S.W. (1995) F-actin binding site masked by the intramolecular association of vinculin head and tail domains. *Nature*, **373** (6511), 261–264.

155 Chen, H., Choudhury, D.M. and Craig, S.W. (2006) Coincidence of actin filaments and talin is required to activate vinculin. *The Journal of Biological Chemistry*, **281** (52), 40389–40398.

156 Kelly, D.F., Taylor, D.W., Bakolitsa, C., Bobkov, A.A., Bankston, L., Liddington, R.C. and Taylor, K.A. (2006) Structure of the alpha-actinin-vinculin head domain complex determined by cryo-electron microscopy. *Journal of Molecular Biology*, **357** (2), 562–573.

157 Raucher, D. and Sheetz, M.P. (2000) Cell spreading and lamellipodial extension rate is regulated by membrane tension. *The Journal of Cell Biology*, **148** (1), 127–136.

158 Sawada, Y., Nakamura, K., Doi, K., Takeda, K., Tobiume, K., Saitoh, M., Morita, K., Komuro, I., De Vos, K., Sheetz, M. and Ichijo, H. (2001) Rap1 is involved in cell stretching modulation of p38 but not ERK or JNK MAP kinase. *Journal of Cell Science*, **114** (Pt 6), 1221–1227.

159 Kirchner, J., Kam, Z., Tzur, G., Bershadsky, A.D. and Geiger, B. (2003) Live-cell monitoring of tyrosine phosphorylation in focal adhesions

following microtubule disruption. *Journal of Cell Science*, **116** (Pt 6), 975–986.

160 Tamada, M., Sheetz, M.P. and Sawada, Y. (2004) Activation of a signaling cascade by cytoskeleton stretch. *Developmental Cell*, **7** (5), 709–718.

161 Ballestrem, C., Erez, N., Kirchner, J., Kam, Z., Bershadsky, A. and Geiger, B. (2006) Molecular mapping of tyrosine-phosphorylated proteins in focal adhesions using fluorescence resonance energy transfer. *Journal of Cell Science*, **119** (Pt 5), 866–875.

162 Bunnell, S.C., Hong, D.I., Kardon, J.R., Yamazaki, T., McGlade, C.J., Barr, V.A. and Samelson, L.E. (2002) T cell receptor ligation induces the formation of dynamically regulated signaling assemblies. *The Journal of Cell Biology*, **158** (7), 1263–1275.

163 Jin, Z.G., Ueba, H., Tanimoto, T., Lungu, A.O., Frame, M.D. and Berk, B.C. (2003) Ligand-independent activation of vascular endothelial growth factor receptor 2 by fluid shear stress regulates activation of endothelial nitric oxide synthase. *Circulation Research*, **93** (4), 354–363.

164 Shimizu, N., Yamamoto, K., Obi, S., Kumagaya, S., Masumura, T., Shimano, Y., Naruse, K., Yamashita, J.K., Igarashi, T. and Ando, J. (2008) Cyclic strain induces mouse embryonic stem cell differentiation into vascular smooth muscle cells by activating PDGF receptor beta. *Journal of Applied Physiology (Bethesda, Md: 1985)*, **104** (3), 766–772.

165 Fisher, T.E., Marszalek, P.E., Oberhauser, A.F., Carrion-Vazquez, M. and Fernandez, J.M. (1999) The micro-mechanics of single molecules studied with atomic force microscopy. *The Journal of Physiology*, **520** (Pt 1), 5–14.

166 Ridley, A.J., Schwartz, M.A., Burridge, K., Firtel, R.A., Ginsberg, M.H., Borisy, G., Parsons, J.T. and Horwitz, A.R. (2003) Cell migration: integrating signals from front to back. *Science*, **302** (5651), 1704–1709.

167 Zaidel-Bar, R., Cohen, M., Addadi, L. and Geiger, B. (2004) Hierarchical assembly of cell-matrix adhesion complexes. *Biochemical Society Transactions*, **32** (Pt 3), 416–420.

168 Dobereiner, H.G., Dubin-Thaler, B.J., Giannone, G. and Sheetz, M.P. (2005) Force sensing and generation in cell phases: analyses of complex functions. *Journal of Applied Physiology (Bethesda, Md: 1985)*, **98** (4), 1542–1546.

169 Bershadsky, A.D., Ballestrem, C., Carramusa, L., Zilberman, Y., Gilquin, B., Khochbin, S., Alexandrova, A.Y., Verkhovsky, A.B., Shemesh, T. and Kozlov, M.M. (2006) Assembly and mechanosensory function of focal adhesions: experiments and models. *European Journal of Cell Biology*, **85** (3–4), 165–173.

170 Lele, T.P., Thodeti, C.K. and Ingber, D.E. (2006) Force meets chemistry: analysis of mechanochemical conversion in focal adhesions using fluorescence recovery after photobleaching. *Journal of Cellular Biochemistry*, **97** (6), 1175–1183.

171 Miyamoto, S., Akiyama, S.K. and Yamada, K.M. (1995) Synergistic roles for receptor occupancy and aggregation in integrin transmembrane function. *Science*, **267** (5199), 883–885.

172 Humphries, J.D., Wang, P., Streuli, C., Geiger, B., Humphries, M.J. and Ballestrem, C. (2007) Vinculin controls focal adhesion formation by direct interactions with talin and actin. *The Journal of Cell Biology*, **179** (5), 1043–1057.

173 von Wichert, G., Haimovich, B., Feng, G.S. and Sheetz, M.P. (2003) Force-dependent integrin-cytoskeleton linkage formation requires downregulation of focal complex dynamics by Shp2. *The EMBO Journal*, **22** (19), 5023–5035.

174 von Wichert, G., Jiang, G., Kostic, A., De Vos, K., Sap, J. and Sheetz, M.P. (2003) RPTP-alpha acts as a transducer of mechanical force on alphav/beta3-integrin-cytoskeleton linkages. *The Journal of Cell Biology*, **161** (1), 143–153.

175 Critchley, D.R., Holt, M.R., Barry, S.T., Priddle, H., Hemmings, L. and Norman, J. (1999) Integrin-mediated cell adhesion: the cytoskeletal connection. *Biochemical Society Symposium*, **65**, 79–99.

176 Gu, J., Tamura, M., Pankov, R., Danen, E.H., Takino, T., Matsumoto, K. and Yamada, K.M. (1999) Shc and FAK differentially regulate cell motility and directionality modulated by PTEN. *The Journal of Cell Biology*, **146** (2), 389–403.

177 Sieg, D.J., Hauck, C.R. and Schlaepfer, D.D. (1999) Required role of focal adhesion kinase (FAK) for integrin-stimulated cell migration. *Journal of Cell Science*, **112** (Pt 16), 2677–2691.

178 Selhuber-Unkel, C., Lopez-Garcia, M., Kessler, H. and Spatz, J.P. (2008) Cooperativity in adhesion cluster formation during initial cell adhesion. *Biophysical Journal*.

179 Beningo, K.A., Dembo, M., Kaverina, I., Small, J.V. and Wang, Y.L. (2001) Nascent focal adhesions are responsible for the generation of strong propulsive forces in migrating fibroblasts. *The Journal of Cell Biology*, **153** (4), 881–888.

180 Hirata, H., Tatsumi, H. and Sokabe, M. (2008) Mechanical forces facilitate actin polymerization at focal adhesions in a zyxin-dependent manner. *Journal of Cell Science*.

181 Zaidel-Bar, R., Ballestrem, C., Kam, Z. and Geiger, B. (2003) Early molecular events in the assembly of matrix adhesions at the leading edge of migrating cells. *Journal of Cell Science*, **116** (Pt 22), 4605–4613.

182 Lele, T.P., Pendse, J., Kumar, S., Salanga, M., Karavitis, J. and Ingber, D.E. (2005) Mechanical forces alter zyxin unbinding kinetics within focal adhesions of living cells. *Journal of Cellular Physiology*.

183 Zamir, E., Katz, B.Z., Aota, S., Yamada, K.M., Geiger, B. and Kam, Z. (1999) Molecular diversity of cell-matrix adhesions. *Journal of Cell Science*, **112** (Pt 11), 1655–1669.

184 Katz, B.Z., Zamir, E., Bershadsky, A., Kam, Z., Yamada, K.M. and Geiger, B. (2000) Physical state of the extracellular matrix regulates the structure and molecular composition of cell-matrix adhesions. *Molecular Biology of the Cell*, **11** (3), 1047–1060.

185 Wu, C., Keivens, V.M., O'Toole, T.E., McDonald, J.A. and Ginsberg, M.H. (1995) Integrin activation and cytoskeletal interaction are essential for the assembly of a fibronectin matrix. *Cell*, **83** (5), 715–724.

186 Cali, G., Mazzarella, C., Chiacchio, M., Negri, R., Retta, S.F., Zannini, M., Gentile, F., Tarone, G., Nitsch, L. and Garbi, C. (1999) RhoA activity is required for fibronectin assembly and counteracts beta1B integrin inhibitory effect in FRT epithelial cells. *Journal of Cell Science*, **112** (Pt 6), 957–965.

187 Pankov, R., Cukierman, E., Katz, B.Z., Matsumoto, K., Lin, D.C., Lin, S., Hahn, C. and Yamada, K.M. (2000) Integrin dynamics and matrix assembly: tensin-dependent translocation of alpha(5)beta (1) integrins promotes early fibronectin fibrillogenesis. *The Journal of Cell Biology*, **148** (5), 1075–1090.

188 Chen, B.H., Tzen, J.T., Bresnick, A.R. and Chen, H.C. (2002) Roles of Rho-associated kinase and myosin light chain kinase in morphological and migratory defects of focal adhesion kinase-null cells. *The Journal of Biological Chemistry*, **277** (37), 33857–33863.

189 Burridge, K. and Wennerberg, K. (2004) Rho and Rac take center stage. *Cell*, **116** (2), 167–179.

190 McBeath, R., Pirone, D.M., Nelson, C.M., Bhadriraju, K. and Chen, C.S. (2004) Cell shape, cytoskeletal tension, and RhoA regulate stem cell lineage commitment. *Developmental Cell*, **6** (4), 483–495.

191 Yoneda, A., Multhaupt, H.A. and Couchman, J.R. (2005) The Rho kinases I and II regulate different aspects of myosin II activity. *The Journal of Cell Biology*, **170** (3), 443–453.

192 Barry, S.T., Flinn, H.M., Humphries, M.J., Critchley, D.R. and Ridley, A.J. (1997) Requirement for Rho in integrin signalling. *Cell Adhesion and Communication*, **4** (6), 387–398.

193 Zhong, C., Chrzanowska-Wodnicka, M., Brown, J., Shaub, A., Belkin, A.M. and Burridge, K. (1998) Rho-mediated contractility exposes a cryptic site in fibronectin and induces fibronectin matrix assembly. *The Journal of Cell Biology*, **141** (2), 539–551.

194 Danen, E.H., Sonneveld, P., Brakebusch, C., Fassler, R. and Sonnenberg, A. (2002) The fibronectin-binding integrins alpha5beta1 and alphavbeta3 differentially modulate RhoA-GTP loading, organization of cell matrix adhesions, and fibronectin fibrillogenesis. *The Journal of Cell Biology*, **159** (6), 1071–1086.

195 Miao, H., Li, S., Hu, Y.L., Yuan, S., Zhao, Y., Chen, B.P., Puzon-McLaughlin, W., Tarui, T., Shyy, J.Y., Takada, Y., Usami, S. and Chien, S. (2002) Differential regulation of Rho GTPases by beta1 and beta3 integrins: the role of an extracellular domain of integrin in intracellular signaling. *Journal of Cell Science*, **115** (Pt 10), 2199–2206.

196 Giannone, G. and Sheetz, M.P. (2006) Substrate rigidity and force define form through tyrosine phosphatase and kinase pathways. *Trends in Cell Biology*, **16** (4), 213–223.

197 Kostic, A. and Sheetz, M.P. (2006) Fibronectin rigidity response through Fyn and p130Cas recruitment to the leading edge. *Molecular Biology of the Cell*, **17** (6), 2684–2695.

198 Kostic, A., Sap, J. and Sheetz, M.P. (2007) RPTPalpha is required for rigidity-dependent inhibition of extension and differentiation of hippocampal neurons. *Journal of Cell Science*, **120** (Pt 21), 3895–3904.

199 Geiger, R.C., Taylor, W., Glucksberg, M.R. and Dean, D.A. (2006) Cyclic stretch-induced reorganization of the cytoskeleton and its role in enhanced gene transfer. *Gene Therapy*, **13** (8), 725–731.

200 Filipenko, N.R., Attwell, S., Roskelley, C. and Dedhar, S. (2005) Integrin-linked kinase activity regulates Rac- and Cdc42-mediated actin cytoskeleton reorganization via alpha-PIX. *Oncogene*, **24**, 5837–5849.

201 Legate, K.R., Montanez, E., Kudlacek, O. and Fassler, R. (2006) ILK, PINCH and parvin: the tIPP of integrin signalling. *Nature Reviews - Molecular Cell Biology*, **7** (1), 20–31.

202 Sakai, T., Li, S., Docheva, D., Grashoff, C., Sakai, K., Kostka, G., Braun, A., Pfeifer, A., Yurchenco, P.D. and Fassler, R. (2003) Integrin-linked kinase (ILK) is required for polarizing the epiblast, cell adhesion, and controlling actin accumulation. *Genes and Development*, **17** (7), 926–940.

203 Sheetz, M.P., Felsenfeld, D., Galbraith, C.G. and Choquet, D. (1999) Cell migration as a five-step cycle. *Biochemical Society Symposium*, **65**, 233–243.

204 Wehrle-Haller, B. and Imhof, B. (2002) The inner lives of focal adhesions. *Trends in Cell Biology*, **12** (8), 382–389.

205 Kirfel, G., Rigort, A., Borm, B. and Herzog, V. (2004) Cell migration: mechanisms of rear detachment and the formation of migration tracks. *European Journal of Cell Biology*, **83** (11–12), 717–724.

206 Galbraith, C.G. and Sheetz, M.P. (1997) A micromachined device provides a new bend on fibroblast traction forces. *Proceedings of the National Academy of Sciences of the United States of America*, **94** (17), 9114–9118.

207 Pelham, R.J. Jr and Wang, Y. (1997) Cell locomotion and focal adhesions are regulated by substrate flexibility. *Proceedings of the National Academy of Sciences of the United States of America*, **94** (25), 13661–13665.

208 Sterba, R.E. and Sheetz, M.P. (1998) Basic laser tweezers. *Methods in Cell Biology*, **55**, 29–41.

209 Beningo, K.A. and Wang, Y.L. (2002) Flexible substrata for the detection of cellular traction forces. *Trends in Cell Biology*, **12** (2), 79–84.

210 LeDuc, P., Ostuni, E., Whitesides, G. and Ingber, D. (2002) Use of micropatterned adhesive surfaces for control of cell behaviour. *Methods in Cell Biology*, **69**, 385–401.

211 Prechtel, K., Bausch, A.R., Marchi-Artzner, V., Kantlehner, M., Kessler, H. and Merkel, R. (2002) Dynamic force spectroscopy to probe adhesion strength of living cells. *Physical Review Letters*, **89** (2), 028101.

212 du Roure, O., Saez, A., Buguin, A., Austin, R.H., Chavrier, P., Siberzan, P. and Ladoux, B. (2005) Force mapping in epithelial cell migration. *Proceedings of the National Academy of Sciences of the United States of America*, **102** (7), 2390–2395.

213 Choquet, D., Felsenfeld, D.P. and Sheetz, M.P. (1997) Extracellular matrix rigidity causes strengthening of integrin-cytoskeleton linkages. *Cell*, **88** (1), 39–48.

214 Geiger, P.C., Cody, M.J., Macken, R.L., Bayrd, M.E., Fang, Y.H. and Sieck, G.C. (2001) Mechanisms underlying increased force generation by rat diaphragm muscle fibers during development. *Journal of Applied Physiology (Bethesda, Md: 1985)*, **90** (1), 380–388.

215 Tseng, Y., Kole, T.P. and Wirtz, D. (2002) Micromechanical mapping of live cells by multiple-particle-tracking microrheology. *Biophysical Journal*, **83** (6), 3162–3176.

216 Mack, P.J., Kaazempur-Mofrad, M.R., Karcher, H., Lee, R.T. and Kamm, R.D. (2004) Force-induced focal adhesion translocation: effects of force amplitude and frequency. *American Journal of Physiology. Cell Physiology*, **287** (4), C954–C962.

217 Ballestrem, C. and Geiger, B. (2005) Application of microscope-based FRET to study molecular interactions in focal adhesions of live cells. *Methods in Molecular Biology (Clifton, NJ)*, **294**, 321–334.

218 Katsumi, A., Naoe, T., Matsushita, T., Kaibuchi, K. and Schwartz, M.A. (2005) Integrin activation and matrix binding mediate cellular responses to mechanical stretch. *The Journal of Biological Chemistry*, **280** (17), 16546–16549.

219 Vallotton, P., Danuser, G., Bohnet, S., Meister, J.J. and Verkhovsky, A.B. (2005) Tracking retrograde flow in keratocytes: news from the front. *Molecular Biology of the Cell*, **16** (3), 1223–1231.

220 Sniadecki, N.J., Desai, R.A., Ruiz, S.A. and Chen, C.S. (2006) Nanotechnology for cell-substrate interactions. *Annals of Biomedical Engineering*, **34** (1), 59–74.

221 Sniadecki, N.J., Anguelouch, A., Yang, M.T., Lamb, C.M., Liu, Z., Kirschner, S.B., Liu, Y., Reich, D.H. and Chen, C.S. (2007) Magnetic microposts as an approach to apply forces to living cells. *Proceedings of the National Academy of Sciences of the United States of America*, **104** (37), 14553–14558.

222 Cai, Y., Biais, N., Giannone, G., Tanase, M., Jiang, G., Hofman, J.M., Wiggins, C.H., Silberzan, P., Buguin, A., Ladoux, B. and Sheetz, M.P. (2006) Nonmuscle myosin IIA-dependent force inhibits cell spreading and drives F-actin flow. *Biophysical Journal*, **91** (10), 3907–3920.

223 Sheetz, M.P. (2001) Cell control by membrane-cytoskeleton adhesion. *Nature Reviews - Molecular Cell Biology*, **2** (5), 392–396.

224 Wang, N., Butler, J.P. and Ingber, D.E. (1993) Mechanotransduction across the cell surface and through the cytoskeleton. *Science*, **260** (5111), 1124–1127.

225 Ganz, A., Lambert, M., Saez, A., Silberzan, P., Buguin, A., Mege, R.M. and Ladoux, B. (2006) Traction forces exerted through N-cadherin contacts. *Biology of the Cell/Under the Auspices of the European Cell Biology Organization*, **98** (12), 721–730.

226 Ehrlich, J.S., Hansen, M.D. and Nelson, W.J. (2002) Spatio-temporal regulation of Rac1 localization and lamellipodia dynamics during epithelial cell-cell

adhesion. *Developmental Cell*, **3** (2), 259–270.

227 Davidson, L.A., Marsden, M., Keller, R. and Desimone, D.W. (2006) Integrin alpha5beta1 and fibronectin regulate polarized cell protrusions required for *Xenopus* convergence and extension. *Current Biology*, **16** (9), 833–844.

228 Swartz, M.A., Tschumperlin, D.J., Kamm, R.D. and Drazen, J.M. (2001) Mechanical stress is communicated between different cell types to elicit matrix remodelling. *Proceedings of the National Academy of Sciences of the United States of America*, **98** (11), 6180–6185.

229 Giannone, G., Dubin-Thaler, B.J., Dobereiner, H.G., Kieffer, N., Bresnick, A.R. and Sheetz, M.P. (2004) Periodic lamellipodial contractions correlate with rearward actin waves. *Cell*, **116** (3), 431–443.

230 Lo, C.M., Wang, H.B., Dembo, M. and Wang, Y.L. (2000) Cell movement is guided by the rigidity of the substrate. *Biophysical Journal*, **79** (1), 144–152.

231 Saez, A., Buguin, A., Silberzan, P. and Ladoux, B. (2005) Is the mechanical activity of epithelial cells controlled by deformations or forces? *Biophysical Journal*, **89** (6), L52–L54.

232 Saez, A., Ghibaudo, M., Buguin, A., Silberzan, P. and Ladoux, B. (2007) Rigidity-driven growth and migration of epithelial cells on microstructured anisotropic substrates. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 8281–8286.

233 Jiang, G., Huang, A.H., Cai, Y., Tanase, M. and Sheetz, M.P. (2006) Rigidity sensing at the leading edge through alphavbeta3 integrins and RPTPalpha. *Biophysical Journal*, **90** (5), 1804–1809.

234 Guo, Y., Hsu, S., Sawhney, H.S., Kumar, R. and Shan, Y. (2007) Robust object matching for persistent tracking with heterogeneous features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **29** (5), 824–839.

235 Webb, D.J., Donais, K., Whitmore, L.A., Thomas, S.M., Turner, C.E., Parsons, J.T. and Horwitz, A.F. (2004) FAK-Src signalling through paxillin, ERK and MLCK regulates adhesion disassembly. *Nature Cell Biology*, **6** (2), 154–161.

236 Laukaitis, C.M., Webb, D.J., Donais, K. and Horwitz, A.F. (2001) Differential dynamics of alpha 5 integrin, paxillin, and alpha-actinin during formation and disassembly of adhesions in migrating cells. *The Journal of Cell Biology*, **153** (7), 1427–1440.

237 Bausch, A.R., Ziemann, F., Boulbitch, A.A., Jacobson, K. and Sackmann, E. (1998) Local measurements of viscoelastic parameters of adherent cell surfaces by magnetic bead microrheometry. *Biophysical Journal*, **75** (4), 2038–2049.

238 Kiosses, W.B., Shattil, S.J., Pampori, N. and Schwartz, M.A. (2001) Rac recruits high-affinity integrin alphavbeta3 to lamellipodia in endothelial cell migration. *Nature Cell Biology*, **3** (3), 316–320.

239 Arthur, W.T., Quilliam, L.A. and Cooper, J.A. (2004) Rap1 promotes cell spreading by localizing Rac guanine nucleotide exchange factors. *The Journal of Cell Biology*, **167** (1), 111–122.

240 Di Stefano, P., Cabodi, S., Boeri Erba, E., Margaria, V., Bergatto, E., Giuffrida, M.G., Silengo, L., Tarone, G., Turco, E. and Defilippi, P. (2004) P130Cas-associated protein (p140Cap) as a new tyrosine-phosphorylated protein involved in cell spreading. *Molecular Biology of the Cell*, **15** (2), 787–800.

241 Nishizaka, T., Shi, Q. and Sheetz, M.P. (2000) Position-dependent linkages of fibronectin-integrin-cytoskeleton. *Proceedings of the National Academy of Sciences of the United States of America*, **97** (2), 692–697.

242 Oganov, V.S. (2004) Modern analysis of bone loss mechanisms in microgravity. *The Journal of Gravitational Physiology*, **11** (2), P143–P150.

243 Dallas, S.L., Chen, Q. and Sivakumar, P. (2006) Dynamics of assembly and

reorganization of extracellular matrix proteins. *Current Topics in Developmental Biology*, **75**, 1–24.

244 Meshel, A.S., Wei, Q., Adelstein, R.S. and Sheetz, M.P. (2005) Basic mechanism of three-dimensional collagen fibre transport by fibroblasts. *Nature Cell Biology*, **7** (2), 157–164.

245 Giannone, G., Dubin-Thaler, B.J., Rossier, O. *et al*. (2007) Lamellipodial actin mechanically links myosin activity with adhesion-site formation. *Cell*, **128**, 561–575.

246 Astrof, S., Crowley, D., George, E.L., Fukuda, T., Sekiguchi, K., Hanahan, D. and Hynes, R.O. (2004) Direct test of potential roles of EIIIA and EIIIB alternatively spliced segments of fibronectin in physiological and tumor angiogenesis. *Molecular and Cellular Biology*, **24** (19), 8662–8670.

247 Brown, R.A., Blunn, G.W. and Ejim, O.S. (1994) Preparation of orientated fibrous mats from fibronectin: composition and stability. *Biomaterials*, **15** (6), 457–464.

248 Baneyx, G. and Vogel, V. (1999) Self-assembly of fibronectin into fibrillar networks underneath dipalmitoyl phosphatidylcholine monolayers: role of lipid matrix and tensile forces. *Proceedings of the National Academy of Sciences of the United States of America*, **96** (22), 12518–12523.

249 Takahashi, S., Leiss, M., Moser, M., Ohashi, T., Kitao, T., Heckmann, D., Pfeifer, A., Kessler, H., Takagi, J., Erickson, H.P. and Fassler, R. (2007) The RGD motif in fibronectin is essential for development but dispensable for fibril assembly. *The Journal of Cell Biology*, **178** (1), 167–178.

250 Little, W.C., Smith, M.L., Ebneter, U. and Vogel, V. (2008) Assay to mechanically tune and optically probe fibrillar fibronectin conformations from fully relaxed to breakage. *Matrix Biology*, **27**, 451–465.

251 Klotzsch, E., Smith, M.L., Kubow, K.E., Muntwyler, S., Little, W.C., Beyeler, F., Gourdon, D., Nelson, B.J. and Vogel, V. (2009) Fibronectin self-assembles when stretched into the most elastic biological fibers diplaying force-regulated molecular recognition switches. submitted.

252 Wierzbicka-Patynowski, I., Mao, Y. and Schwarzbauer, J.E. (2007) Continuous requirement for pp60-Src and phosphopaxillin during fibronectin matrix assembly by transformed cells. *Journal of Cellular Physiology*, **210** (3), 750–756.

253 Paci, E. and Karplus, M. (1999) Forced unfolding of fibronectin type 3 modules: an analysis by biased molecular dynamics simulations. *Journal of Molecular Biology*, **288** (3), 441–459.

254 Craig, D., Krammer, A., Schulten, K. and Vogel, V. (2001) Comparison of the early stages of forced unfolding for fibronectin type III modules. *Proceedings of the National Academy of Sciences of the United States of America*, **98** (10), 5590–5595.

255 Oberhauser, A.F., Badilla-Fernandez, C., Carrion-Vazquez, M. and Fernandez, J.M. (2002) The mechanical hierarchies of fibronectin observed with single-molecule AFM. *Journal of Molecular Biology*, **319** (2), 433–447.

256 Craig, D., Gao, M., Schulten, K. and Vogel, V. (2004) Tuning the mechanical stability of fibronectin type III modules through sequence variations. *Structure (Camb.)*, **12** (1), 21–30.

257 Ng, S.P., Rounsevell, R.W., Steward, A., Geierhaas, C.D., Williams, P.M., Paci, E. and Clarke, J. (2005) Mechanical unfolding of TNfn3: the unfolding pathway of a fnIII domain probed by protein engineering, AFM and MD simulation. *Journal of Molecular Biology*, **350** (4), 776–789.

258 Krammer, A., Lu, H., Isralewitz, B., Schulten, K. and Vogel, V. (1999) Forced unfolding of the fibronectin type III module reveals a tensile molecular recognition switch. *Proceedings of the National Academy of Sciences of the United States of America*, **96** (4), 1351–1356.

259 Marszalek, P.E., Lu, H., Li, H., Carrion-Vazquez, M., Oberhauser, A.F., Schulten,

K. and Fernandez, J.M. (1999) Mechanical unfolding intermediates in titin modules. *Nature*, **402** (6757), 100–103.

260 Gao, M., Craig, D., Vogel, V. and Schulten, K. (2002) Identifying unfolding intermediates of FN-III(10) by steered molecular dynamics. *Journal of Molecular Biology*, **323** (5), 939–950.

261 Krammer, A., Craig, D., Thomas, W.E., Schulten, K. and Vogel, V. (2002) A structural model for force regulated integrin binding to fibronectin's RGD-synergy site. *Matrix Biology*, **21** (2), 139–147.

262 Gao, M., Craig, D., Lequin, O., Campbell, I.D., Vogel, V. and Schulten, K. (2003) Structure and functional significance of mechanically unfolded fibronectin type III1 intermediates. *Proceedings of the National Academy of Sciences of the United States of America*, **100** (25), 14784–14789.

263 Andresen, M., Wahl, M.C., Stiel, A.C., Grater, F., Schafer, L.V., Trowitzsch, S., Weber, G., Eggeling, C., Grubmuller, H., Hell, S.W. and Jakobs, S. (2005) Structure and mechanism of the reversible photoswitch of a fluorescent protein. *Proceedings of the National Academy of Sciences of the United States of America*, **102** (37), 13070–13074.

264 Grater, F., Shen, J., Jiang, H., Gautel, M. and Grubmuller, H. (2005) Mechanically induced titin kinase activation studied by force-probe molecular dynamics simulations. *Biophysical Journal*, **88** (2), 790–804.

265 Lee, E.H., Hsin, J., Mayans, O. and Schulten, K. (2007) Secondary and tertiary structure elasticity of titin Z1Z2 and a titin chain model. *Biophysical Journal*, **93** (5), 1719–1735.

266 Ejim, O.S., Blunn, G.W. and Brown, R.A. (1993) Production of artificial-orientated mats and strands from plasma fibronectin: a morphological study. *Biomaterials*, **14** (10), 743–748.

267 Wojciak-Stothard, B., Denyer, M., Mishra, M. and Brown, R.A. (1997) Adhesion, orientation, and movement of cells cultured on ultrathin fibronectin fibers. *In Vitro Cellular & Developmental Biology - Animal*, **33** (2), 110–117.

268 Ahmed, Z. and Brown, R.A. (1999) Adhesion, alignment, and migration of cultured Schwann cells on ultrathin fibronectin fibres. *Cell Motility and the Cytoskeleton*, **42** (4), 331–343.

269 Antia, M., Baneyx, G., Kubow, K.E. and Vogel, V. (2008) Fibronectin in aging extracellular matrix fibrils is progressively unfolded by cells and elicits an enhanced rigidity response. *Faraday Discsussions*, **139**, 229–249.

270 Wolf, K. and Friedl, P. (2005) Functional imaging of pericellular proteolysis in cancer cell invasion. *Biochimie*, **87** (3–4), 315–320.

271 Zaman, M.H., Trapani, L.M., Siemeski, A., Mackellar, D., Gong, H., Kamm, R.D., Wells, A., Lauffenburger, D.A. and Matsudaira, P. (2006) Migration of tumor cells in 3D matrices is governed by matrix stiffness along with cell-matrix adhesion and proteolysis. *Proceedings of the National Academy of Sciences of the United States of America*, **103** (29), 10889–10894.

272 Wolf, K., Wu, Y.I., Liu, Y., Geiger, J., Tam, E., Overall, C., Stack, M.S. and Friedl, P. (2007) Multi-step pericellular proteolysis controls the transition from individual to collective cancer cell invasion. *Nature Cell Biology*, **9** (8), 893–904.

273 Janmey, P.A. and Weitz, D.A. (2004) Dealing with mechanics: mechanisms of force transduction in cells. *Trends in Biochemical Sciences*, **29** (7), 364–370.

274 Felsenfeld, D.P., Schwartzberg, P.L., Venegas, A., Tse, R. and Sheetz, M.P. (1999) Selective regulation of integrin--cytoskeleton interactions by the tyrosine kinase Src. *Nature Cell Biology*, **1** (4), 200–206.

275 Volberg, T., Romer, L., Zamir, E. and Geiger, B. (2001) pp60(c-src) and related tyrosine kinases: a role in the assembly and reorganization of matrix adhesions.

Journal of Cell Science, **114** (Pt 12), 2279–2289.

276 Wang, H.B., Dembo, M., Hanks, S.K. and Wang, Y. (2001) Focal adhesion kinase is involved in mechanosensing during fibroblast migration. *Proceedings of the National Academy of Sciences of the United States of America*, **98** (20), 11295–11300.

277 Munevar, S., Wang, Y.L. and Dembo, M. (2001) Distinct roles of frontal and rear cell-substrate adhesions in fibroblast migration. *Molecular Biology of the Cell*, **12** (12), 3947–3954.

278 Perrin, B.J. and Huttenlocher, A. (2002) Calpain. *The International Journal of Biochemistry and Cell Biology*, **34** (7), 722–725.

279 Kole, T.P., Tseng, Y., Jiang, I., Katz, J.L. and Wirtz, D. (2005) Intracellular mechanics of migrating fibroblasts. *Molecular Biology of the Cell*, **16** (1), 328–338.

280 Pankov, R., Endo, Y., Even-Ram, S., Araki, M., Clark, K., Cukierman, E., Matsumoto, K. and Yamada, K.M. (2005) A Rac switch regulates random versus directionally persistent cell migration. *The Journal of Cell Biology*, **170** (5), 793–802.

281 Frame, M.C. and Brunton, V.G. (2002) Advances in Rho-dependent actin regulation and oncogenic transformation. *Current Opinion in Genetics and Development*, **12** (1), 36–43.

282 Ghosh, M., Song, X., Mouneimne, G., Sidani, M., Lawrence, D.S. and Condeelis, J.S. (2004) Cofilin promotes actin polymerization and defines the direction of cell motility. *Science*, **304** (5671), 743–746.

283 Danen, E.H., van Rheenen, J., Franken, W., Huveneers, S., Sonneveld, P., Jalink, K. and Sonnenberg, A. (2005) Integrins control motile strategy through a Rho-cofilin pathway. *The Journal of Cell Biology*, **169** (3), 515–526.

284 Mouneimne, G., DesMarais, V., Sidani, M., Scemes, E., Wang, W., Song, X., Eddy, R. and Condeelis, J. (2006) Spatial and temporal control of cofilin activity is required for directional sensing during chemotaxis. *Current Biology*, **16** (22), 2193–2205.

285 Sidani, M., Wessels, D., Mouneimne, G., Ghosh, M., Goswami, S., Sarmiento, C., Wang, W., Kuhl, S., El-Sibai, M., Backer, J.M., Eddy, R., Soll, D. and Condeelis, J. (2007) Cofilin determines the migration behavior and turning frequency of metastatic cancer cells. *The Journal of Cell Biology*, **179** (4), 777–791.

286 Levinson, A.D., Oppermann, H., Levintow, L., Varmus, H.E. and Bishop, J.M. (1978) Evidence that the transforming gene of avian sarcoma virus encodes a protein kinase associated with a phosphoprotein. *Cell*, **15** (2), 561–572.

287 Parker, R.C., Varmus, H.E. and Bishop, J.M. (1984) Expression of v-src and chicken c-src in rat cells demonstrates qualitative differences between pp60v-src and pp60c-src. *Cell*, **37** (1), 131–139.

288 Guck, J., Schinkinger, S., Lincoln, B., Wottawah, F., Ebert, S., Romeyke, M., Lenz, D., Erickson, H.M., Ananthakrishnan, R., Mitchell, D., Kas, J., Ulvick, S. and Bilby, C. (2005) Optical deformability as an inherent cell marker for testing malignant transformation and metastatic competence. *Biophysical Journal*, **88** (5), 3689–3698.

289 Ruoslahti, E. (1999) Fibronectin and its integrin receptors in cancer. *Advances in Cancer Research*, **76**, 1–20.

290 Gimona, M. (2008) The microfilament system in the formation of invasive adhesions. *Seminars in Cancer Biology*, **18** (1), 23–34.

291 Mierke, C.T., Rosel, D., Fabry, B. and Brabek, J. (2008) Contractile forces in tumor cell migration. *European Journal of Cell Biology*, **87**, 669–676.

292 Huang, S. and Ingber, D.E. (2005) Cell tension, matrix mechanics, and cancer development. *Cancer Cell*, **8** (3), 175–176.

293 Paszek, M.J. and Weaver, V.M. (2004) The tension mounts: mechanics meets morphogenesis and malignancy. *Journal*

of Mammary Gland Biology and Neoplasia, **9** (4), 325–342.

294 Paget, S. (1889) The distribution of secondary growths in cancer of the breast. *Cancer Metastasis Reviews*, **8** (2), 98–102.

295 Beacham, D.A. and Cukierman, E. (2005) Stromagenesis: the changing face of fibroblastic microenvironments during tumor progression. *Seminars in Cancer Biology*, **15** (5), 329–341.

296 Mitra, S.K. and Schlaepfer, D.D. (2006) Integrin-regulated FAK-Src signaling in normal and cancer cells. *Current Opinion in Cell Biology*, **18** (5), 516–523.

297 Dorssers, L.C., Grebenchtchikov, N., Brinkman, A. *et al.* (2004) The prognostic value of BCAR1 in patients with primary breast cancer. *Clinical Cancer Research*, **10**, 6194–6202.

298 Riggins, R.B., De Berry, R.M., Toosarvandani, M.D. and Bouton, A.H. (2003) Src-dependent association of Cas and p85 phosphatidylinositol 3′-kinase in v-crk-transformed cells. *Molecular Cancer Research*, **1**, 428–437.

299 Nakamoto, T., Sakai, R., Honda, H., Ogawa, S., Ueno, H., Suzuki, T., Aizawa, S., Yazaki, Y. and Hirai, H. (1997) Requirements for localization of p130cas to focal adhesions. *Molecular and Cellular Biology*, **17** (7), 3884–3897.

300 Nievers, M.G., Birge, R.B., Greulich, H., Verkleij, A.J., Hanafusa, H., van Bergen en Henegouwen, P. M. (1997) v-Crk-induced cell transformation: changes in focal adhesion composition and signalling. *Journal of Cell Science*, **110** (Pt 3), 389–399.

301 Sakai, R., Nakamoto, T., Ozawa, K., Aizawa, S. and Hirai, H. (1997) Characterization of the kinase activity essential for tyrosine phosphorylation of p130Cas in fibroblasts. *Oncogene*, **14** (12), 1419–1426.

302 Kirsch, K., Kensinger, M., Hanafusa, H. and August, A. (2002) A p130Cas tyrosine phosphorylated substrate domain decoy disrupts v-crk signalling. *BMC Cell Biology*, **3**, 18.

303 Gotoh, T., Cai, D., Tian, X., Feig, L.A. and Lerner, A. (2000) p130Cas regulates the activity of AND-34, a novel Ral, Rap1, and R-Ras guanine nucleotide exchange factor. *Journal of Biological Chemistry*, **275**, 30118.

304 Brabek, J., Constancio, S.S., Siesser, P.F., Shin, N.Y., Pozzi, A. and Hanks, S.K. (2005) Crk-associated substrate tyrosine phosphorylation sites are critical for invasion and metastasis of SRC-transformed cells. *Molecular Cancer Research*, **3**, 307–315.

305 Ingber, D.E. and Folkman, J. (1989) Mechanochemical switching between growth and differentiation during fibroblast growth factor-stimulated angiogenesis in vitro: role of extracellular matrix. *The Journal of Cell Biology*, **109** (1), 317–330.

306 Larsen, M., Wei, C. and Yamada, K.M. (2006) Cell and fibronectin dynamics during branching morphogenesis. *Journal of Cell Science*, **119** (Pt 16), 3376–3384.

307 Heil, M. and Schaper, W. (2007) Insights into pathways of arteriogenesis. *Current Pharmaceutical Biotechnology*, **8** (1), 35–42.

308 Mammoto, A., Mammoto, T. and Ingber, D.E. (2008) Rho signaling and mechanical control of vascular development. *Current Opinion in Hematology*, **15** (3), 228–234.

309 Furuya, M. and Yonemitsu, Y. (2008) Cancer neovascularization and proinflammatory microenvironments. *Current Cancer Drug Targets*, **8** (4), 253–265.

310 Mahabeleshwar, G.H., Feng, W., Phillips, D.R. and Byzova, T.V. (2006) Integrin signaling is critical for pathological angiogenesis. *The Journal of Experimental Medicine*, **203** (11), 2495–2507.

311 Wijelath, E.S., Murray, J., Rahman, S., Patel, Y., Ishida, A., Strand, K., Aziz, S., Cardona, C., Hammond, W.P., Savidge, G.F., Rafii, S. and Sobel, M. (2002) Novel vascular endothelial growth factor binding domains of fibronectin enhance

vascular endothelial growth factor biological activity. *Circulation Research*, **91** (1), 25–31.

312 Wijelath, E.S., Rahman, S., Namekata, M., Murray, J., Nishimura, T., Mostafavi-Pour, Z., Patel, Y., Suda, Y., Humphries, M.J. and Sobel, M. (2006) Heparin-II domain of fibronectin is a vascular endothelial growth factor-binding domain: enhancement of VEGF biological activity by a singular growth factor/matrix protein synergism. *Circulation Research*, **99** (8), 853–860.

313 Miralem, T., Steinberg, R., Price, D. and Avraham, H. (2001) VEGF(165) requires extracellular matrix components to induce mitogenic effects and migratory response in breast cancer cells. *Oncogene*, **20** (39), 5511–5524.

314 Wijelath, E.S., Rahman, S., Murray, J., Patel, Y., Savidge, G. and Sobel, M. (2004) Fibronectin promotes VEGF-induced CD34 cell differentiation into endothelial cells. *Journal of Vascular Surgery*, **39** (3), 655–660.

315 Georges, P.C. and Janmey, P.A. (2005) Cell type-specific response to growth on soft materials. *Journal of Applied Physiology*, **98**, 1547–1553.

316 Goerges, A.L. and Nugent, M.A. (2004) pH regulates vascular endothelial growth factor binding to fibronectin: a mechanism for control of extracellular matrix storage and release. *The Journal of Biological Chemistry*, **279** (3), 2307–2315.

317 Nikolopoulos, S.N., Blaikie, P., Yoshioka, T., Guo, W. and Giancotti, F.G. (2004) Integrin beta4 signaling promotes tumor angiogenesis. *Cancer Cell*, **6** (5), 471–483.

318 Bon, G., Folgiero, V., Bossi, G., Felicioni, L., Marchetti, A., Sacchi, A. and Falcioni, R. (2006) Loss of beta4 integrin subunit reduces the tumorigenicity of MCF7 mammary cells and causes apoptosis upon hormone deprivation. *Clinical Cancer Research*, **12** (11 Pt 1), 3280–3287.

319 Steinbock, F.A. and Wiche, G. (1999) Plectin: a cytolinker by design. *Biological Chemistry*, **380** (2), 151–158.

320 Yi, M., Sakai, T., Fassler, R. and Ruoslahti, E. (2003) Antiangiogenic proteins require plasma fibronectin or vitronectin for in vivo activity. *Proceedings of the National Academy of Sciences of the United States of America*, **100** (20), 11435–11438.

321 Akerman, M.E., Pilch, J., Peters, D. and Ruoslahti, E. (2005) Angiostatic peptides use plasma fibronectin to home to angiogenic vasculature. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 2040–2045.

322 Vogel, V. and Baneyx, G. (2003) The tissue engineering puzzle: a molecular perspective. *Annual Review of Biomedical Engineering*, **5**, 441–463.

323 Badylak, S.F. (2002) The extracellular matrix as a scaffold for tissue reconstruction. *Seminars in Cell and Developmental Biology*, **13**, 377–383.

324 Urech, L., Bittermann, A.G., Hubbell, J.A. and Hall, H. (2005) Mechanical properties, proteolytic degradability and biological modifications affect angiogenic process extension into native and modified fibrin matrices in vitro. *Biomaterials*, **26**, 1369–1379.

325 Leor, J., Gerecht, S., Cohen, S., Miller, L., Holbova, R., Ziskind, A., Shachar, M., Feinberg, M.S., Guetta, E. and Itskovitz-Eldor, J. (2007) Human embryonic stem cell transplantation to repair the infarcted myocardium. *Heart*, **93**, 1278–1284.

326 Miyahara, Y., Nagaya, N., Kataoka, M., Yanagawa, B., Tanaka, K., Hao, H., Ishino, K., Ishida, H., Shimizu, T., Kangawa, K., Sano, S., Okano, T., Kitamura, S. and Mori, H. (2006) Monolayered mesenchymal stem cells repair scarred myocardium after myocardial infarction. *Nature Medicine*, **12** (4), 459–465.

327 Gilbert, T.W., Stewart-Akers, A.M. and Badylak, S.F. (2007) A quantitative method for evaluating the degradation of biologic scaffold materials. *Biomaterials*, **28** (2), 147–150.

328 Kubow, K.E., Klotzsch, E., Smith, M.L., Gourdon, D., Little, W., Vogel, V. Rigidity and not fibronectin conformation

controls extracellular matrix assembly by fibroblasts reseeded into de-cellularized ECM scaffolds, submitted.

329 Wipff, P.J. and Hinz, B. (2008) Integrins and the activation of latent transforming growth factor beta1: An intimate relationship. *European Journal of Cell Biology*, **87** (8–9), 601–615.

330 Dallas, S.L., Sivakumar, P., Jones, C.J., Chen, Q., Peters, D.M., Mosher, D.F., Humphries, M.J. and Kielty, C.M. (2005) Fibronectin regulates latent transforming growth factor-beta (TGF beta) by controlling matrix assembly of latent TGF beta-binding protein-1. *Journal of Biological Chemistry*, **280**, 18871–18880.

331 Wells, R.G. and Discher, D.E. (2008) Matrix elasticity, cytoskeletal tension, and TGF-beta: the insoluble and soluble meet. *Science Signaling*, **1** (10), pe13.

332 Sarikaya, A., Record, R., Wu, C.C., Tullius, B., Badylak, S. and Ladisch, M. (2002) Antimicrobial activity associated with extracellular matrices. *Tissue Engineering*, **8** (1), 63–71.

333 Schmid-Schonbein, G.W. and Hugli, T.E. (2005) A new hypothesis for microvascular inflammation in shock and multiorgan failure: self-digestion by pancreatic enzymes. *Microcirculation (New York, NY: 1994)* **12** (1), 71–82.

334 Liu, S., Calderwood, D.A. and Ginsberg, M.H. (2000) Integrin cytoplasmic domain-binding proteins. *Journal of Cell Science*, **113** (Pt 20), 3563–3571.

335 Brakebusch, C. and Fassler, R. (2003) The integrin-actin connection, an eternal love affair. *The EMBO Journal*, **22** (10), 2324–2333.

336 Wilhelmsen, K., Litjens, S.H., Kuikman, I., Tshimbalanga, N., Janssen, H., van den Bout, I., Raymond, K. and Sonnenberg, A. (2005) Nesprin-3, a novel outer nuclear membrane protein, associates with the cytoskeletal linker protein plectin. *The Journal of Cell Biology*, **171** (5), 799–810.

337 Geiger, B. (2006) A role for p130Cas in mechanotransduction. *Cell*, **127** (5), 879–881.

10
Stem Cells and Nanomedicine: Nanomechanics of the Microenvironment

Florian Rehfeldt, Adam J. Engler, and Dennis E. Discher

10.1
Introduction

Tissue cells in our body adhere to other cells and matrix and have evolved to require such attachment. While it has been known for some time that adhesion is needed for viability and normal function of most tissue cells, it has only recently been appreciated that adhesive substrates *in vivo* – namely other cells and the extracellular matrix (ECM) – are generally compliant. The only rigid substrate in most mammals is calcified bone. While the biochemical milieu for a given cell generally contains a wide range of important and distinctive soluble factors (e.g. neuronal growth factor, epidermal growth factor, fibroblast growth factor, erythropoietin), the physical environment may also possess very different *elasticity* from one tissue to another. It is well accepted that cells can 'smell' or 'taste' the soluble factors and respond via specific receptor pathways; however, it is also increasingly clear that cells 'feel' the mechanical properties of their surroundings. Regardless of the adhesion mechanism – that is, cadherins binding to adjacent cells or integrins binding to the ECM – cells engage their contractile actin/myosin cytoskeleton to exert forces on the environment, and this drives a *feedback* with responses that range from structural remodeling to differentiation. In this chapter, we aim to provide a brief overview of the diversity of *in vivo* micro/nano-environments in the human body, and also seek to describe some *mechanosensitive phenomena*, particularly with regards to adult stem cells cultured in *in vitro* systems and intended to mimic the elastic properties of native tissues.

10.2
Stem Cells in Microenvironment

10.2.1
Adult Stem Cells

Adult stem cells are distinct from embryonic stem cells (ESCs). Two properties are required for a stem cell: *self-renewal* and *pluripotency*. Stem cells must be able to

Nanotechnology, Volume 5: Nanomedicine. Edited by Viola Vogel

ISBN: 978-3-527-31736-3

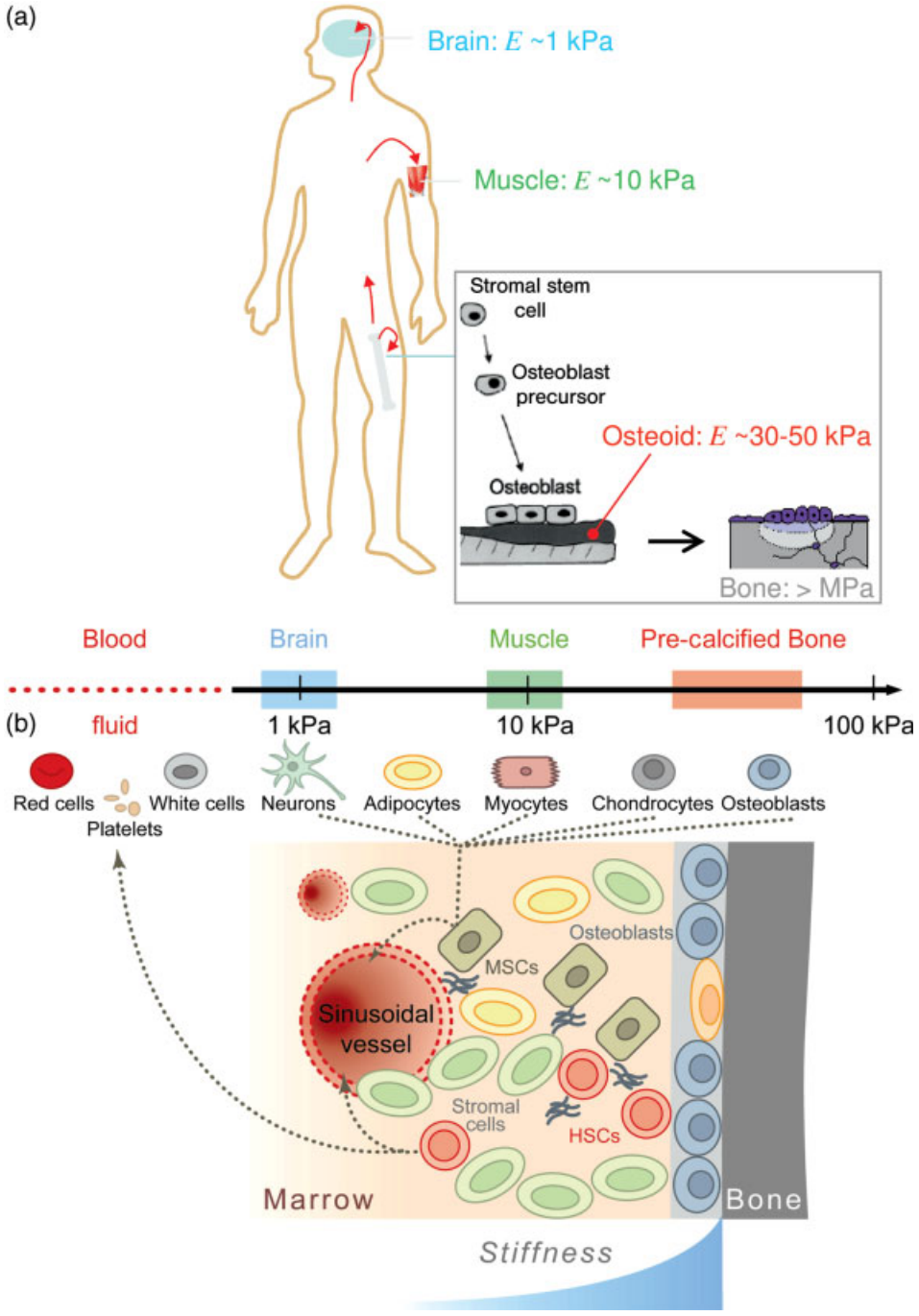
divide 'indefinitely' (or at least many times compared to other cells) and also maintain their undifferentiated state. They must be potent to differentiate into various lineages. The fertilized egg cell is *totipotent* because all possible cell types in the body are derived from it. Adult stem cells or somatic stem cells are *multipotent* and are not derived directly from eggs, sperm or early embryos, as are ESCs. Among the adult stem cells found in fully developed organisms, two classes are of paramount importance for both basic scientific inquiry and possible medical application: mesenchymal stem cells (MSCs) and hematopoietic stem cells (HSCs), both of which can be obtained from adult bone marrow (Figure 10.1).

For more than 50 years, it has been known that HSCs are present in the bone marrow and can differentiate into all of the different blood cell types. Becker and coworkers identified a certain type of cell in mouse bone marrow that, when transplanted into mice which had been heavily irradiated to kill any endogenous cells, would reconstitute the various HSCs: red blood cells, white blood cells, platelets, and so on [1]. Since then, the transplantation of HSCs has become a routine medical treatment for many blood diseases, such as leukemia. However, controversies persist regarding the differentiation and de-differentiation of HSCs, particularly whether or not they can become pluripotent, being able to differentiate into neurons, muscle or bone [2–5].

MSCs also reside in the bone marrow, and can certainly differentiate into various types of solid tissue cells such as muscle, bone, cartilage and fat. From human bone marrow, Pittenger and coworkers successfully isolated truly multipotent MSCs and also demonstrated *in vitro* differentiation into various lineages [6]. Using media cocktails – often composed of glucocorticoids such as dexamethasone – many other groups have since standardized their differentiation into different tissue lineages [7–13].

Unlike the adult stem cells, ESCs are derived from a newly fertilized egg that has divided sufficiently to form a blastocyst. Once isolated from the inner mass of the blastocyst, the ESCs are then cultured and expanded *in vitro* to produce a sufficient number of cells for study. These ESCs are pluripotent in that *all* organismal cell types in the developing embryo emerge from them. This cell type is therefore considered to be the most promising for cell therapy and regenerative medicine. On the other hand, major problems such as immune rejection are significant obstacles as ESCs will generally not be from the same organism.

Figure 10.1 *In vivo* microenvironments of adult stem cells: the physiological range of stiffness. (a) Range of physiological elasticity of native cells and tissue: mesenchymal stem cells (MSCs) reside in the bone marrow (see panel (b)), but can egress from their niche into the bloodstream and travel to different tissues and organs, facing new environments with a wide range of stiffness. Nerve cells and brain tissue (around and below 1 kPa) are softest, adipocytes are assumed to be slightly stiffer, while muscle has an intermediate stiffness (~10 kPa). The elasticity of chondrocytes is speculated to fall between myocytes and osteoids (precalcified bone precursors) that are very stiff ($E \sim 30$–50 kPa) prior to mineralization into bone (~MPa to GPa); (b) Hematopoietic stem cells (HSCs), as distinct from the MSC differentiation pathway, that also reside in the bone marrow niche are the precursor cells of all blood lineage-type cells. Within the bone marrow and between the marrow and the blood there are various gradients, for example, oxygen concentration, biochemical factors and, of course, viscoelasticity.


(a)
Brain: E ~1 kPa
Muscle: E ~10 kPa
Stromal stem cell
Osteoblast precursor
Osteoid: E ~30-50 kPa
Osteoblast
Bone: > MPa
Blood
Brain
Muscle
Pre-calcified Bone
(b)
fluid
1 kPa
10 kPa
100 kPa
Red cells
Platelets
White cells
Neurons
Adipocytes
Myocytes
Chondrocytes
Osteoblasts
Osteoblasts
MSCs
Sinusoidal vessel
Stromal cells
HSCs
Marrow
Bone
Stiffness

Regardless of the type of stem cell, it is essential to understand how the stem cell differentiates and in which ways it interacts with its environment, both within the niche and potential target tissues. Biochemical factors are key but not singular factors: mechanical cues in the microenvironment have only recently been recognized as contributing to the fate of MSCs.

In this chapter we focus on the nanomechanics of adult stem cells, as these cells interact with a substrate of well-defined stiffness. Stiffness – or, more formally, Young's elastic modulus, E (measured in Pascal (Pa)) – is a potential issue because of its physiological variation. That is, the only tissue in the body that is rigid is bone – all other solid tissues are soft, with elastic moduli in the range of 0.1 to 100 kPa (Figure 10.1a). Over the past decade, it has become particularly evident that the ability of an adherent cell to exert forces and build up tension reflects this elastic resistance of the surrounding microenvironment. This mechanosensitivity is based on the tension generated by ubiquitously expressed non-muscle myosins II (NMM II) which are the molecular motors that drive transduction. As with other class II myosins, such as skeletal muscle myosin (that moves all of our limbs against everyday), NMM II assembles into filaments. These NMM II minifilaments of ~300 nm length are bipolar with heads on either side that bind and actively traverse actin filaments [14]. In many situations involving moderately stiff substrates, actin–myosin assemblies are visible in cells as stress fibers, which is a prototypical contractile unit seen at least within cells grown on glass coverslips and other rigid substrates.

Of the three isoforms of NMM II (a–c), only non-muscle myosin IIa (NMM·IIa) appears prominent near the leading edge of crawling cells where cells probe their microenvironment, and is responsible for the bulk of force generation in non-muscle cells [15]. NMM IIa is also the only isoform that is essential in the developing embryo: [16] NMM IIa knockout mice fail to exhibit any functional differentiation beyond germ layers, in that they do not undergo gastrulation, and the null embryos die by day 7 *in utero*. Embryoid bodies grown in suspension culture also appear flat and flaccid, rather than spherical, which suggests weak and unstable cell–cell contacts, even though cadherins are clearly expressed. Such results highlight the fact that cell motors – as an active part of cell mechanics – are important even in the earliest stages of differentiation and development.

10.2.2
Probing the Nanoelasticity of Cell Microenvironments

Cells probe the elasticity of their microenvironment on the micro- and nanometer scale, whether the surroundings are tissue, ECM or an artificial culture substrate. Since this micro/nano-length scale is the range that cells can feel their surrounding, an appropriate experimental tool is needed to measure the elasticity on the same length scale. Perhaps the most suitable and pervasive technique for this is atomic force microscopy (AFM), which has been most widely used for imaging but can also make accurate force and elasticity measurements. The atomic force microscope exploits a microfabricated cantilever with a probe tip of well-defined geometry. This tip is pressed into the sample and the indentation depth as well as the required

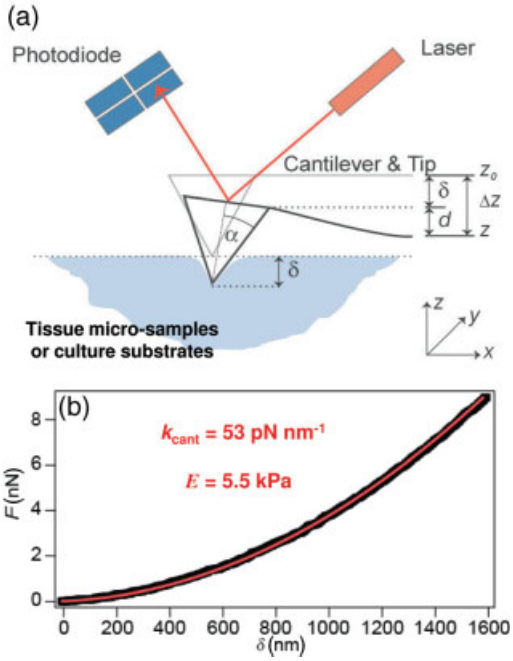


Figure 10.2 Probing the microelasticity by AFM. (a) Schematic of the measurement principle of Young's elastic modulus (E) of a hydrogel with AFM. A probe of well-defined geometry located at the tip of a cantilever with known spring constant k is pressed into the sample, and the deflection of a reflected laser beam is recorded with a four-quadrant photodiode. The required force F can be calculated by multiplying the deflection by the spring constant; (b) Force–indentation curve of a polyacrylamide gel with an elasticity of 5.5 kPa. The black points depict measured data; the solid red line represents the best fit of a modified Hertz model. (Reprinted with permission from Rehfeldt, F. *et al.* (2007) Cell responses to the mechano-chemical microenvironment – Implications for regenerative medicine and drug delivery. *Advanced Drug Delivery Reviews*, **59**, 1329–1339; © 2007, Elsevier.)

force is measured. The Young's elastic modulus E of the material surface can then be calculated from classical expressions and compared to bulk measurements, such as results from classic tensile tests.

AFM was developed during the 1980s by Binnig and coworkers [17], and was originally designed to investigate surfaces at the atomic scale with nanometer resolution. The instrument's ability to operate in a fluid environment has made it increasingly important for biological samples, and today it is used frequently to measure the nanomechanical properties of fre sh tissue samples [18], hydrogels for cell culture [18], and even living cells [19] as well as single proteins [20]. Many commercial AFM instruments allow for a precise measurement of forces, and have the ability to raster the sample at nanometer resolution, which permits mapping of a sample's elasticity.

The basic principle of AFM for determining the elasticity of a sample is sketched in Figure 10.2a. The cantilever's tip is pressed into the surface and the deflection of

the bent cantilever is monitored by a laser beam reflected onto a four-quadrant photodiode. The exerted force F can be calculated by multiplying the calibrated spring constant k of the cantilever with the measured deflection d.

$$F = k \cdot d$$

The fundamental problem of the deformation of two elastic solids was first described by Hertz in 1881 [21], and the classic Hertz model was subsequently refined and modified by Sneddon to take special geometries into account [22]. A pyramidal tip that is commonly used for imaging but also works for elasticity measurements is most simply approximated as a conical probe with an opening angle α. For this geometry, E can be calculated from

$$E = \frac{\pi(1-\nu^2)F}{\delta^2\, 2\tan\alpha}.$$

Here, F is the applied force of the tip, δ is the indentation depth of the probe into the sample, and ν is the Poisson ratio of the sample that is separately measured or can be estimated typically as 0.3–0.5.

Figure 10.2b shows a representative force indentation curve for a polyacrylamide hydrogel with a modulus of $E = 5.5$ kPa, an elasticity typical of soft tissues. The measured data points (black thick line) span a range of surface indentation (0–2000 nm) and surface forces (0–10 nN) that are typical of matrix displacements and cell-generated forces at focal adhesions [23]. AFM experiments involve measurements of the same surface THAT a cell would engage, and the results fit very well to a modified Hertz model (thin red line). The elastic modulus determined by this type of experiment only reflects the low frequency ($\sim$0.1–10 Hz) or quasi-static elasticity of the material, which is relevant to studies such as cardiac myocyte beating [23]. Additional techniques can address frequency-dependent viscoelasticity, which can be important for the interactions of cells and their surroundings. Several studies have shown a differential behavior of cells subject to an external static or oscillating force field, and a simple theoretical model seems to describe the cell response [24]. Though it seems unlikely that frequencies in the MHz range or higher will have an effect, timescales from milliseconds to hours are likely relevant when compared to processes of assembly and disassembly of actin filaments, microtubules, focal contacts and focal adhesions [25].

Frequency-dependent rheology measurements that not only encompass the static elasticity as a low-frequency limit but also measure dynamic viscous properties, are commonly performed with bulk samples using rheometers. Here, the hydrogel sample is placed between two parallel plates or a plate and a cone with very small angle (around 1°) and a well-defined stress is applied while the strain within the sample is measured. Using such a rheometer, Storm *et al.* investigated the complex rheology of several biopolymers (collagen, fibrin, neurofilaments, actin, vimentin) and polyacrylamide (PA) hydrogels, and found a nonlinear increase of the complex shear modulus with higher strain – so-called *strain-hardening behavior* [26]. In contrast to the biological gels, PA hydrogels exhibit a constant shear modulus over a large strain range. The nonlinear behavior found for biopolymers has clear, albeit unproven,

implications for cell–matrix interactions. With these instruments the bulk measures of the storage and loss modulus of the material can be determined, but they do not access the rheology on the cellular scale. AFM can be similarly used if, after the probe is indented into the sample (cells, gels, etc.), a sinusoidal signal is superimposed to measure the frequency-dependent viscoelasticty. Mahaffy *et al.* used this technique to determine the viscoelastic parameters for cells and PA gels, and found good agreement with bulk measurements, at least for PA gels [27]. Additional particle-based techniques include magnetic tweezers or magnetic twisting rheology, optical tweezers and two-particle passive rheology. Although beyond the scope of this chapter, Hoffman *et al.* have shown that the use of all these different microrheology tools converges towards a "...consensus mechanics of cultured mammalian cells" [28].

10.2.3
Physical Properties of *Ex-Vivo* Microenvironments

The microelasticity of freshly isolated tissue samples can also be determined using AFM, revealing tissue inhomogeneities vis-à-vis a lateral mapping of mechanical properties that macroscopic measurements cannot address. Thus, AFM serves as a mechanical analogue to histology as it could thus reveal microelasticity differences across diseased tissues, such as fibrotic regions.

At the whole-body scale, the elasticities of normal, soft tissue vary considerably (see Figure 10.1 and Table 10.1). Aside from bodily fluids such as blood, that obviously have zero elasticity, perhaps the softest solid tissue is the brain, with an elastic modulus of just $E \sim 0.1$–1 kPa [29–31]. Native mammary gland tissue has a similar elasticity ($E \sim 0.2$ kPa) [32]. The lateral elasticity of muscle, in its relaxed state, is significantly stiffer, with AFM probing yielding $E \sim 10$ kPa. Even at the subcellular level, myofibrils isolated from rabbit skeletal muscle have $E = 11.5 \pm 3.5$ kPa in the relaxed state [33]; this is consistent with the transverse stiffness of rat skeletal muscle measured under *in vivo* conditions ($E = 15.6 \pm 5.4$ kPa) [34], as well as the elasticity of both *ex-vivo* mouse *extensor digitorum longus* (EDL) muscle ($E \approx 12 \pm 4$ kPa) and one-week cultures of C2C12 myotubes ($E \approx 12$–15 kPa) [35]. Bone is of course the stiffest material in the body, with a modulus in the region of GPa after calcification; however, bone is a composite of protein plus mineral, and precalcified bone or 'osteoid' is

Table 10.1 Elastic modulus E of normal and diseased* tissues.

Type of tissue	Elastic modulus E [kPa]	Reference(s)
Brain (macroscopic)	0.1–1	[29–31]
Mammary gland tissue	~0.2	[32]
Muscle (passive, lateral)	~10	[33,34]
Osteoid (secreted film in culture)	30–50	[36–38]
*Mammary gland tumor tissue	~4	[32]
*Dystrophic muscle	~18	[39]
*Infarcted myocardium (surface)	~55	[42]
*Granulation tissue	~50	[41]

a heavily crosslinked network of collagen-I plus other matrix proteins such as osteocalcin [36–38] with an elastic modulus of 30–50 kPa. All of the measurements above indicate that the physiological range of the elasticity of soft solid tissues ranges from 0.1 kPa up to 100 kPa.

Abnormal or diseased tissue can exhibit an elasticity which is significantly different from normal tissue. 'Sclerosis' is Greek for 'hardening of tissue', and is a descriptor of many diseases, such as atherosclerosis, that refers to a hardening of the arteries. In the context of muscle, one AFM study has shown that the muscle of mice with dystrophy has an elevated stiffness of 18 ± 6 kPa in comparison to $E \approx 12 \pm 4$ kPa for healthy tissue [39], a difference which previously had only been described in bulk [40]. Commonly found granulation tissue after wound healing exhibits an elastic modulus of $E = 50 \pm 30$ kPa, which is three- to fivefold higher than that of normal tissue [41]. Following a myocardial infarction, the myocardium is also remodeled with fibrosis that stiffens the tissue significantly: AFM measurements have shown healthy rats with a normal myocardial $E = 10$–20 kPa that increases threefold to $E = 55 \pm 15$ kPa within the infarcted region [42]. Another significant increase in tissue stiffness from the native to the diseased state is observed in tumorigenesis. The average mammary gland tumor stiffens 20-fold to $E \approx 4 \pm 1$ kPa compared to 0.2 kPa for the normal tissue [32]. In addition to fibrosis, cell tension generally stiffens the cells and tissues; this is apparent even with isolated myofibrils that exhibit an order of magnitude increase in stiffness (relaxed: $E = 11.5 \pm 3.5$ kPa to rigor: $E = 84.0 \pm 18.1$ kPa) [33]. Whether hardening is part of the key initial causes, or simply a late after-effect of a disease, is presently unclear. Nonetheless, it seems reasonable to hypothesize that hardening contributes to the cycle of disease in many cases, with a likely basis in how cells feel normal soft tissue versus abnormally hard (sclerotic) tissue. Addressing tissue hardening in regenerative medicine would seem an important aspect of therapy.

10.3 *In Vitro* Microenvironments

Different tissue cells reside in a range of very different microenvironments *in vivo*. In order to assess biological questions in rigor, however, it is necessary to culture cells, and the first reports of culturing tissue cells *in vitro* date back to the start of the twentieth century. With time, techniques and protocols have become increasingly sophisticated, allowing the growth and maintenance of a wide variety of primary cells and cell lines. The introduction of serum-free media with well-defined soluble supplements during the 1970s opened up the possibility to investigate the effects of growth factors and other factors on cultured cells. Polystyrene (PS) cell culture flasks facilitated large-scale sterile cultures, while glass coverslips have allowed very high-resolution microscopy of live cells in culture. However, both types of material substrates – as ubiquitous as they are – are very rigid with an elastic modulus in the range of MPa or GPa, which is many orders of magnitude higher than the mechanical properties that cells encounter *in vivo* (see Figure 10.1 and Table 10.1). The difference

contributes to the distinct differences between cells cultured *in vitro* compared to those cultured *in vivo*. For example, cells grown on tissue culture plastic or glass very often exhibit 'stress fibers' that are not found *in vivo* and seem to reflect the mechanical stresses applied isometrically to rigid substrates. In the same way that more sophisticated media cocktails have been formulated and continue to be generated in order to dissect the different biochemical stimuli that affect tissue cells, there is growing effort with different substrates to better mimic the various physical and mechanical properties that cells encounter in soft tissues.

10.3.1
Cells Probe and Feel their Mechanical Microenvironment

In 1980, Harris and coworkers demonstrated that most cell types actively exert forces on their substrates [43]. The culture of 3T3 fibroblasts on thin silicone rubber films showed that these cells actively deform these films, generating a wrinkling pattern (Figure 10.3a). Opas demonstrated a decade later that chick retinal pigmented epithelial (RPE) cells exhibit a differential response to substrates that are rigid or viscoelastic, despite a similar surface composition [44]. On a thick compliant Matrigel substrate, the cells did not spread and remained heavily pigmented. In contrast, on the rigid glass substrate that was covalently coated with soluble basement membrane (Matrigel), the cells spread, developed stress fibers, vinculin- and talin-rich focal contacts, and expressed the dedifferentiated phenotype. These results were perhaps the first to suggest the mechanosensitivity of cells to substrate flexibility, although the study was far from conclusive: rather, it was limited to only two conditions with

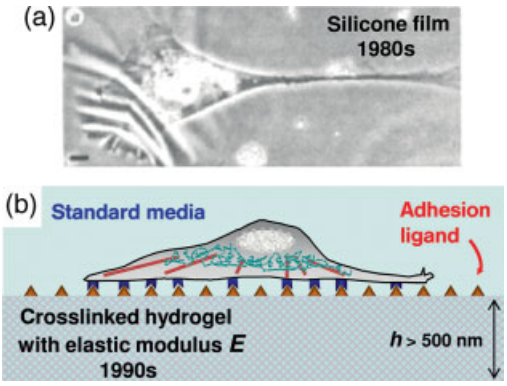


Figure 10.3 Cells pull and feel their mechanical environment. (a) 3T3 Fibroblasts cultured on a thin silicone rubber film exert forces that result in substrate wrinkling. Scale bar = 10 µm. (Reprinted with permission from Harris, A.K., Stopak, D. and Wild, P. (1981) Fibroblast traction as a mechanism for collagen morphogenesis. *Nature*, **290** (5803), 249–251; © 1981, Macmillan Publishers Ltd); (b) *In vitro* model system for cell culture on a flexible substrate. The elastic substrate with its Young's modulus *E* is coated with a ligand (e.g. collagen-I) for cell attachment. The cell attaches to this ligand and senses the elasticity via tension of actin–myosin complexes and/or stress fibers that are coupled to the substrate via integrins and other cell-surface receptors.

obvious compositional differences, namely rigid functionalized glass and soft Matrigel, and there were no quantitative measurements of substrate elasticity.

The study of cell mechanics on flexible substrates was significantly advanced in 1997 by Pelham and Wang, with their seminal studies on epithelial and fibroblast cell lines cultured on a range of collagen-I coated, elastic PA hydrogels [45]. An adhesive ligand in such studies is always needed because PA gels are not adhesive to cells (i.e. they do not engage integrins); collagen-I is a logical first choice for an ECM ligand because it is one of the most abundant proteins in mammals – which means of course that cells are very likely to encounter this protein in an organism. By using different crosslinker concentrations, a set of gels ranging in elastic modulus with an otherwise identical surface was generated. With this *in vitro* culture system, distinct differences were exhibited by cells on soft and stiff matrices: cells were seen to spread more on stiffer substrates, and also exhibited more typical focal adhesions on stiffer substrates. It was also clear that non-muscle myosin is a key player in generating force in the mechanosensitivity. When exposed to the common myosin inhibitors of the time (2,3-butanedione monoxime or KT5926), the cells could no longer distinguish between soft and stiff substrates. The effects on cell motility also have become clear: fibroblast cells that approached the transition from the soft side could easily migrate to the stiffer side, with a simultaneous increase in their spread area and traction forces [46]. In comparison, cells on the stiffer side of the gel often turned around or retracted as they reached the border. This dependence of cell movement on purely physical properties of the substrate has been termed *durotaxis*, and clearly shows that cells probe and feel the mechanics of their microenvironment.

Although these initial studies very elegantly demonstrated the differential responses of cells to substrate elasticity, the precise connections to *in vivo* microenvironments, as well as the role that diseases play in matrix stiffness, remained unclear and required further exploration. Measurements of elasticity were also only approximate: they were made by estimating indentations with steel balls of known weight. Different tissues often exhibit characteristic elasticities and can have significant alterations in disease (as discussed in Section 10.1). Engler and coworkers cultured myoblasts on the same collagen-I-coated PA gel systems with a wide range of AFM-determined elasticities, and showed that fusion into myotubes was not significantly affected, whereas myosin–actin striation was most prominent within cells grown on substrates with $E = 12 \pm 5$ kPa, which corresponds to the native muscle elasticity [39]. Gel substrates that were too stiff, as well as rigid glass, inhibited the formation of striated actin–myosin fibers. Striation was weak in myotubes on 18 kPa gels emulating dystrophic muscle, suggesting a significant counterinfluence against differentiation in disease.

In the context of wound healing, Goffin *et al*. examined fibroblast adhesion and cytoskeletal organization in cells on surfaces of different rigidity [41]. These cells play an essential role in wound healing and tissue remodeling as they migrate to wounded tissue and can develop stress fibers and tension to facilitate wound closure and healing. Using substrates that simulate normal soft tissue and stiffened wounded tissue, more contractile and differentiated myofibroblasts were only seen on the stiffer substrates. In addition, 'supermature' focal adhesions (suFAs) were found to develop only on rigid substrates and to exert a fourfold higher stress on the

matrix than was exerted under more typical focal adhesions formed on 11 kPa gels. It was proposed that this is a means by which the matrix influences the tension that the cells apply and therefore helps to steer the wound-healing process.

Similar findings were recently reported for liver-derived portal fibroblasts and their differentiation to myofibroblasts *in vitro* on PA substrates in the presence or absence of transforming growth factor-β (TGF-β) [47]. When these fibroblasts differentiate towards myofibroblasts – as occurs in response to an acute liver injury – they start to express α-smooth muscle actin (α-SMA) and form stress fibers on rigid surfaces (>3 kPa), but not on very soft (400 Pa) gels that resemble the elasticity of native rat or human liver tissue. When treated with 100 pM TGF-β the portal fibroblasts began to express α-SMA even on the soft matrix, although they did not develop organized stress fibers; stiffer matrices were required for cell spreading and stress fiber organization. Cells treated with 5 μ*M* TGF-β receptor kinase inhibitor did not differentiate on any of the substrates, which suggests that TGF-β functions as an essential contractile inducer in these cells (opposite to myosin inhibitors), leading to higher α-SMA expression and stress fiber organization as stiffer substrates. Both, biochemical and biophysical stimuli are thus part of the complex interplay of mechanosensing.

10.3.2
Cells React to External Forces

The responses of cells to physical cues in their microenvironment – namely elasticity and geometry – are not the only physical factors of importance. Cells also react to external forces, and forces are found throughout an organism. Muscles contract and relax, joints are compressed during standing, walking and running, and the average human heart beats 72 times each minute to keep our blood circulating, which leads to shear stresses on the surfaces of endothelial cells. The effects of external forces on cells have been widely studied, although it is often not clear to what extent the force-generation capabilities of cells are again part of the response.

Several studies have documented the influence of both static and dynamic strains on cells. For example, C2C12 mouse myoblasts cultured for several days on a substrate that is subject to a continuous, unidirectional stretching leads to alignment and elongation of the cells [35]. This applied static strain mimics the *in vivo* conditions of long bone growth and muscle development. Another investigation of fibroblast morphological changes in collagen matrices under a mechanical load [48] revealed cell alignment with the direction of the external force to minimize their exposure to the strain. In contrast to the parallel alignment in the case of static strain, several studies have also reported a more perpendicular orientation for rapidly oscillating external forces [49–52]. Experimental results such as these provide insight into potential mechanisms in development and repair of connective tissue.

Experiments are many, but theories are few and would benefit in understanding and predicting. Safran and coworkers have modeled the cell as a contractile dipole in an external force field [24]. This coarse-grain model of a cell aligns parallel to a static or quasistatic external oscillating force field, but it orients perpendicular to the field if the frequency is too high to follow. This is analogous to an electric dipole in an

oscillating electromagnetic field. The model not only agrees with experimental evidence but also demonstrates the applicability of basic physical concepts to cell mechanics.

10.3.3
Adult Stem Cell Differentiation

The impact of substrate elasticity on cell behavior is now evident in many studies. One last – but central – example for this chapter perhaps highlights the potent influence of elastic matrix effects, namely the differentiation of adult stem cells (MSCs) [53]. The usual method for inducing the differentiation of MSCs towards any particular lineage (e.g. adipocytes, chondrocytes, myocytes, osteocytes) is to use media cocktails based on steroids and/or growth factors [6–13]. Our approach has been to use a single, 10% serum-containing media and to vary only the stiffness of the culture substrate in sparsely plated cultures. These cells are exposed to serum *in vivo*, but during natural processes of emigration from the marrow to repair and maintain tissue, they also encounter different micromechanical environments. It is this latter aspect of environment that we sought to mimic.

MSCs were plated on collagen-I PA hydrogels of different elasticity E (Figure 10.3b) and found to exhibit after just 4 h a significantly different morphology that becomes even more pronounced over the next 24 to 96 h (Figure 10.4a). The cells spread more with increasing substrate stiffness, as found with other cells [45], but they also take on different morphologies. As the cells are multipotent, it was of further interest to assess whether substrate mechanics could also influence gene transcription, and therefore differentiation. Immunostaining for early lineage specific proteins indeed revealed that the neurogenic marker, β3 tubulin was only present on the softest 1 kPa gels, the myogenic transcription factor MyoD was most prominent on the 11 kPa gels, and an early osteogenic transcription factor, CBFα1, was detectable only on the stiffest 34 kPa substrates. Remarkably, these stiffnesses that induced differentiation correspond to the elasticities that the various lineages would experience in their respective native tissues: quantitative analyses of differentiation markers emphasizes the finding that adult stem cells adjust their lineage towards the stiffness of their surrounding tissues (Figure 10.4b).

Figure 10.4 (*Continued*) increase of any of the three proteins. Dashed green and orange curves depict the substrate-dependent upregulation of the markers for already committed cells [C2C12 (muscle) and hFOB (bone)] exhibiting the same qualitative behavior at a higher intensity due to their committed nature; (c) Transcription profiling array shows selective upregulation of several lineage-specific genes due to matrix elasticity. Values for MSCs cultured on PA gels for one week were normalized by β-actin and then further normalized with data obtained from naïve, undifferentiated MSCs before plating. Red depicts relative upregulation; green shows downregulation. Gene transcripts of the different lineages are upregulated only on the substrates with the appropriate stiffness; blebbistatin treatment inhibits this upregulation and thus differentiation. (Reprinted with permission from Engler, A.J. *et al.* (2006) Matrix elasticity directs stem cell lineage specification. *Cell*, **126**, 677–689; © 2006, Elsevier).

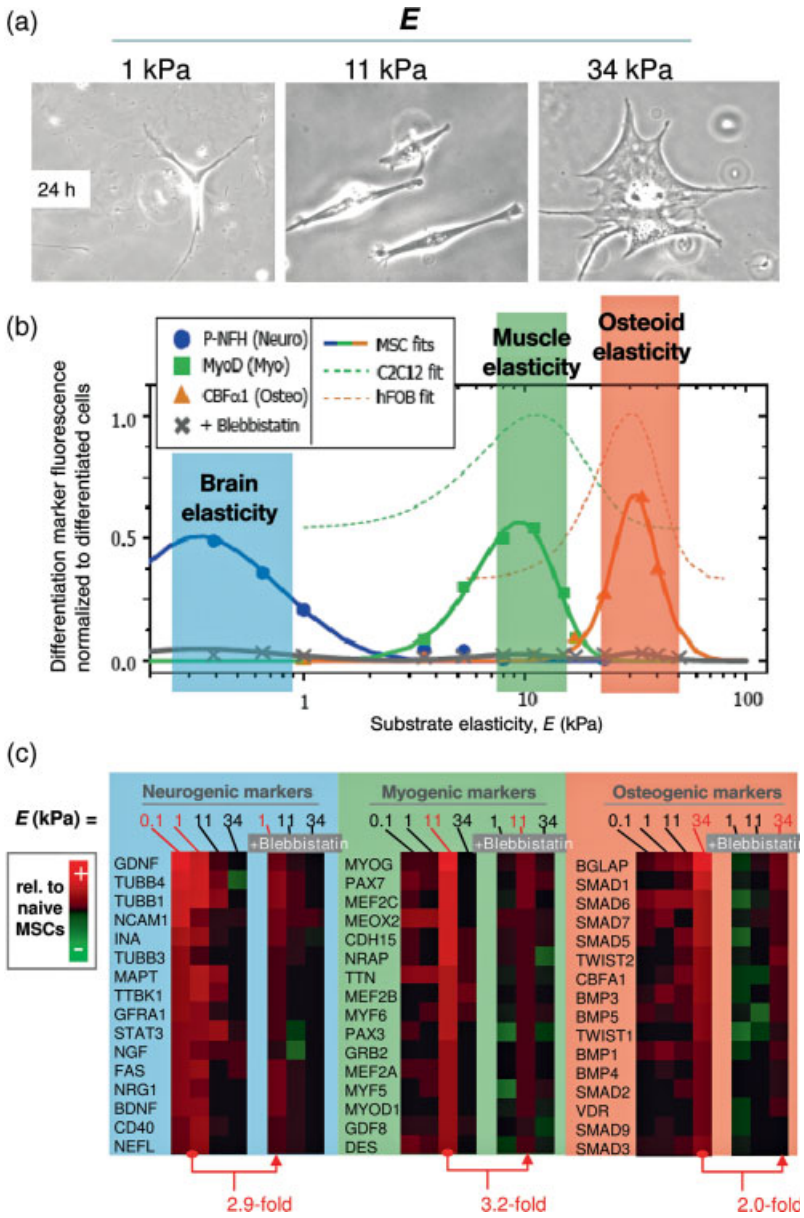


Figure 10.4 Differentiation of adult stem cells guided by matrix elasticity. (a) Mesenchymal stem cells (MSCs) on collagen-I coated PA gels with different elasticities ($E = 1$ kPa; 11 kPa; 34 kPa) show distinct morphology at 24 h after plating; (b) Quantitative immunofluorescence of the lineage markers (blue, P-NFH (neuro); green, MyoD (muscle); orange, CBFα1 (osteo)) reveals the stiffness-dependent differentiation of the MSCs. The multipotent stem cells upregulate differentiation markers only on substrates yielding a stiffness in the range of the native tissue, respectively. The gray curve of blebbistatin-treated cells shows no selective

Treatment of the MSCs with blebbistatin, a potent, recently synthesized NMM II inhibitor, largely blocked the expression of any of the differentiation markers, and again highlighted the key role of this motor in sensing the substrate in mechanoguided differentiation. Repeating the same experiment with two cell lines that were already committed (C2C12 mouse myoblasts and hFOB, human osteoblasts) showed a similar upregulation of the differentiation marker according to the tissue-level elasticity, but there was also a higher, baseline level of expression that reflected the fact that these cells were already committed. This led to a new hypothesis of differentiation mechanisms suggesting that both biochemical and biophysical stimuli influenced the differentiation of these *multipotent* adult stem cells.

Transcript profiling of some of the most commonly accepted lineage markers was used to more broadly assess lineage specification by matrix. The top-16 genes profiled for neuro-, myo- and osteo-genesis show selective upregulation of several relative to the naïve undifferentiated MSCs before plating (Figure 10.4c). Consistent with protein markers, it can be shown that β3 tubulin is the sixth-ranked gene on the softest 0.1–1 kPa gels, MyoD is the 14th-ranked gene on the 11 kPa gels, and CBFα1 is the seventh-ranked gene on the stiffest 34 kPa matrices. Also consistent with the downregulation of protein with blebbistatin treatment, the transcripts also exhibited a downregulation of about two to threefold.

Further examination of differentially expressed genes is revealing. Neural growth factors such as glial-derived neurotrophic factor (GDNF) and nerve growth factor (NGF) are upregulated on softer matrices. GDNF is interesting because its most prominent feature is its ability to support the survival of dopaminergic and motor neurons. The latter neuronal populations die during the course of amyotrophic lateral sclerosis (ALS). Myostatin (GDF8) is upregulated on the 11 kPa myogenic matrix and secreted by skeletal muscle cells; it is understood to circulate and act as a negative regulator of muscle mass, slowing down the myogenesis of muscle stem cells. Several bone growth factors (e.g. bone morphogenetic proteins: BMP 3, 4, 5) are upregulated on the stiffest osteogenic matrices. These proteins are interesting as potent osteoinductive growth factors belonging to the TGF-β superfamily, which was described in Section 10.2.1 as promoting stress fibers in fibroblasts on stiff matrices. This is very consistent with stress fiber assembly seen also in the MSCs [53]. Follow-up studies are certainly required to assess the secretion of these factors as well as autocrine–paracrine loops, although matrix elasticity is clearly the initiating factor throughout. Additionally, the many transcription factor genes listed (e.g. STATs, MYFs, MEFs and SMADs), as well as the many cytoskeletal and adhesion transcripts (e.g. NCAM, TTN and BGLAP (or osteocalcin)) make for a compelling story of how these MSCs physically interact with their microenvironment and reprogram their gene expression accordingly.

10.3.4
Implications for Regenerative Medicine

MSCs are believed to have considerable potential for cell therapies and regenerative medicine. Taking into account the impact of the microenvironment (as described

above), it perhaps becomes clear how important it is to carefully assess potential applications of these cells.

One application which currently is undergoing major exploration is the injection of purified and enriched MSCs into a stiffened infarct of the heart – a technique known as *cellular cardiomyoplasty*. The hope is that these adult stem cells will differentiate to cardiomyocytes and improve contractile function, although recent animal models and even clinical trials have yielded mixed results at best [42,54–56]. For example, in one rat infarct model, the injection of human-MSCs was found to marginally improve myocardial compliance as determined using AFM [42], but the MSCs did not regenerate the infarcted heart muscle tissue. Working strictly with a mouse model, Fleischmann and coworkers also injected MSCs into an infarcted myocardium [55] and, two to four weeks after injection, identified encapsulated calcifications and ossifications in the infarcted zone. These compartments were clearly restricted to the scarred region of the infarct where the elastic modulus E is much higher than that of native cardiac muscle. Interestingly, when MSCs were injected into intact non-infarcted hearts, calcifications and ossifications occurred only on the scar tissue along the injection channel. These findings were in excellent agreement with the *in vitro* studies of Engler *et al.* [53], where osteogenesis of MSCs was found on matrices having an elasticity in the range of 30–50 kPa (Figure 10.4) – that is, the stiffness of postinfarct scar tissue.

For future experiments and clinical trials, it will be of paramount importance to clearly dissect all of the possible cell stimuli in order to at least avoid negative implications for the patient such as calcifications. Our cells live in a 'world' of biophysics and biochemistry, and it seems necessary to understand and control parameters of both sides.

10.4
Future Perspectives

This chapter could only highlight some of many studies on the implications of the *mechano*-chemical environment of cells, even this small selection underscores the importance of a better understanding of the interactions between cells and environment to improve the design of therapeutic applications. Adult stem cells are probably one of the most promising candidates for successful tissue regeneration, given their multipotency, availability and limited ethical considerations, although their interactions with the microenvironment must be taken into account. Further studies must elucidate the complex interplay of biochemistry and biophysics, and should identify ways to influence either side with stimuli from the other. As a prime example, approaches to repair the infarcted heart reveal how new strategies are needed to overcome the physical limitations of a fibrotic tissue. Perhaps it is possible to alter the cell's perception of the surrounding stiffness so that adult stem cells could develop towards a suitable myogenic lineage (instead of osteogenic)? This is clearly a large playground for future studies of what are ultimately diseases that couple to cell mechanics.

Acknowledgments

F.R. gratefully acknowledges the Feodor Lynen fellowship from the Alexander von Humboldt foundation, and thanks André E.X. Brown for critical reading of the manuscript and Andrea Rehfeldt for help with the illustrations. A.J.E. and D.E.D. acknowledge the NIH and NSF for support via NRSA and R01 funding, respectively.

References

1 Becker, A.J., Till, J.E. and McCulloch, E.A. (1963) *Nature*, **197**, 452.
2 Corbel, S.Y., Lee, A., Yi, L., Duenas, J., Brazelton, T.R., Blau, H.M. and Rossi, F.M.V. (2003) *Nature Medicine*, **9**, 1528.
3 Hess, D.C., Abe, T., Hill, W.D., Studdard, A.M., Carothers, J., Masuya, M., Fleming, P.A., Drake, C.J. and Ogawa, M. (2004) *Experimental Neurology*, **186**, 134.
4 Roybon, L., Ma, Z., Asztely, F., Fosum, A., Jacobsen, S.E.W., Brundin, P. and Li, J.Y. (2006) *Stem Cells (Dayton, Ohio)*, **24**, 1594.
5 Deten, A., Volz, H.C., Clamors, S., Leiblein, S., Briest, W., Marx, G. and Zimmer, H.G. (2005) *Cardiovascular Research*, **65**, 52.
6 Pittenger, M.F., Mackay, A.M., Beck, S.C., Jaiswal, R.K., Douglas, R., Mosca, J.D., Moorman, M.A., Simonetti, D.W., Craig, S. and Marshak, D.R. (1999) *Science*, **284**, 143.
7 Caplan, A.I. (1991) *Journal of Orthopaedic Research*, **9**, 641.
8 Hofstetter, C.P., Schwarz, E.J., Hess, D., Widenfalk, J., El Manira, A., Prockop, D.J. and Olson, L. (2002) *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 2199.
9 Kondo, T., Johnson, S.A., Yoder, M.C., Romand, R. and Hashino, E. (2005) *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 4789.
10 McBeath, R., Pirone, D.M., Nelson, C.M., Bhadriraju, K. and Chen, C.S. (2004) *Developmental Cell*, **6**, 483.
11 Kuznetsov, S.A., Krebsbach, P.H., Satomura, K., Kerr, J., Riminucci, M., Benayahu, D. and Robey, P.G. (1997) *Journal of Bone and Mineral Research*, **12**, 1335.
12 Prockop, D.J. (1997) *Science*, **276**, 71.
13 Yoo, J.U., Barthel, T.S., Nishimura, K., Solchaga, L., Caplan, A.I., Goldberg, V.M. and Johnstone, B. (1998) *Journal of Bone and Joint Surgery - American Volume*, **80**, 1745.
14 Verkhovsky, A.B., Svitkina, T.M. and Borisy, G.G. (1995) *Journal of Cell Biology*, **131**, 989.
15 Cai, Y.F., Biais, N., Giannone, G., Tanase, M., Jiang, G.Y., Hofman, J.M., Wiggins, C.H., Silberzan, P., Buguin, A., Ladoux, B. and Sheetz, M.P. (2006) *Biophysical Journal*, **91**, 3907.
16 Conti, M.A., Even-Ram, S., Liu, C.Y., Yamada, K.M. and Adelstein, R.S. (2004) *Journal of Biological Chemistry*, **279**, 41263.
17 Binnig, G., Quate, C.F. and Gerber, C. (1986) *Physical Review Letters*, **56**, 930.
18 Engler, A.J., Richert, L., Wong, J.Y., Picart, C. and Discher, D.E. (2004) *Surface Science*, **570**, 142.
19 Radmacher, M. (2002) Measuring the elastic properties of living cells by the atomic force microscope, in *Methods in Cell Biology* (eds B.P. Jena and H.J.K. Horber), Vol. 68, Academic Press, San Diego, pp. 67–90.
20 Ludwig, M., Rief, M., Schmidt, L., Li, H., Oesterhelt, F., Gautel, M. and Gaub, H.E. (1999) *Applied Physics A - Materials Science and Processing*, **68**, 173.
21 Hertz, H. (1881) *Journal für Die Reine und Angewandte Mathematik*, **92**, 156.

22 Sneddon, I.N. (1965) *International Journal of Engineering Science*, **3**, 47.

23 Balaban, N.Q., Schwarz, U.S., Riveline, D., Goichberg, P., Tzur, G., Sabanay, I., Mahalu, D., Safran, S., Bershadsky, A., Addadi, L. and Geiger, B. (2001) *Nature Cell Biology*, **3**, 466.

24 De, R. Zemel, A. and Safran, S.A. (2007) *Nature Physics*, **3**, 655.

25 von Wichert, G., Haimovich, B., Feng, G.S. and Sheetz, M.P. (2003) *EMBO Journal*, **22**, 5023.

26 Storm, C., Pastore, J.J., MacKintosh, F.C., Lubensky, T.C. and Janmey, P.A. (2005) *Nature*, **435**, 191.

27 Mahaffy, R.E., Shih, C.K., MacKintosh, F.C. and Kas, J. (2000) *Physical Review Letters*, **85**, 880.

28 Hoffman, B.D., Massiera, G., Van Citters, K.M. and Crocker, J.C. (2006) *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 10259.

29 Gefen, A. and Margulies, S.S. (2004) *Journal of Biomechanics*, **37**, 1339.

30 Lu, Y.B., Franze, K., Seifert, G., Steinhauser, C., Kirchhoff, F., Wolburg, H., Guck, J., Janmey, P., Wei, E.Q., Kas, J. and Reichenbach, A. (2006) *Proceedings of the National Academy of Sciences of the United States of America* **103**, 17759.

31 Georges, P.C., Miller, W.J., Meaney, D.F., Sawyer, E.S. and Janmey, P.A. (2006) *Biophysical Journal*, **90**, 3012.

32 Paszek, M.J., Zahir, N., Johnson, K.R., Lakins, J.N., Rozenberg, G.I., Gefen, A., Reinhart-King, C.A., Margulies, S.S., Dembo, M., Boettiger, D., Hammer, D.A. and Weaver, V.M. (2005) *Cancer Cell*, **8**, 241.

33 Yoshikawa, Y., Yasuike, T., Yagi, A. and Yamada, T. (1999) *Biochemical and Biophysical Research Communications* **256**, 13.

34 Bosboom, E.M.H., Hesselink, M.K.C., Oomens, C.W.J., Bouten, C.V.C., Drost, M.R. and Baaijens, F.P.T. (2001) *Journal of Biomechanics*, **34**, 1365.

35 Collinsworth, A.M., Torgan, C.E., Nagda, S.N., Rajalingam, R.J., Kraus, W.E. and Truskey, G.A. (2000) *Cell and Tissue Research*, **302**, 243.

36 Morinobu, M., Ishijima, M., Rittling, S.R., Tsuji, K., Yamamoto, H., Nifuji, A., Denhardt, D.T. and Noda, M. (2003) *Journal of Bone and Mineral Research*, **18**, 1706.

37 Andrades, J.A., Santamaria, J.A., Nimni, M.E. and Becerra, J. (2001) *International Journal of Developmental Biology*, **45**, 689.

38 Holmbeck, K., Bianco, P., Caterina, J., Yamada, S., Kromer, M., Kuznetsov, S.A., Mankani, M., Robey, P.G., Poole, A.R., Pidoux, I., Ward, J.M. and Birkedal-Hansen, H. (1999) *Cell*, **99**, 81.

39 Engler, A.J., Griffin, M.A., Sen, S., Bönnemann, C.G., Sweeney, H.L. and Discher, D.E. (2004) *Journal of Cell Biology*, **166**, 877.

40 Stedman, H.H., Sweeney, H.L., Shrager, J.B., Maguire, H.C., Panettieri, R.A., Petrof, B., Narusawa, M., Leferovich, J.M., Sladky, J.T. and Kelly, A.M. (1991) *Nature*, **352**, 536.

41 Goffin, J.M., Pittet, P., Csucs, G., Lussi, J.W., Meister, J.J. and Hinz, B. (2006) *The Journal of Cell Biology*, **172**, 259.

42 Berry, M.F., Engler, A.J., Woo, Y.J., Pirolli, T.J., Bish, L.T., Jayasankar, V., Morine, K.J., Gardner, T.J., Discher, D.E. and Sweeney, H.L. (2006) *American Journal of Physiology - Heart and Circulatory Physiology*, **290**, H2196.

43 Harris, A.K. Wild, P. and Stopak, D. (1980) *Science*, **208**, 177.

44 Opas, M. (1989) *Developmental Biology*, **131**, 281.

45 Pelham, R.J. and Wang, Y.L. (1997) *Proceedings of the National Academy of Sciences of the United States of America*, **94**, 13661.

46 Lo, C.M., Wang, H.B., Dembo, M. and Wang, Y.L. (2000) *Biophysical Journal*, **79**, 144.

47 Li, Z.D., Dranoff, J.A., Chan, E.P., Uemura, M., Sevigny, J. and Wells, R.G. (2007) *Hepatology (Baltimore, Md)*, **46**, 1246.

48 Eastwood, M., Mudera, V.C., McGrouther, D.A. and Brown, R.A. (1998) *Cell Motility and the Cytoskeleton*, **40**, 13.

49 Shirinsky, V.P., Antonov, A.S., Birukov, K.G., Sobolevsky, A.V., Romanov, Y.A., Kabaeva, N.V., Antonova, G.N. and Smirnov, V.N. (1989) *Journal of Cell Biology*, **109**, 331.

50 Hayakawa, K. Sato, N. and Obinata, T. (2001) *Experimental Cell Research*, **268**, 104.

51 Kurpinski, K., Chu, J., Hashi, C. and Li, S. (2006) *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 16095.

52 Cha, J.M., Park, T.N., Noh, T.H. and Suh, T. (2006) *Artificial Organs*, **30**, 250.

53 Engler, A.J., Sen, S., Sweeney, H.L. and Discher, D.E. (2006) *Cell*, **126**, 677.

54 Lee, M.S. Lill, M. and Makkar, R.R. (2004) *Reviews in Cardiovascular Medicine*, **5**, 82.

55 Breitbach, M., Bostani, T., Roell, W., Xia, Y., Dewald, O., Nygren, J.M., Fries, J.W.U., Tiemann, K., Bohlen, H., Hescheler, J., Welz, A., Bloch, W., Jacobsen, S.E.W. and Fleischmann, B.K. (2007) *Blood*, **110**, 1362.

56 Murry, C.E., Soonpaa, M.H., Reinecke, H., Nakajima, H., Nakajima, H.O., Rubart, M., Pasumarthi, K.B.S., Virag, J.I., Bartelmez, S.H., Poppa, V., Bradford, G., Dowell, J.D., Williams, D.A. and Field, L.J. (2004) *Nature*, **428**, 664.

11
The Micro- and Nanoscale Architecture of the Immunological Synapse

Iain E. Dunlop, Michael L. Dustin, and Joachim P. Spatz

11.1
Introduction

In vivo, biological cells come into direct physical contact with other cells, and with extracellular matrices in a wide variety of contexts. These contact events are in turn used to pass an enormous variety of cell signals, often by bringing ligand–receptor pairs on adjacent cells into contact with each other. Whereas, some traditional outlooks on cell signaling arguably focused strongly on these individual ligation events as triggers for signaling cascades, it is now becoming clear that this is insufficient. Rather, in some cases where signal-activating ligands are found on cell or matrix surfaces *in vivo*, the properties of each surface as a whole need to be considered if the events leading to signaling are to be fully understood. That is, the strength of signaling – and whether signaling occurs at all – can depend on factors such as the spatial distribution of signaling-inducing ligands that are presented on a surface, the mobility of these ligands, the stiffness of the substrate, and the force and contact time between the surface and the cell being stimulated [1]. The effects of such surface properties on the activation of cell signaling pathways can often be studied by bringing the cells into contact with artificial surfaces, the properties of which can be controlled and systematically varied, so that the effects of such properties on signaling pathway activation can be observed. These studies have been successfully conducted in the context of signaling pathways associated with cell behaviors such as fibroblast adhesion to the extracellular matrix (ECM) [2, 3] and rolling adhesion of leukocytes [4, 5]. One important system in which cell–cell communication has been studied is the immunological synapse formed between T lymphocytes and tissue cells at multiple stages of the immune response.

We first introduce the role of T cells in the immune response and the concept of an immunological synapse (for an introduction to immunological concepts, see Ref. [6]). T cells are an important component of the mammalian adaptive immune system, and each of the billions of T cells in a mammal expresses a unique receptor generated by the recombination of variable genomic segments. This can then serve as a substrate

Nanotechnology, Volume 5: Nanomedicine. Edited by Viola Vogel

ISBN: 978-3-527-31736-3

for the selection of pathogen-specific T cells suitable for combating only infections by identical or similar pathogens, the proteomes of which share a particular short peptide sequence, known as the T cell's *agonist peptide*. There are three main subclasses of T cell, classified according to their effector functions: helper-, killer- and regulatory T cells. Broadly speaking, helper T cells act to stimulate and maintain the immune response in the vicinity of an infection, whereas killer T cells are responsible for detecting and destroying virus-infected cells. Regulatory T cells play a role in the prevention of autoimmune disease. In this chapter we will concern ourselves almost entirely with the activation of helper T cells. As the number of possible pathogens is enormous, the body does not maintain large stocks of T cells of a wide variety of specificities, but rather maintains small numbers of inactive T cells of each possible specificity in locations such as the lymph nodes and the spleen. When a pathogen is detected in the body, specialist antigen-presenting cells (APCs) travel to these locations and locate the correct T cells to combat the infection. This causes the T cells to become *activated*, whereupon they proliferate, producing a large number of T cells that travel to the infected tissues to carry out their antipathogen roles. Activation of the T cell occurs during direct physical contact between the T cell and the APC, and proceeds via the formation of a stable contact region between the T cell and the APC, known as the immunological synapse. (The term 'synapse' was applied due to a number of shared features with neurological synapses, such as stability, a role for adhesion molecules and directed transfer of information between cells [7].) It is known that one of the central requirements for activation is the ligation of T-cell receptors (TCRs) on the T-cell surface by peptide-major histocompatibility complex protein (p-MHC) complexes on the APC surface. The MHC may be thought of as a molecule in which the APC mounts short peptides made by breaking down all proteins in the cytoplasm (MHC class I) or in the endocytic pathway (MHC class II), including both pathogenic and 'self' proteins. As MHC class II molecules are relevant for helper T cells, we will focus on these from here on. Each TCR molecule strongly binds MHC molecules that mount the agonist peptide, and weakly binds MHC molecules that mount a subset of self-peptides. These strong and weak interactions synergize to generate sustained signals in the T cell. Thus, the APC activates only the correct T cells to combat the particular infection that is under way due to the necessary role of the agonist peptide in T-cell activation, but does so with great sensitivity (early in infection) due to the contribution of the self-peptides [8].

In addition to the initial activation process, helper T cells may encounter other agonist p-MHC-bearing APCs later in the infection process with which they can also form immunological synapses. This is particularly important in the infected tissues, where helper T cells coordinate responses by many immune cell types. The signaling from these synapses effectively informs the T cells that the infection is still in progress, encouraging them to continue countering it locally. Although there are differences between the initial activation process and these subsequent restimulations, similar signaling methods may underlie them both, and we will usually not be concerned with such distinction in this chapter. Although most of the agonist peptide-specific T cells will die when the infectious agent has been eliminated from the body, a small subset will live on as 'memory' T cells and can facilitate the mounting of a response to

reinfection with the same or closely related agents at a later time [9–11]. In fact, memory T cells are the basis of vaccination, and the process by which they are reactivated is likely to be similar in its requirements for immunological synapse formation.

Although TCR–p-MHC ligation is necessary for T-cell activation, there is evidence that the structure of the T cell–APC contact zone on a wide variety of length scales from tens of micrometers down to one to a few nanometers plays a role in determining the strength of activation signaling. Artificial surfaces functionalized with p-MHC and other immune cell proteins have been used to study structures that arise in the contact zone, and their effect on the activation process.

In this chapter, we will review the emerging body of work in which surface functionalization and lithography techniques have been used to produce artificial surfaces that have shed light on the nature and dynamics of the immunological synapse. We will first consider the structure of the immunological synapse on the micrometer scale, including the so-called supramolecular activation clusters (SMACs). These are essentially segregated areas in which different ligated receptor species predominate. Although SMACs have been widely studied, it now seems unlikely that they are the critical structures in providing activation signaling, with smaller-scale microclusters consisting of around 5–20 TCR–p-MHC pairs bound closely together being of greater significance [12]; these microclusters will also be discussed. We will then describe experiments that demonstrate the importance of the spatial distribution of molecules on the nanometer scale – that is, one to a few protein molecules – using materials such as soluble p-MHC oligomers to stimulate T cells. By illustrating the importance of the nanoscale, these results should motivate future studies in which T cells are brought into contact with surfaces that are patterned on a nanometer scale with p-MHC and other immunological synapse molecules. We will then discuss an emerging nanolithography technique that could plausibly be used to perform such studies, namely block copolymer micellar nanolithography. Finally, we will consider the possibility of making direct therapeutic use of micro- and nanopatterned T-cell-activating surfaces, and conclude that the most likely application is in *adoptive cell therapy*. In this method, T cells are removed from a patient, expanded *in vitro*, and returned to the patient to combat a disease – most commonly a cancerous tumor. It has been suggested that the success of adoptive cell therapy can depend heavily on the detailed phenotype of the returned T cells; the use of micro- and nanopatterned surfaces for *in vitro* T-cell activation could help to control their phenotype.

11.2
The Immunological Synapse

11.2.1
Large-Scale Structure and Supramolecular Activation Clusters (SMACs)

The immunogical synapse is a complex structure, which features a number of important ligand–receptor interactions in addition to the crucial TCR–p-MHC

interactions. The artificial substrates discussed here are based on a simplified model of the synapse that includes two of the most significant ligand–receptor pairs: TCR with p-MHC, and lymphocyte function-associated antigen 1 (LFA-1) with intracellular adhesion molecule 1 (ICAM-1). LFA-1 is an integrin-family protein, the function of which is to control T cell to APC adhesion. LFA-1 is expressed on T cell surfaces and binds ICAM-1 on the APC surface.

A major contribution to the understanding of the immunological synapse has been derived from studies in which T cells are allowed to settle on glass substrates on which lipid bilayers have been deposited. These bilayers contain some lipid molecules that are bound to protein constructs containing the extracellular portions of p-MHC and ICAM-1, respectively. Due to the fluidity of the lipid bilayer, the p-MHC and ICAM-1 are mobile, creating a simplified model of the APC surface (see for example, Refs [12–15]). Although a simplified system, this model reproduces features of immunological synapses observed *in vivo* with some types of APC, for example the so-called *B cells*; differences between these synapses and those observed between T cells and so-called *dendritic cells* (another type of APC) may be due to the dendritic cell cytoskeleton's restricting and controlling of p-MHC and/or ICAM-1 motion [16].

On the largest length scales, the evolution of the synapse can be seen to proceed in three stages, as illustrated in Figure 11.1 [13]. In the first stage, the T cell is migrating over the model bilayer surface; this corresponds to an *in vivo* T cell forming transient contacts with passing APCs. A central core of adhesive LFA-1–ICAM-1 contacts forms, around which the cytoskeleton deforms to produce an area of very close contact between the cell and the substrate, in which TCR with agonist p-MHC pairs can readily form. This cytoskeletal deformation is important, as TCR and p-MHC molecules are both rather short and consequently easily prevented from coming into contact by abundant larger glycosolated membrane proteins [13]. Signaling arising from the formation of TCR–p-MHC pairs causes LFA-1 molecules to change their

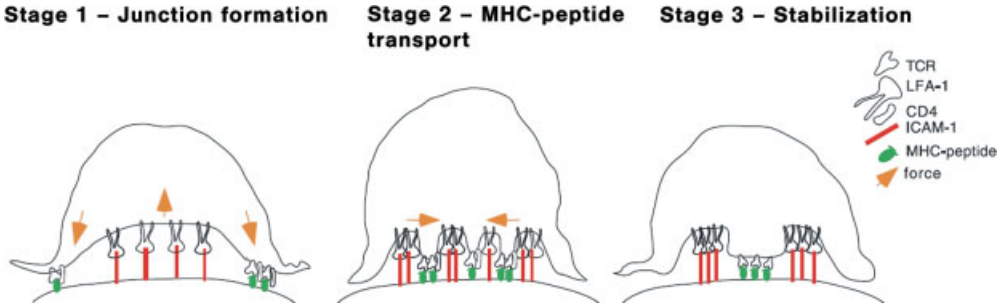


Figure 11.1 Schematic showing the three stages in the formation of the immunological synapse. A detailed description is provided in the text. Briefly, in stage 1, a central area of ICAM-1 ligating LFA-1 forms, around which the cytoskeleton rearranges to give a narrow zone in which p-MHCs can readily ligate TCRs. In stage 2, p-MHC-ligated TCRs move to the center of the contact zone, leading to stage 3, where a central area rich in TCRs and p-MHCs (the cSMAC) is surrounded by an annular area rich in ICAM-1-ligated LFA-1 (the pSMAC). (In the present chapter, to simplify the discussion, the role of CD4 is not described.) (Reproduced from *Science* **1999**, *285*, 221–227 [13]. Reprinted with permission from AAAS.)

shape so that they bind ICAM-1 strongly, which in turn causes the cell to stop migrating and thus stabilizes the synapse. *In vivo*, this mechanism enables the APCs to adhere strongly to T cells of the correct specificity, for the periods of up to several hours that may be necessary for full activation to take place. However, it simultaneously prevents the APCs from forming time-wasting, long-lasting contacts with other T cells.

In the second stage of immunological synapse evolution, p-MHC-ligated TCRs migrate to the middle of the contact zone, possibly due to actin-based transport, leading to the third stage, where a stable central region of bound TCR–p-MHC pairs, the central supramolecular activation cluster (cSMAC), is surrounded by a ring of ICAM-1 bound to LFA-1, the peripheral supramolecular activation cluster (pSMAC), which in turn is surrounded by an area of very close contact between the cell and the surface, suitable for the formation of new TCR–p-MHC pairs, the distal supramolecular activation cluster (dSMAC). As the primary purpose of the LFA-1–ICAM-1 bond is to bind the T cell to the APC, most of the lines of adhesive force between the cells pass through the dSMAC where these molecules are highly present, as shown by the arrows in Figure 11.1. Close examination of the structure and dynamics of cytoskeletal actin in the cell, as well as the distribution of LFA-1–ICAM-1 pairs, has shown that the dSMAC and pSMAC may be thought of as respectively analogous to *lamellapodium* and *lammella* structures that are exhibited by many motile cells [17], such as fibroblasts moving across the ECM. In the case of the fibroblast, $\alpha_v\beta_3$ integrin molecules (analogous to LFA-1) bind to the ECM surface in the lamellapodium, which is pushed out in the direction of desired motion. A characteristic feature is that the actin in the lamellapodium/dSMAC is organized into two stacked layers, whereas that in the lamella/pSMAC is organized into one layer only. By a combination of actin polymerization at the periphery and depolymerization at the cell center, the cell center effectively pulls on the anchored integrin molecules and thus moves towards them. In the case of the immunological synapse, the same actin polymerization and depolymerization occur, but because the dSMAC extends out in all directions, and because the ICAM-1 molecules are mobile in the APC lipid membrane (in contrast to integrin-binding elements of the ECM), the center of the T cell remains stationary, although there is a constant motion of actin towards the center of the cell [18]. Some important implications of this effect will be described in Section 11.2.2.

In a recent study conducted by Doh and Irvine, photolithographic methods were used to produce a substrate that encouraged T cells to form a cSMAC/pSMAC-like structure, but without using mobile ligands [19]. Rather, a surface was patterned with shaped patches of anti-CD3, a type of antibody that binds TCR and can thus simulate the effect of p-MHC, against a background of ICAM-1. This study employed a novel method for patterning surfaces with two proteins using photolithography, by using a photoresist that can be processed using biological buffers [20]. The photoresist used in this method is a random terpolymer with a methacrylate backbone, and methyl, *o*-nitrobenzyl and poly(ethylene glycol) (PEG) ($M_n \approx 600$ Da) side groups randomly distributed along the chain, where some of the PEG side chains are terminated with biotin. The PEG chains make the polymer somewhat hydrophilic, while the *o*-nitrobenzyl group can be cleaved to a carboxylic acid-bearing group by ultraviolet

(UV) light. The photoresist is spincoated onto a catonic substrate (in this case aminosilane-functionalized glass) so that, if the resist is exposed to UV light, and then rinsed with pH 7.4 buffer so that it contains negatively charged carboxylic acid groups, the negative charge of these groups causes the majority of the photoresist to be soluble and thus to be washed away. This will leave behind a thin layer of resist molecules, the negatively charged groups of which are ionically bound to positively charged amine groups on the glass surface. The sequence of events in preparing the patterned surface used in the T-cell studies [19] is shown in Figure 11.2. The photoresist layer was first exposed to UV through a photomask and then washed with pH 7.4 buffer to remove all but the thin residual polymer layer from the areas to be patterned with anti-CD3. After further UV irradiation of the whole surface, streptavidin followed by biotinylated anti-CD3 was deposited at pH 6.0 (at which the resist is stable) with the streptavidin binding the anti-CD3 to the biotin sites on the resist polymer surfaces. A second wash at pH 7.4 removed most of the resist from the non-anti-CD3 functionalized area, leaving a thin biotinylated polymer layer to which streptavidin followed by biotinylated ICAM-1 could be attached.

The T cells that were allowed to settle on surfaces bearing widely spaced circles of anti-CD3 6 μm in diameter against a background of ICAM-1 (prepared using this method) migrated until they encountered an anti-CD3 circle, and then formed a central cSMAC-like area of TCRs bound to anti-CD3, surrounded by a pSMAC-like ring of LFA-1 bound to ICAM-1 [19]. Molecules known to be associated respectively

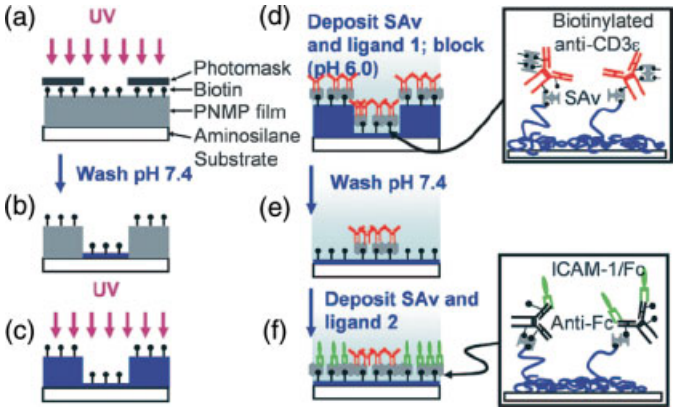


Figure 11.2 Schematic of the photolithographic production of surfaces patterned with two proteins on a micrometer scale, using a novel photoresist that can be processed in biological buffers [19]. A detailed description is provided in the text. Briefly: (a) A resist layer created by spincoating onto a cationic substrate was UV-irradiated through a photomask; (b) A wash at pH 7.4 removed all but a molecularly thin layer of resist from the irradiated areas; (c) The sample was uniformly UV-irradiated; (d) Biotinylated anti-CD3 was bound to the biotin functional groups in the resist layer via streptavidin (SAv); (e) Washing at pH 7.4 removed all but a molecularly thin layer of resist from the originally unirradiated areas; (f) ICAM-1-Fc was bound to the biotin functional groups in the resist layer via biotinylated anti-Fc and streptavidin. (Reproduced from *Proc. Nat. Acad. Sci. U.S.A.* **2006**, *103*, 5700 [19]. Copyright 2006 National Academy of Sciences, U.S.A.)

with LFA-1 (talin) and TCR (protein kinase C, θ isoform (PKC-θ)) signaling localized in the 'pSMAC'- and 'cSMAC'-like regions respectively, in accordance with some previous observations of T cell–APC contact zones [21]. T cells that formed these model 'cSMAC–pSMAC' structures showed elevated levels of intracellular free calcium (an early sign of activation), and eventually proliferated and showed increased cytokine production with respect to cells on control surfaces, thus confirming that full activation had taken place.

Although the cSMAC–pSMAC–dSMAC model corresponds well to *in vivo* images of some T cell–APC contacts (notably where a so-called B cell is used as the APC [21]), when the important dendritic cell APC type is used, a different structure is seen, which may be conceptualized as a multifocal pattern with several smaller cSMAC-like zones. This type of pattern, which could conceivably arise from the dendritic cell cytoskeleton's imposing constraints on the mobility of TCR [16], was reproduced by Doh and Irvine [19], by using their photolithographic technology to produce groups of four small anti-CD3 spots (spots 2 μm diameter, spot centers placed at corners of a ~5 μm cube). The T cells that encountered such groups indeed formed multifocal contacts.

Similar multifocal contacts have also been produced by Mossmann *et al.*, who used electron-beam lithography (EBL) to produce chromium 'walls' (about 100 nm wide and 5 nm high) on a silicon dioxide substrate on which a lipid bilayer containing ICAM-1 and p-MHC was then deposited. In this way, the ICAM-1 and p-MHC were able to move freely laterally up to, but not through, the walls. The p-MHC molecules could thus be confined to several large regions, resulting in the formation of a multifocal pattern with several miniature cSMAC-like regions [15, 16], as shown in Figure 11.3.

The T-cell structures produced on the anti-CD3-patterned substrates of Doh and Irvine [19] might be thought of as a good model of an activating T cell, with the cSMAC as the principal source of activation signaling from ligated TCR. However, as will be seen below, evidence has emerged which suggests that the cSMAC is not an important source of activation signaling, which rather comes primarily from TCR–p-MHC microclusters, and the physiological relevance of the model substrate of Doh and Irvine [19] may be questioned in this respect. It is possible that the interfacial line between anti-CD3- and ICAM-1-coated areas in the studies of Doh and Irvine [19] served the same function as the dSMAC in immunological synapses formed on B cells on supported planar bilayers. The important generalization is that T cells may be highly adaptable as part of their evolution to navigate a wide variety of anatomic sites and interact with essentially any cell in the body to combat continually evolving pathogens. Hence, one important role of nanotechnology may be to test the limits of adaptability and understand the fundamental recognition elements and how they may be manipulated.

11.2.2
TCR–p-MHC Microclusters as Important Signaling Centers

In the model shown in Figure 11.1, an early stage of immunological synapse formation was the ligation of TCR by p-MHC in the peripheral area around the

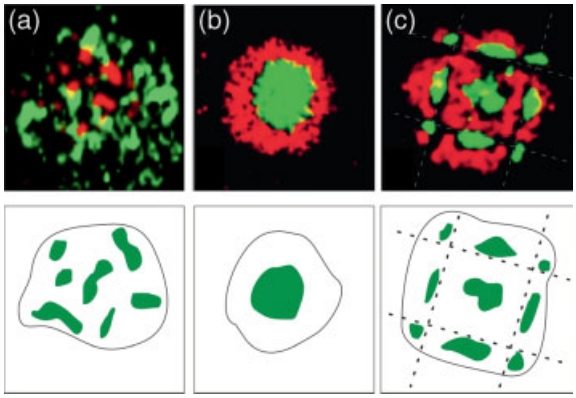


Figure 11.3 Fluorescence micrographs of immunological synapses, with TCRs labeled green (top row) and schematics (bottom row, green show TCR locations, solid back lines show cell outline). (a) A fixed synapse between a T cell and a dendritic cell (red shows PKC-θ, which is not important for our purposes here). TCRs gather at several separate focal points; (b) A synapse formed between a T cell and a supported lipid bilayer, where the lateral motion of p-MHC and ICAM-1 in the bilayer is unconstrained; red shows ICAM-1. TCRs gather in one large cluster, the cSMAC; (c) A synapse formed between a T cell and a supported lipid bilayer where the lateral mobility of p-MHC and ICAM-1 in the bilayer is constrained by chromium 'walls', indicated by dashed lines (white in micrograph, black in schematic) . (ICAM-1 is labeled red.) TCRs gather at several focal points. Note the presence of multiple TCR foci in (a) and (c), which suggests that a reduced lateral mobility of p-MHC on the dendritic cell surface might be responsible for the multifocal nature of the T cell with dendritic cell synapse (a). The central square formed by the chromium 'walls' in (c) has dimensions $2 \times 2\,\mu m$. (Micrographs reproduced from *Curr. Opin. Immunol.*, *18*, 512–516, M.L. Dustin, S.Y. Tseng, R. Varma, G. Campi, T cell-dendritic cell immunological synapses. Copyright (2006), with permission from Elsevier [16]; panels (b) and (c) are originally from **2005**, *310*, 1191–1193. Reprinted with permission from AAAS [15].)

central region of initial ICAM-1 adhesion, with the resulting TCR–p-MHC pairs then migrating to the center of the contact zone to finally form the cSMAC. A closer examination of this system in fact shows the TCR–p-MHC pairs combining to form microclusters throughout the contact zone, which then combine to form the cSMAC [22]. By using highly sensitive total internal reflection fluorescence microscopy (TIRFM), it has also proved possible to image a subsequent continuous 'rain' of microclusters, each consisting of between approximately five and 20 p-MHC–TCR pairs, that form in the peripheral dSMAC region, and then move radially inwards eventually joining the cSMAC [12, 23]. This motion is likely to occur because the TCR are indirectly connected to actin filaments, which are moving continuously inwards in the dSMAC and pSMAC (as discussed in Section 11.2.1). Experiments in which an antibody that disrupts TCR binding to p-MHC was added after the T cells had formed a stable cSMAC on a lipid bilayer surface (such that at early times the formation of new microclusters was prevented but the cSMAC was not yet disrupted) suggest strongly that activation signaling arises from the microclusters rather than from the cSMAC, as signaling almost completely ceased at a time when the cSMAC was still intact [12]. It therefore seems likely that, rather than functioning as a signaling device, the cSMAC in fact plays other roles. In particular, it has been observed that significant numbers of

TCR are endocytosed in the cSMAC. Some of these may be recycled through the cell for reincorporation into the dSMAC, ensuring a continuous supply of TCR and thus enabling signaling to continue for a long time, while others may be degraded [14, 24].

If TCR–p-MHC microclusters arising in the peripheral dSMAC give rise to activation signaling, which switches off as they join the cSMAC, there are two possible hypotheses: the initial signaling may decrease either with time, or with proximity to the center of the contact zone. It proved possible to distinguish between these hypotheses by using the chromium 'walls' of Mossmann *et al.* to divide the contact zone into many small areas, thus preventing microclusters from moving large distances. Using this approach, it was shown that, while the signaling from each microcluster has a finite lifetime, such lifetime decreased strongly with proximity to the center of the contact zone. This showed that spatial factors do play a role, and help to confirm the picture of the cSMAC as a non-signaling region [15].

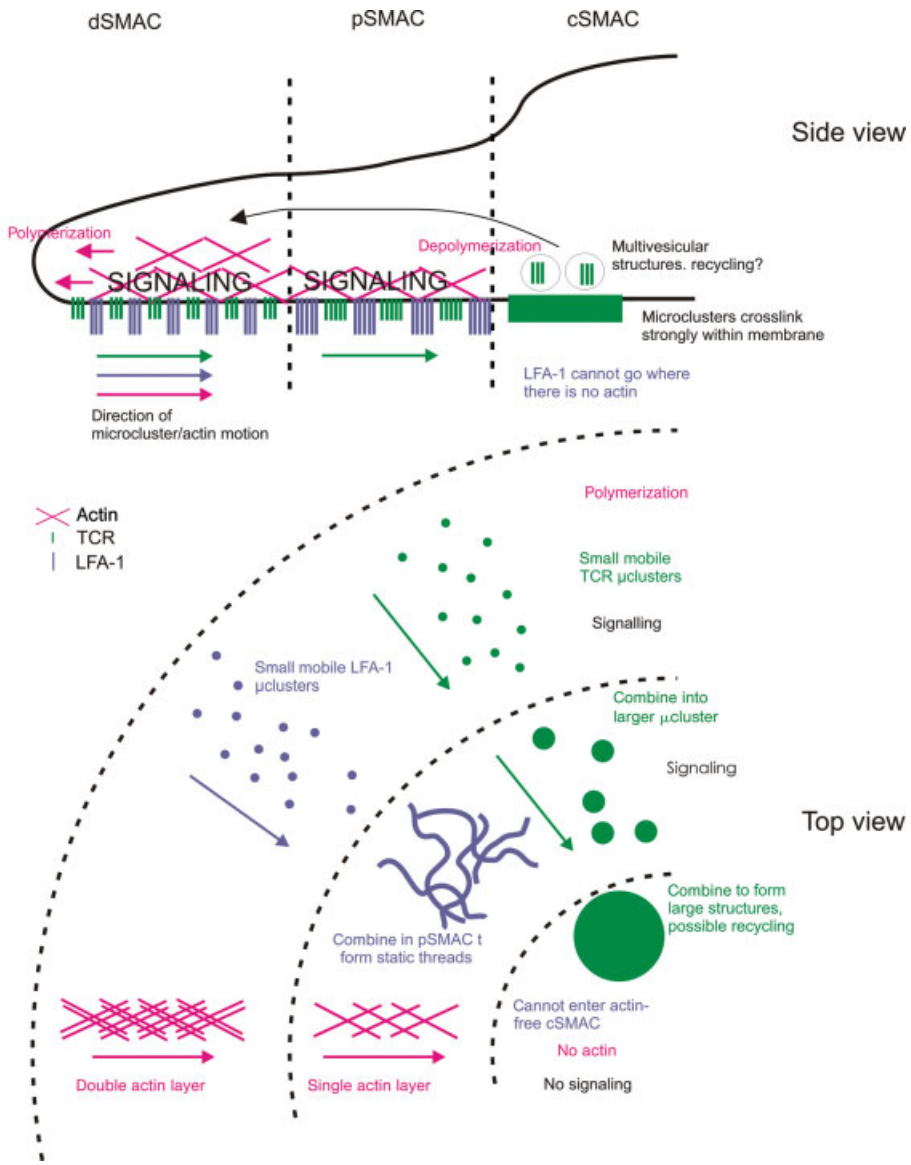
In contrast to the studies just mentioned, in which TCR–p-MHC microclusters formed spontaneously by the coming together or pulling together of mobile p-MHC molecules in a lipid bilayer [12, 15], Anikeeva *et al.* effectively created artificial TCR–p-MHC microclusters by exposing T cells to a solution of *quantum dots*. These are fluorescent semiconductor nanocrystals, to which p-MHC molecules have been bound, with the binding mechanism being the ligation of zinc ions on the semiconductor surfaces by carboxylic acid groups belonging to six histidine residues inserted at the base of the p-MHC molecule [25]. Approximately 12 p-MHC molecules were found per quantum dot, as determined by the measurement of nonradiative energy transfer between the quantum dot and fluorophores bound to the p-MHC molecules. This suggested that the number of ligated TCRs in the artificial microclusters might have been about six, comparable to the size of the smallest signaling microclusters observed in one of the previously mentioned bilayer studies [12]. The stimulation of T cells with appropriate p-MHC-functionalized quantum dots caused activation signaling to occur. Although this study does not relate directly to our theme of T-cell activation by artificial substrates, we mention it here because it is indicative of the potential value of studies performed using p-MHC molecules bound to nanospheres. In fact, it will be seen below that surfaces functionalized with nanospheres may play an important role in future studies.

In addition to TCR–p-MHC microclusters, LFA-1–ICAM-1 microclusters have also been observed; the latter seem to form in the dSMAC and to move inwards before eventually joining thread-like LFA-1–ICAM-1 structures in the pSMAC [18]. Figure 11.4 summarizes schematically the localization of TCRs, p-MHCs, LFA-1 and ICAM-1 in the three SMACs. The structure of the cytoskeletal actin in these regions, as discussed in Section 11.2.1, is also shown.

11.3
The Smallest Activating Units? p-MHC Oligomers

As discussed in Section 11.2, signaling from p-MHC-ligated TCRs seems to depend on the ligated TCRs coming together to form microclusters, rather than ligation alone

being enough for signaling. It transpires that activation signaling can indeed not be initiated by a single ligated TCR, but that the coming together of at least two ligated TCRs is necessary for signaling. This was demonstrated by Boniface *et al.*, who combined biotin-functionalized p-MHCs with naturally tetravalent streptavidin molecules to produce p-MHC monomers, dimers, trimers and tetramers. T cells exposed to the p-MHC monomer showed no activation signaling, whereas significant signaling was already present in the case of the dimer, and the strength increased



when trimer or tetramer was used [26]. This suggests that some degree of TCR 'clustering' is necessary for T-cell activation signaling. This could conceivably indicate that part of the signaling mechanism requires the close proximity of the cytoplasmic parts of neighboring TCRs.

Interestingly, doubt was cast on the finding that TCR clustering is required for the activation signal when, in an experiment using APCs in which all of the agonist p-MHC molecules had been fluorescently labeled, activation signaling was observed to be initiated by a T cell where the contact zone with the APC surface contained only one agonist p-MHC molecule [27]. This apparent contradiction may have been resolved by Krogsgaard *et al.*, who obtained a T-cell activation signal by stimulating cells with a synthetic heterodimer consisting of one MHC molecule with agonist peptide and one with self-peptide (i.e. peptide found within the proteome of the T cell-producing mammal, in this case a mouse) [28]. Krogsgaard *et al.* argued that such heterodimers may play a role in *in vivo* activation. This controversy and its resolution underlines the roles that molecules other than agonist p-MHC may play in *in vivo* T-cell activation; one of the principal advantages of experiments performed on artificial substrates is that 'clean' experiments can be performed, without the possible intrusion of unknown ligands. The ability of T cells to respond to mixed stimulations by agonist and self-peptide-loaded MHC molecules may be important for the functioning of the immune system, as it could increase the likelihood of T-cell activation by APC surfaces that present only small amounts of agonist peptide [8].

If TCR clustering is indeed required for T-cell activation, then it is interesting to ask how close together the TCRs need to be drawn in order for signaling to occur. A significant contribution towards answering this question was made by Cochran *et al.*, who used p-MHC molecules genetically engineered to contain free cystine residues to produce p-MHC dimers; the dimers were created by reacting the cystine residues

Figure 11.4 Summary of our current understanding of the structure of the immunological synapse. Schematic top and side views of the T cell only are provided here: it is assumed that all or most TCRs and LFA-1 molecules shown are ligated by p-MHCs or ICAM-1 on the opposing APC surface (not shown). In the top view, TCRs, LFA-1 and actin are shown in separate locations purely for visual clarity. In the dSMAC, which is analogous to a lamellapodium in a migrating cell, and thus contains two stacked layers of cytoskeletal actin, microclusters of TCR–p-MHC and LFA-1–ICAM-1 form and are transported towards the cell center by the inward motion of actin filaments, as indicated by arrows below the cell. The actin filament motion is caused by depolymerization at the edge of the cSMAC, and polymerization at the edge of the cell. The direction of actin filament growth is indicated by arrows within the cell. In the pSMAC, TCR–p-MHC microclusters merge to form somewhat larger microclusters, and continue migrating inwards, whereas LFA-1–ICAM-1 microclusters merge into a thread-like structure of mutually associated LFA–ICAM-1 pairs, and thus cease moving. The pSMAC is analogous to a lamella in a migrating cell, and thus contains only a single layer of cytoskeletal actin. In the cSMAC, TCR–pMHC pairs merge into a large mass of mutually associated TCR–pMHC pairs. Significant quantities of TCRs are endocytosed by the T cell, and some of these may be recycled through the cell back to the dSMAC, where the process can begin again, enabling signaling to be maintained over a long period of time. LFA-1–ICAM-1 pairs do not enter the actin-free cSMAC in significant quantities. TCR–pMHC microclusters in the dSMAC and pSMAC participate in T-cell activation signaling; there is no signaling due to TCR–p-MHC in the cSMAC.

with maleidomide groups on polypeptide crosslinkers of various lengths. The activation response from T cells decreased as the length of the spacer between the bases of the p-MHC molecules was increased from <1 to 9 nm [29].

11.4
Molecular-Scale Nanolithography

It can be seen from the above discussions that the clustering of p-MHC-ligated TCRs is critical to the initiation of T cell activation signaling. In this section, we describe possible future experiments aimed at further examining these effects, using the technology of *block copolymer micellar nanolithography*. This has recently become available, and enables surfaces to be patterned on the nanometer scale with single-protein molecules such as p-MHCs. Here, we will describe the technique in detail and review its previous uses in cell signaling studies. We will also discuss how the technique can be used, in combination with chemistry and protein engineering methods, to perform experiments to further our understanding of the immunological synapse.

11.4.1
Block Copolymer Micellar Nanolithography

The concept of block copolymer nanolithography is illustrated in Figure 11.5 [30–32]. Here, poly(styrene block 2-vinyl pyridine) forms a micellar solution in toluene with the hydrophilic 2-vinyl pyridine (2VP) block making up the micelle core. Hydrogen tetrachloroaurate (III) ($HAuCl_4$) is added, and complexes the 2VP, producing a gold-rich micelle core. When a flat substrate with a chemically suitable flat surface such as silicon oxide is immersed in the micellar solution and then withdrawn, the approximately spherical micelles form a two-dimensional (2-D) close-packed array on the substrate surface, with the capillary force due to the retreating toluene interface possibly playing a role in forcing them into this configuration. The micelle-coated surface is then exposed to a hydrogen plasma; this removes the polymeric material and reduces the gold ions, producing metallic gold particles at the former sites of the micelle cores. The result is a hexagonal array of gold particles on a (usually) silicon dioxide background. The size of the gold particles can be controlled between approximately 3 nm and 10 nm by varying the amount of added ($HAuCl_4$), while the spacing between adjacent particles can be controlled by varying the length of the styrene block of the original diblock copolymer, and to some extent also by varying the speed of withdrawal of the substrate from the micellar solution. Present investigations have produced interparticle spacings in the range of approximately 15–250 nm, although it seems likely that spacings below 15 nm should also be achievable.

For experiments to study the stimulation of biological cells, the gold nanoparticles can be functionalized with biological ligands using thiol chemistry, while the silicon dioxide surface in between the spheres can be differently functionalized using silane

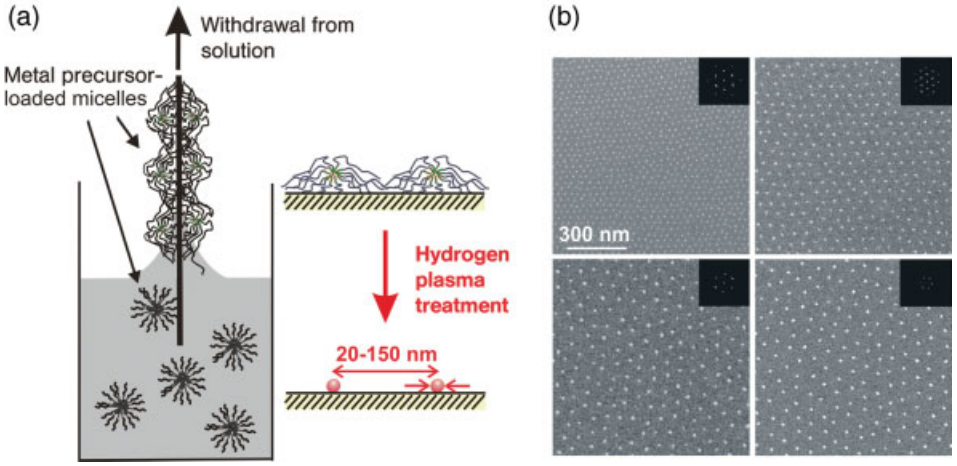


Figure 11.5 Block copolymer micellar nanolithography. Full details are provided in the text. Briefly: (a) Schematic: Micelles of which the cores are loaded with gold ions form a 2-D close-packed layer on a suitable substrate surface that is withdrawn at a suitable speed from the micellar solution. Treatment with a hydrogen plasma removes organic material, resulting in a hexagonally ordered array of gold nanoparticles, with the interparticle spacing being determined by the original polymer block molecular weights and the speed of withdrawal from the solution; (b) Scanning electron microscopy images of surfaces patterned with gold nanoparticles produced using diblock copolymers, where the two blocks have various different molecular weights: the variation in the lattice parameter can be seen. The ordered nature of the patterns is demonstrated by the sharp peaks in the numerically calculated 2-D Fourier transforms of the images (insets at top right of images). ((a) is from R. Glass, M. Arnold, J. Blummel, A. Kuller, M. Moller, J.P. Spatz: Micro-Nanostructured Interfaces Fabricated by the Use of Inorganic Block Copolymer Micellar Monolayers as Negative Resist for Electron-Beam Lithography. *Adv. Funct. Mat.* **2003**, *13*, 569–575 [39]. (b) is from M. Arnold, E.A. Calvalcanti-Adam, R. Glass, J. Blummel, W. Eck, M. Kantlehner, H. Kessler, J.P. Spatz: Activation of Integrin Function by Nanopatterned Adhesive Interfaces. *ChemPhysChem* [2] **2005**, *5*, 383–388. (a) (b) copyright Wiley-VCH Verlag GmbH & Co. KGaA. Reproduced with permission.)

chemistry. The power of this technique is well illustrated by the studies of Arnold *et al.*, who functionalized the gold nanoparticles with cyclic arginine-glycine-aspartate (RGD) peptide molecules that were bound to the gold via a thiol-functionalized linker [2]. These RGD peptides bind strongly to $\alpha_v\beta_3$ integrins, which are membrane-bound receptors that play an important role in the initiation of adhesion by fibroblasts to the ECM. The large size of $\alpha_v\beta_3$ integrins ensured that only one integrin could bind to each gold particle, so that the interparticle spacing could be used as a measure of the minimum separation between adjacent ligated integrin molecules. Experiments showed that fibroblasts adhered readily to substrates with an interparticle spacing of 58 ± 7 nm or less, but did not adhere to substrates with an interparticle spacing of 73 ± 8 nm or more. This suggested that some clustering of ligated integrins is necessary for the initiation of adhesion signaling in fibroblasts, and that the critical spacing below which integrins may be considered to be 'clustered' lies between 58 nm and 73 nm. Additionally, actin-rich protein clusters known as *focal adhesions* that form at sites of $\alpha_v\beta_3$ integrin-mediated adhesion, and which may be considered

as local indicators of adhesion signaling, were observed to form only when the interparticle spacing was higher than this critical value.

It is important to note that, while other lithographic techniques such as EBL [3] and dip-pen nanolithography [33–35], have been used to immobilize biological ligands on surfaces, to the best of our knowledge, block copolymer micellar lithography is the only lithographic method that has thus far been used to spatially isolate individual ligand receptor interactions. This is most likely due to its ability to reliably produce particle sizes as small as 3 nm. This limit compares favorably with, for example, the lower size limit for reliable structure production using EBL with conventional poly (methyl methacrylate) (PMMA) resists, which is about 10 nm [36], and the smallest protein feature that has been created to date using dip-pen nanolithography, which is about 25–40 nm [33].

In view of the apparent requirement for the clustering of ligated TCRs if T-cell activation signaling were to occur, it would clearly be very interesting to perform an analogous experiment to that just described [2], but to study TCR rather than $\alpha_v\beta_3$ integrin clustering. In order to perform such an experiment, each gold nanoparticle would need to be functionalized with a single molecule of p-MHC, and the effect of the interparticle spacing on the activation signaling behavior of T cells brought into contact with such surfaces determined. If TCR clustering were indeed necessary for T-cell activation signaling, then one would expect to observe no signaling when the interparticle spacing was high, with signaling perhaps onsetting as the interparticle spacing was reduced below a critical value. Indeed, the above-mentioned studies of Cochran *et al.* suggested that this spacing should range between 1 and 15 nm [29].

The binding of p-MHCs to the gold dots could be achieved by creating a recombinant MHC construct containing an appropriately located free cystine residue that could react directly with the gold. Alternatively, protein constructs containing multiple consecutive histidine residues have been successfully bound to gold nanospheres on block copolymer micellar nanolithography-patterned surfaces by binding thiol-functionalized nitrilotriacetic acid (NTA) molecules, and allowing the carboxylic acid groups of the NTA and histidine to simultaneously coordinate the same nickel cation. Functionalization of the silicon dioxide surface between the gold nanospheres should also be considered. In the integrin-clustering experiments of Arnold *et al.*, the area between the gold nanoparticles was functionalized with protein-repellent PEG molecules that were end-functionalized with trimethoxysilane groups (this enabled them to bind covalently to the silicon dioxide surface). This PEG functionalization ensured that the cells under study interacted with the surface only via receptor interactions with ligand-functionalized gold particles, and not via nonspecific attractions, as well as resisting the deposition of cell-secreted proteins onto the silicon dioxide surface [2]. In the context of experiments to study T-cell activation, it might be advantageous to functionalize the area between the gold nanoparticles with a combination of PEG molecules to reduce the effect of nonspecific cell-surface attractions, and ICAM-1, to bring about the LFA-1-mediated cell adhesion that is a critical feature of the immunological synapse. This could be achieved by incor-

porating functional groups into the PEG layer that could be bound specifically to ICAM-1; an example would be to incorporate biotin groups into the PEG layer that bind via streptavidin to biotinylated ICAM-1 molecules. The incorporation of biotin into surface-grafted PEG layers has been achieved [37, 38]; I.E. Dunlop *et al.*, unpublished results], and its incorporation into the PEG layer between the gold nanoparticles should be readily achievable.

11.4.2
Micronanopatterning by Combining Block Copolymer Micellar Nanolithography and Electron-Beam Lithography

Surfaces that are structured on both the micrometer and nanometer scales can be produced using a method that combines block copolymer micellar nanolithography with EBL.

The principle of the method is shown schematically in Figure 11.6 [39]. After having deposited a close-packed monolayer of $HAuCl_4$-loaded block copolymer micelles onto the substrate, part of the layer is exposed to an electron beam, which causes the polymer molecules to become highly crosslinked. The substrate is then rinsed with acetone to remove all noncrosslinked polymer, leaving micelles behind only in the area that was exposed to the electron beam. These micelles are then exposed to a hydrogen plasma, which leads to hexagonally arranged gold nanoparticles in the normal manner. It is thus possible, by using a steerable electron beam (such as that in the scanning electron microscope) to pattern only parts of a surface using block copolymer micellar nanolithography, and thus to produce patches of patterning containing controlled numbers of gold nanoparticles (see Figure 11.6).

Surfaces prepared using this method could enable experiments that address questions relating to the number of p-MHC molecules or clusters required for T-cell activation signaling to be addressed. For example, if p-MHC-ligated TCR dimers are sufficient to cause T-cell activation signaling, then it is interesting to ask whether one dimer would be sufficient to produce a detectable activation signal, as suggested by the cell–cell contact experiments of Irvine *et al.* [27], and how the signaling strength would depend on the number of dimers with which a cell interacts. Additionally, the suggestion of Varma *et al.*, that signaling might arise primarily from microclusters of between roughly five and 20 p-MHC-ligated TCRs, suggests that it would be interesting to examine the effect on signaling intensity of the microcluster size, and the number of microclusters per cell. Simulated 'microclusters' containing precisely controlled numbers of molecules could be produced using patches of gold nanoparticles, similar to that shown in Figure 11.6; here, each gold nanoparticle could bear one p-MHC molecule and the interparticle spacing could be chosen sufficiently small that a T cell could 'see' the resulting ligated TCRs as being clustered. Alternatively, microclusters could be simulated by allowing several p-MHCs to bind to one larger nanosphere, as in the experiments of Anikeeva *et al.* mentioned above, in which p-MHCs were bound to soluble quantum dots [25].

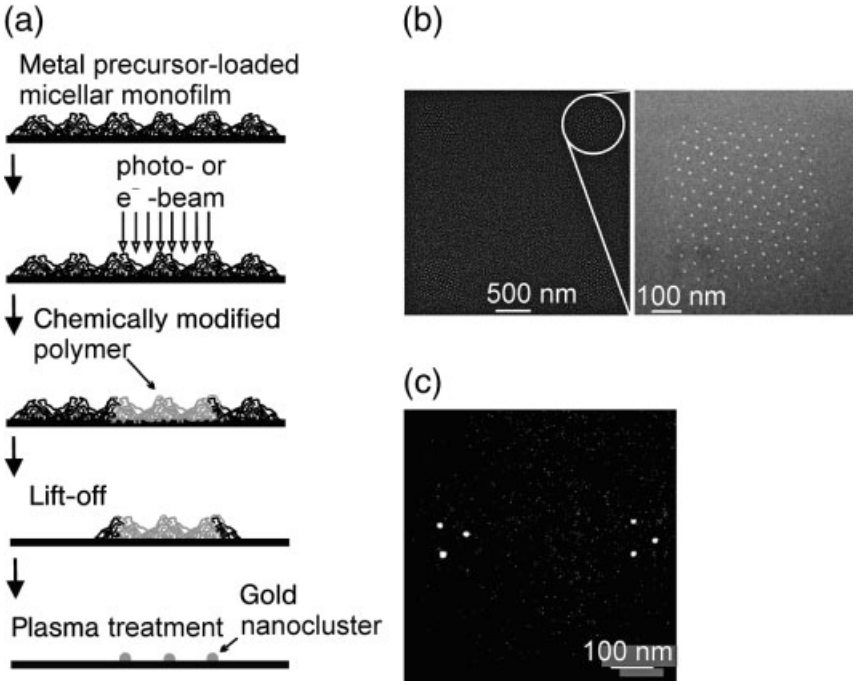


Figure 11.6 Block copolymer micellar nanolithography combined with electron beam lithography to produce patches of substrate nanopatterned with controlled numbers of nanoparticles. (a) Schematic: parts of a close-packed film of gold-loaded micelles are irradiated with a steerable electron beam. Washing in acetone 'lifts off' unirradiated polymer before plasma treatment produces hexagonal arrays of gold particles in the treated areas; (b) (c) Scanning electron microscopy images of different surfaces produced using this method, showing the possibility of producing widely spaced patches of patterned surface, where each patch contains a similar number of gold nanoparticles. ((a) Adapted from R. Glass, M. Arnold, J. Blummel, A. Kuller, M. Moller, J.P. Spatz: Micro-Nanostructured Interfaces Fabricated by the Use of Inorganic Block Copolymer Micellar Monolayers as Negative Resist for Electron-Beam Lithography. *Adv. Funct. Mat.* **2003**, *13*, 569–575. Copyright Wiley-VCH Verlag GmbH & Co. KGaA. Reproduced with permission [39]. (b) (c) Adapted from *Methods Cell Biol.*, *83*, 89–111, J.P. Spatz and B. Geiger: Molecular engineering of cellular environments: Cell adhesion to nano-digital surfaces. Copyright (2007), with permission from Elsevier [53].)

11.5
Therapeutic Possibilities of Immune Synapse Micro- and Nanolithography

So far, we have mostly considered the use of micro- and nanolithographically patterned T-cell-activating surfaces as tools to investigate the functioning of the immune synapse. It is also interesting to consider the potential of these technologies for direct use in clinical therapies. Although it is possible to imagine incorporating T-cell-activating surfaces into medical implants, in order to encourage a specific and local immune response (e.g. against a tumor), it is likely that the first therapeutic use of such surfaces will be in the context of *adoptive cell transfer*.

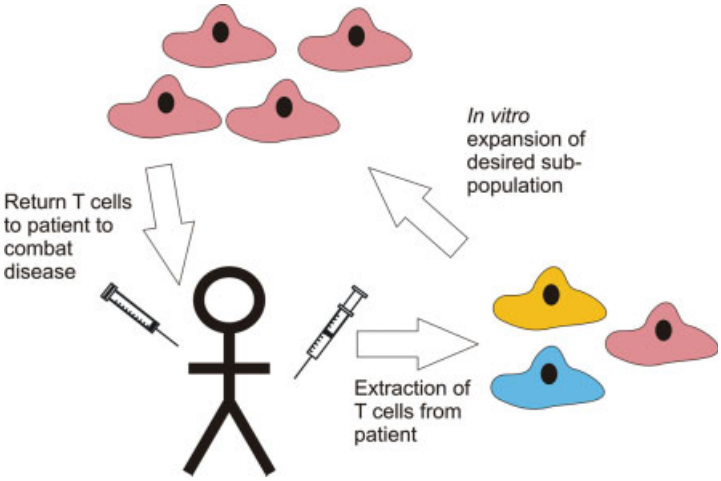


Figure 11.7 The principle of adoptive T-cell therapy.

The principle of the techniques is shown schematically in Figure 11.7. Here, T cells are removed from a patient and a selected subpopulation is deliberately activated, causing it to expand *in vitro*, before being returned to the patient's body. The returned T cells should then produce a strong immune response to the disease being treated. Although adoptive T cell transfer may prove useful in combating certain viral infections, much research has focused on the treatment of cancerous tumors, where the subpopulation of cells to be expanded should clearly be selected to be responsive to tumor-related antigens (for reviews, see Refs [40–42]). One approach is to use extracted tumor cells directly to selectively activate T cells of an appropriate specificity; this leads to a population of T cells that are specific for a variety of epitopes contained in the tumor [43]. Alternatively, epitopes that are known to be tumor-associated can be chosen, and T cells that are specific to those epitopes activated using artificial MHC–peptide constructs [40]. Here, we will focus on the second approach, as micro- and nanopatterned biomimetic surfaces functionalized with p-MHCs and costimulatory molecules might be of value in this context.

The identification of tumor-specific antigens is key if the adoptive cell therapy is to target a tumor, without damaging the healthy tissue: this approach is of particular value in tumors that are virus-induced, where antigens derived from viral proteins can be used [44]. Equally, many tumors express significantly mutated proteins that could be targeted, although the individual genetic analysis of a patient's tumor could prove expensive [41, 45]. Alternatively, antigens can be chosen from proteins that are known to be overexpressed in particular types of tumor, or even from healthy but tissue-specific proteins in tissues that are not necessary for survival, such as the prostate gland [41, 46].

In order to activate T cells using synthetic p-MHCs it is not necessary to use sophisticated spatially patterned substrates; the p-MHCs could simply be bound to a surface with no control of its spatial distribution. However, the results of recent studies of adoptive cell therapy have emphasized that T-cell activation is not a simple

on-off event; depending on the details of the activation method, as well as the prior history of the T cell, a huge variety of subtly different phenotypes can be obtained. Moreover, the differences between these phenotypes can determine the outcome of treatment [40]. An important factor here is the strength of T-cell stimulation with p-MHCs. T cells that are fairly strongly stimulated tend to differentiate to an *effector* or *effector memory* T-cell phenotype which will combat infection but not give rise to a long-lived population of T cells *in vivo*. In contrast, less-strongly stimulated cells tend to differentiate to longer-lived *central memory* T cells, which are more likely to act as progenitors of a large, long-lived T-cell population [40, 47]. It has been suggested that adoptive cell therapy can be more effective if central memory, rather than effector or effector memory, T cells are used [48]. P-MHC micro- and nanopatterned surfaces could clearly be used to control the activation 'dose' delivered to each T cell by, for example, producing spatially separated activating 'patches', each of which contains a given number of p-MHCs, along with appropriate adhesion molecules and cofactors. As discussed above, the spatial distribution of p-MHCs on an activating surface can play a role in determining immunological synapse structure and also the degree of T-cell activation; spatially structured p-MHC-functionalized surfaces may therefore be of use in controlling the phenotype of T cells used for adoptive cell therapy.

A number of other factors, in addition to the nature of the p-MHC stimulus, have been identified as important in the preparation of T cells for adoptive cell therapy [40]. For example, it may be necessary to selectively activate either helper or killer T cells, and it is certainly important not to activate regulatory T cells which act to suppress the immune response to the target epitope [42, 49]. Also, certain effects of *in vitro* culture may cause T cells to senesce in ways that resemble the weakening of the immune system on aging, thus reducing their therapeutic effectiveness [40, 50, 51]. Both of these issues have been addressed by activating T cells using costimulatory molecules simultaneously with p-MHCs. Given the spatial structuring of the immunological synapse, using lithographic methods to determine the relative positions of p-MHCs and costimulatory molecules (as described above for p-MHCs and ICAM-1) might well lead to a better control of the final T-cell phenotype. Recently, when using microcontact printing to generate patterns of costimulatory anti-CD28 and TCR-activating anti-CD3, it was shown that multiple peripheral anti-CD28 foci were better than one large central spot with the same amount of anti-CD28 for the stimulation of T-cell interleukin-2 production [52].

To summarize, adoptive cell therapy based on the *ex vivo* activation of T cells shows promise as an anti-cancer therapy, but better control of the detailed phenotype of the activated T cells is desired. Lithographic patterning of activating surfaces with p-MHCs and costimulatory molecules may contribute to attaining this control.

11.6 Conclusions

Studies performed by bringing T cells into contact with artificial surfaces that mimic aspects of the APC surfaces have contributed greatly to our understanding of the

immunological synapse, and such surfaces may be of therapeutic value in the future. Among the most informative experiments have been those performed using substrates bearing lipid bilayers that contain mobile ICAM-1 and p-MHCs. Photolithographic methods have been used to control the mobility of molecules within such bilayers, stimulating possible effects of the APC cytoskeleton and enabling the effects of reduced mobility on signaling by p-MHC-ligated TCR microclusters to be investigated. In separate studies, photolithographic methods that enable the patterning of surfaces with multiple proteins have been used to bring about artificial SMAC-like structures. The importance of studying p-MHC-ligated TCR clustering effects at the nanometer scale is attested to by evidence from several studies in which T cells were stimulated with soluble p-MHC oligomers, and substrates that are patterned with single p-MHC molecules on the nanometer scale will accordingly be required for the next generation of such studies. Block copolymer micellar nanolithography represents a suitable technique for generating such substrates and, when combined with EBL. will enable the production of surfaces patterned on both nanometer and micrometer length scales. T cell activation experiments performed on such substrates are likely to play a role in extending our understanding of the immunological synapse. Both, micro- and nanopatterned substrates may also be used for *ex vivo* T cell activation in the context of T cell adoptive immunotherapy, where T cells removed from a patient are activated and expanded *ex vivo* before being returned to combat disease, notably cancer. The use of these substrates may also help to gain close control of the phenotypes of *ex vivo*-activated T cells, leading to more effective treatments.

Acknowledgments

The authors thank Thomas O. Cameron and Rajat Varma for useful discussions. This chapter was partially supported by the National Institutes of Health through the NIH Roadmap for Medical Research (PN2 EY016586) (I.E.D., M.L.D., J.P.S.) and by the Max Planck Society (I.E.D., J.P.S.). I.E.D. acknowledges a Humboldt Research Fellowship.

References

1 Vogel, V. and Sheetz, M. (2006) *Nature Reviews Molecular Cell Biology*, **7**, 265.

2 Arnold, M., Cavalcanti-Adam, E.A., Glass, R., Blummel, J., Eck, W., Kantlehner, M., Kessler, H. and Spatz, J.P. (2004) *Chemphyschem*, **5**, 383.

3 Cherniavskaya, O., Chen, C.J., Heller, E., Sun, E., Provezano, J., Kam, L., Hone, J., Sheetz, M.P. and Wind, S.J. (2005) *Journal of Vacuum Science & Technology B*, **23**. 2972.

4 Taite, L.J., Rowland, M.L., Ruffino, K.A., Smith, B.R.E., Lawrence, M.B. and West, J.L. (2006) *Annals of Biomedical Engineering*, **34**, 1705.

5 Chen, S.Q., Alon, R., Fuhlbrigge, R.C. and Springer, T.A. (1997) *Proceedings of the National Academy of Sciences of the United States of America*, **94**, 3172.

6 Janeway, C.A.J., Travers, P., Walport, M. and Schlomchik, M.J. (2005) *Immunobiology: The Immune System in*

Health and Disease, 6th edn, Garland Science Publishing, New York, N.Y., USA.

7 Dustin, M.L. and Colman, D.R. (2002) *Science*, **298**, 785.

8 Davis, M.M., Krogsgaard, M., Huse, M., Huppa, J.B., Lillemeier, B.F. and Li, Q.J. (2007) *Annual Review of Immunology*, **25**, 681.

9 Pulendran, B. and Ahmed, R. (2006) *Cell*, **124**, 849.

10 Gourley, T.S., Wherry, E.J., Masopust, D. and Ahmed, R. (2004) *Seminars in Immunology*, **16**, 323.

11 Crotty, S. and Ahmed, R. (2004) *Seminars in Immunology*, **16**, 197.

12 Varma, R., Campi, G., Yokosuka, T., Saito, T. and Dustin, M.L. (2006) *Immunity*, **25**, 117.

13 Grakoui, A., Bromley, S.K., Sumen, C., Davis, M.M., Shaw, A.S., Allen, P.M. and Dustin, M.L. (1999) *Science*, **285**, 221.

14 Lee, K.H., Dinner, A.R., Tu, C., Campi, G., Raychaudhuri, S., Varma, R., Sims, T.N., Burack, W.R., Wu, H., Kanagawa, O., Markiewicz, M., Allen, P.M., Dustin, M.L., Chakraborty, A.K. and Shaw, A.S. (2003) *Science*, **302**, 1218.

15 Mossman, K.D., Campi, G., Groves, J.T. and Dustin, M.L. (2005) *Science*, **310**, 1191.

16 Dustin, M.L., Tseng, S.Y., Varma, R. and Campi, G. (2006) *Current Opinion in Immunology*, **18**, 512.

17 Sims, T.N., Soos, T.J., Xenias, H.S., Dubin-Thaler, B., Hofman, J.M., Waite, J.C., Cameron, T.O., Thomas, V.K., Varma, R., Wiggins, C.H., Sheetz, M.P., Littman, D.R. and Dustin, M.L. (2007) *Cell*, **129**, 773.

18 Kaizuka, Y., Douglass, A.D., Varma, R., Dustin, M.L. and Vale, R.D. (2007) *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 20296.

19 Doh, J. and Irvine, D.J. (2006) *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 5700.

20 Doh, J. and Irvine, D.J. (2004) *Journal of the American Chemical Society*, **126**, 9170.

21 Monks, C.R.F., Freiberg, B.A., Kupfer, H., Sciaky, N. and Kupfer, A. (1998) *Nature*, **395**, 82.

22 Freiberg, B.A., Kupfer, H., Maslanik, W., Delli, J., Kappler, J., Zaller, D.M. and Kupfer, A. (2002) *Nature Immunology*, **3**, 911.

23 Campi, G. Varma, R. and Dustin, M.L. (2005) *Journal of Experimental Medicine*, **202**, 1031.

24 Lee, K.H., Holdorf, A.D., Dustin, M.L., Chan, A.C., Allen, P.M. and Shaw, A.S. (2002) *Science*, **295**, 1539.

25 Anikeeva, N., Lebedeva, T., Clapp, A.R., Goldman, E.R., Dustin, M.L., Mattoussi, H. and Sykulev, Y. (2006) *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 16846.

26 Boniface, J.J., Rabinowitz, J.D., Wulfing, C., Hampl, J., Reich, Z., Altman, J.D., Kantor, R.M., Beeson, C., McConnell, H.M. and Davis, M.M. (1998) *Immunity*, **9**, U7.

27 Irvine, D.J., Purbhoo, M.A., Krogsgaard, M. and Davis, M.M. (2002) *Nature*, **419**, 845.

28 Krogsgaard, M., Li, Q.J., Sumen, C., Huppa, J.B., Huse, M. and Davis, M.M. (2005) *Nature*, **434**, 238.

29 Cochran, J.R., Cameron, T.O., Stone, J.D., Lubetsky, J.B. and Stern, L.J. (2001) *Journal of Biological Chemistry*, **276**, 28068.

30 Spatz, J.P. Sheiko, S. and Moller, M. (1996) *Macromolecules*, **29**, 3220.

31 Spatz, J.P. Roescher, A. and Moller, M. (1996) *Advanced Materials*, **8**, 337.

32 Spatz, J.P., Mossmer, S., Hartmann, C., Moller, M., Herzog, T., Krieger, M., Boyen, H.G., Ziemann, P. and Kabius, B. (2000) *Langmuir*, **16**, 407.

33 Li, B., Zhang, Y., Hu, J. and Li, M.Q. (2005) *Ultramicroscopy*, **105**, 312.

34 Lee, K.B. Lim, J.H. and Mirkin, C.A. (2003) *Journal of the American Chemical Society*, **125**, 5588.

35 Lee, K.B., Kim, E.Y., Mirkin, C.A. and Wolinsky, S.M. (2004) *Nano Letters*, **4**, 1869.

36 Vieu, C., Carcenac, F., Pepin, A., Chen, Y., Mejias, M., Lebib, A., Manin-Ferlazzo, L., Couraud, L. and Launois, H. (2000) *Applied Surface Science*, **164**, 111.
37 Morgenthaler, S., Zink, C., Stadler, B., Voros, J., Lee, S., Spencer, N.D. and Tosatti, S.G.P. (2006) *Biointerphases*, **1**, 156.
38 You, Y..-Z. and Oupicky, D. (2007) *Biomacromolecules*, **8**, 98.
39 Glass, R., Arnold, M., Blummel, J., Kuller, A., Moller, M. and Spatz, J.P. (2003) *Advanced Functional Materials*, **13**, 569.
40 June, C.H. (2007) *Journal of Clinical Investigation*, **117**, 1204.
41 June, C.H. (2007) *Journal of Clinical Investigation*, **117**, 1466.
42 Gattinoni, L., Powell, D.J., Rosenberg, S.A. and Restifo, N.P. (2006) *Nature Reviews Immunology*, **6**, 383.
43 Milone, M.C. and June, C.H. (2005) *Clinical Immunology*, **117**, 101.
44 Straathof, K.C.M., Bollard, C.M., Popat, U., Huls, M.H., Lopez, T., Morriss, M.C., Gresik, M.V., Gee, A.P., Russell, H.V., Brenner, M.K., Rooney, C.M. and Heslop, H.E. (2005) *Blood*, **105**, 1898.
45 Sjoblom, T., Jones, S., Wood, L.D., Parsons, D.W., Lin, J., Barber, T.D., Mandelker, D., Leary, R.J., Ptak, J., Silliman, N., Szabo, S., Buckhaults, P., Farrell, C., Meeh, P., Markowitz, S.D., Willis, J., Dawson, D., Willson, J.K.V., Gazdar, A.F., Hartigan, J., Wu, L., Liu, C.S., Parmigiani, G., Park, B.H., Bachman, K.E., Papadopoulos, N., Vogelstein, B., Kinzler, K.W. and Velculescu, V.E. (2006) *Science*, **314**, 268.
46 Pardoll, D.M. (1999) *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 5340.
47 Sallusto, F. Geginat, J. and Lanzavecchia, A. (2004) *Annual Review of Immunology*, **22**, 745.
48 Klebanoff, C.A., Gattinoni, L., Torabi-Parizi, P., Kerstann, K., Cardones, A.R., Finkelstein, S.E., Palmer, D.C., Antony, P.A., Hwang, S.T., Rosenberg, S.A., Waldmann, T.A. and Restifo, N.P. (2005) *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 9571.
49 Curiel, T.J., Coukos, G., Zou, L.H., Alvarez, X., Cheng, P., Mottram, P., Evdemon-Hogan, M., Conejo-Garcia, J.R., Zhang, L., Burow, M., Zhu, Y., Wei, S., Kryczek, I., Daniel, B., Gordon, A., Myers, L., Lackner, A., Disis, M.L., Knutson, K.L., Chen, L.P. and Zou, W.P. (2004) *Nature Medicine*, **10**, 942.
50 Zhou, J.H., Shen, X.L., Huang, J.P., Hodes, R.J., Rosenberg, S.A. and Robbins, P.F. (2005) *Journal of Immunology*, **175**, 7046.
51 Monteiro, J., Batliwalla, F., Ostrer, H. and Gregersen, P.K. (1996) *Journal of Immunology*, **156**, 3587.
52 Shen, K., Thomas, V.K., Dustin, M.L. and Kam, L.C. (2008) *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 7791.
53 Spatz, J.P. and Geiger, B. (2007) *Methods in Cell Biology*, **83**, 89.

12
Bone Nanostructure and its Relevance for Mechanical Performance, Disease and Treatment

Peter Fratzl, Himadri S. Gupta, Paul Roschger, and Klaus Klaushofer

12.1
Introduction

The human skeleton not only serves as an ion reservoir for calcium homeostasis but also has an obvious mechanical function in supporting and protecting the body. These functions place serious requirements on the mechanical properties of bone, which should be stiff enough to support the body's weight and tough enough to prevent easy fracturing. Such outstanding mechanical properties are achieved by a very complex hierarchical structure of bone tissue, which has been described in a number of reviews [1–3]. Starting from the macroscopic structural level, bones can have quite diverse shapes, depending on their respective function. Long bones, such as the femur or the tibia, are found in the body's extremities and provide stability against bending and buckling. In other cases, for example in the vertebra or the head of the femur, the applied load is mainly compressive, and in such cases the bone shell is filled with highly porous cancellous bone (see Figure 12.1). Several levels of hierarchy are visible in this figure, with trabeculae or osteons in the hundred-micron range (Figure 12.1b and c), a lamellar structure in the micron range (Figure 12.1d), collagen fibrils of 50–200 nm diameter (Figure 12.1e), and collagen molecules as well as bone mineral particles with just a few nanometers thickness.

This hierarchical structure is largely responsible for the outstanding mechanical properties of bone. At the nanoscale, both collagen and mineral – and also their structural arrangement – play a crucial role. In this chapter we review the structure of bone at the nanoscale, and describe some recent findings concerning the influence of bone on deformation and fracture. We also outline some approaches to studying biopsy specimens in diseases and in treatments that are known to influence bone at the nanoscale.

Nanotechnology, Volume 5: Nanomedicine. Edited by Viola Vogel

ISBN: 978-3-527-31736-3

12.2
Nanoscale Structure of Bone

At the nanometer scale, bone is a composite of a collagen-rich organic matrix and mineral nanoparticles made from carbonated hydroxyapatite. The structure and properties of bone have recently been reviewed [2]. The basic building block of the bone material is a mineralized collagen fibril of between 50 and 200 nm diameter (Figure 12.1e and f). Collagen type I is the organic constituent of these fibrils in bone and in many biological tissues, including tendon, ligaments skin and cornea. The collagen molecules are triple helices with a length of about 300 nm, and are assembled within the cell. After secretion, the globular ends of the molecules are cleaved off enzymatically and the (apart from short telopeptide ends) triple helical molecules [4, 5] undergo a self-assembly process that leads to a staggered arrangement of parallel molecules. This in turn creates a characteristic pattern of low-density gap zones that are 35 nm long and high-density overlap zones 32 nm long within the fibril [6]; hence, the effective periodicity (D) will be 67 nm (Figure 12.1f). The collagen fibrils are filled and coated by mineral crystallites [7, 8]; the latter are mainly flat plates [9] that are mostly arranged parallel to each other in a fibril, and parallel to the long axis of the collagen fibrils [10]; however, they may not always be parallel between different fibrils [7]. The crystallites have a periodicity in axial packing density along the fibrils of the same 67 nm dimension [11] by which adjacent collagen molecules are staggered (Figure 12.1f). Crystal formation is triggered by collagen or (more likely) other noncollagenous proteins which act as nucleation centers [12]. After nucleation, the plate-like crystals become elongated but extremely thin [7, 9, 13, 14],

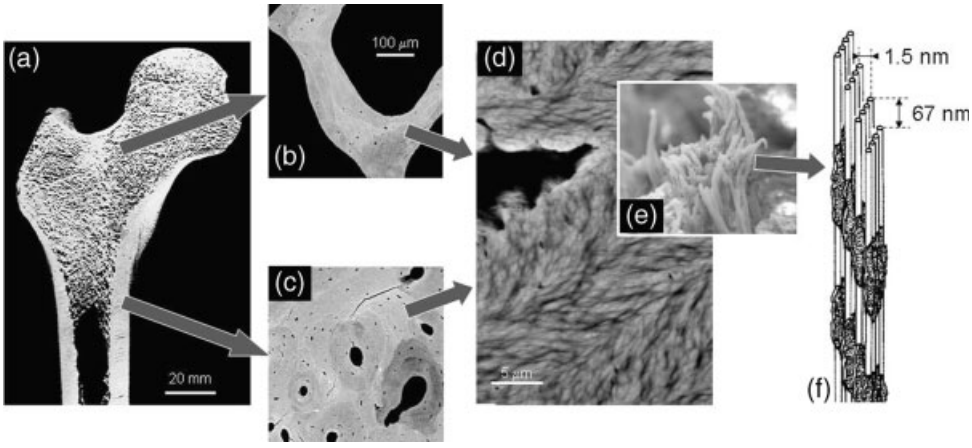


Figure 12.1 Hierarchical structure of a human femur. The femoral head (a) is filled with cancellous bone, consisting of individual trabeculae (b). The cortical bone shell contains osteons (c) where the central Haversian canal is surrounded by concentric lamellae of bone tissue. Lamellar bone (d) consists of thin layers of parallel collagen fibrils with rotating orientation similar to plywood. Collagen fibrils are constituted by parallel collagen molecules with a longitudinal stagger of 67 nm (e) and are reinforced with plate-like mineral particles located inside and on the surface of the fibrils.

and later grow in thickness [15, 16]. Among bone tissues from several different mammalian and nonmammalian species, the bone mineral crystals have thicknesses ranging from 1.5 to 4.5 nm [2, 7, 16–20]. While bone mineral is based mainly on hydroxyapatite ($Ca_5(PO_4)_3OH$), it also typically contains additional elements that replace either the calcium ions or the phosphate or hydroxyl groups; one of the most common such occurrences is replacement of the phosphate group by carbonate [1, 2].

12.3 Mechanical Behavior of Bone at the Nanoscale

The fracture resistance of bone results from the ability of its microstructure to dissipate deformation energy, without the propagation of large cracks leading to eventual material failure [21–23]. Different mechanisms have been reported for the dissipation of energy [24], including: the formation of nonconnected microcracks ahead of the crack tip [25, 26]; crack deflection and crack blunting at interlamellar interfaces and cement lines [27]; and crack bridging in the wake zone of the crack [28–30], which was attributed a dominant role [28].

One striking feature of the fracture properties in compact bone is the anisotropy of the fracture toughness, which differs by almost two orders of magnitudes between a crack that propagates parallel or perpendicular to the collagen fibrils [24]. This results in a zig-zag pattern of the crack path, when it needs to propagate perpendicular to the fibril direction (Figure 12.2a). This dependence of fracture properties on collagen orientation underlines the general importance of the organic matrix and

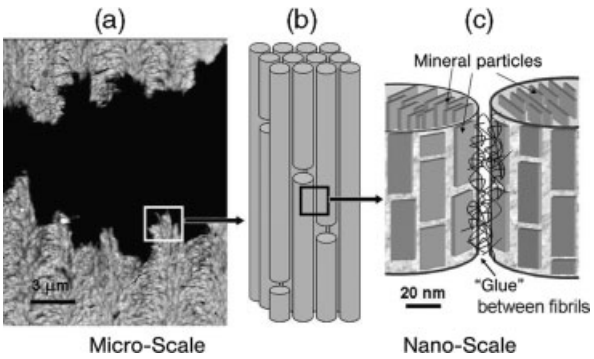


Figure 12.2 Some structural features of bone at the microscale and nanoscale that are responsible for dissipating energy during deformation and fracture. (a) Cracks propagating perpendicularly to the lamellar structure are forced into a zig-zag path, which increases the dissipated energy by about a factor of 30 [24]; (b) Each layer consists of parallel collagen fibrils arranged in a plywood-like structure where the fibril direction rotates along the direction perpendicular to the layer [47, 48]. About one-half of the deformation in a fibril bundle occurs in a glue layer between fibrils [40, 44, 45]. (c) The fibrils are stiffened by mineral particles inside and on the surface of fibrils. The 'glue' layer [43] may contain proteoglycans and phosphorylated proteins, perhaps coordinated by divalent ions, such as calcium [44].

its organization for bone toughness. The organic matrix varies with genetic background, age and disease, and this will clearly influence bone strength and toughness [2, 31–39].

The dominant structural motif at the nanoscale is the *mineralized collagen fibril*. Important contributions to the fracture resistance and defect tolerance of bone composites are believed to arise from these nanometer-scale structural motifs. In recent studies [40], it has been shown that both mineral nanoparticles and the mineralized fibrils deform at first elastically, but to different degrees, in a ratio of 12 : 5 : 2 between tissue, fibrils and mineral particles. These different degrees of deformation of different components arranged in parallel manner within the tissue can be explained by a shear deformation between the components [41]. This means that there is shear deformation within the collagen matrix inside the fibril to accommodate for the difference between the strain in mineral particles and fibrils. In addition, there must also be some shear deformation between adjacent collagen fibrils to accommodate the residual tissue strain. This shear deformation occurs presumably in a 'glue' layer between fibrils (Figure 12.2c), which may consist of proteoglycans and noncollagenous phosphorylated proteins [40, 42–45]. The existence of a glue layer was originally proposed as a consequence of investigations using scanning force microscopy [43]. Beyond the regime of elastic deformation, it is likely that the glue matrix is partially disrupted, and that neighboring fibrils move past each other, breaking and reforming the interfibrillar bonds. An alternative explanation could be the debonding between organic matrix and hydroxyapatite particles (Figure 12.2c) and a modification of the frictional stress between fibril structures [46].

The maximum strain seen in mineral nanoparticles (0.15–0.2%) can reach up to twice the fracture strain calculated for bulk apatite. The origin of this very high strength (~200 MPa) of the mineral particles may result directly from their extremely small size [49]. The strength of brittle materials is known to be controlled by the size of the defects, and of course it can be argued that a defect in a mineral particle cannot be larger than the particle itself. Under such conditions, the strength approaches the theoretical value determined by the chemical bonds between atoms rather than by the defects [49]. Although the nanoparticles in bone are still a way off this value ($\sim E/10$ or 10 GPa), it is believed that the trend towards higher strengths is related to their small size.

As a consequence, it must be concluded that the mechanical properties of bone material are determined by a number of structural features, including:

- the mineral concentration inside the organic matrix, the 'bone mineral density distribution (BMDD)
- the size of mineral particles
- the quality of the collagen, in terms of its amino-acid sequence, crosslinks and hydration
- the quality and composition of the extrafibrillar organic matrix between the collagen fibrils (consisting mostly of noncollagenous organic molecules)
- the orientation distribution of the mineralized collagen fibrils.

Assuming that these parameters are typically optimized in healthy bone material, it is likely that any variation from normal might affect the mechanical performance of the bone. Although these material characteristics cannot typically be determined in a noninvasive manner for patients, they are accessible when studying biopsies using different – and in some cases well-established – technology.

12.4
Bone Mineral Density Distribution in Osteoporosis and Treatments

The mineral concentration inside the organic bone matrix is a major determinant of bone stiffness and strength [2, 33, 50, 51]. However, the mineral content within both, the trabecular and the cortical bone motifs, is far from homogeneous (Figure 12.3). At least two processes that occur in bones over the whole lifetime of an adult individual are responsible for this situation [52]:

- Bone remodeling: The cortical and trabecular bone compartments are continuously remodeled. This means that, during a cycle of about 200 days, areas of bone are resorbed by specific bone cells (*osteoclasts*); this results in *resorption lacunae* which are re-filled with new bone matrix [53] produced by other bone cells (*osteoblasts*). Thus, the bone tissue of an individual adult is on average younger than that adult's chronological age, because the bone turnover time is about

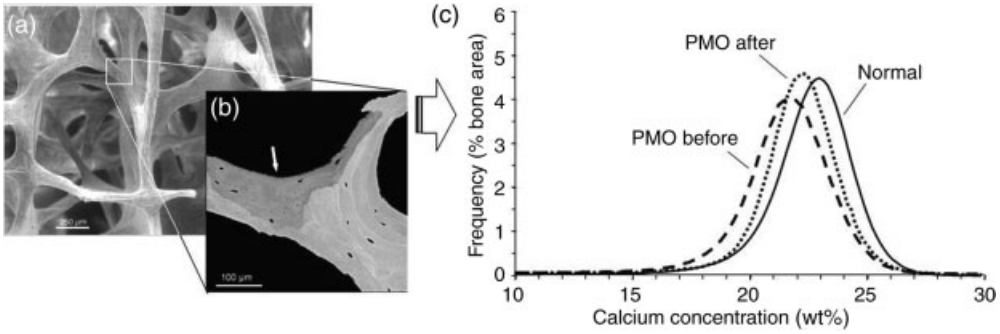


Figure 12.3 Bone is composed of packets of mineralized bone matrix with different mineral concentrations. The distribution of mineral is described by a histogram, called the bone mineralization density distribution (BMDD). (a) Scanning electron microscopy image (secondary electron emission) showing the 3-D structure of trabecular bone; (b) Backscattered electron image of a single trabecula in a bone section, revealing several bone packets with different mineral contents. The dark gray region indicates low mineralization, and the light gray high mineralization of the bone matrix. The arrow shows a newly forming bone packet with a lower mineral content than adjacent packets; (c) Examples of BMDD curves deduced from calibrated backscattered electron images of trabecular bone: NORMAL = healthy individual; PMO before = post-menopausal osteoporotic women before any treatment; PMO after = post-menopausal osteoporotic women after bisphosphonate treatment [57].

five years [54]. In addition, the more such remodeling sites act on the bone surface, the higher will be the bone turnover rate, and more bone packets will be present at a younger stage.

- Kinetics of matrix mineralization: The newly formed bone matrix is initially unmineralized (*osteoid*). However, after an initial maturation time of about 14 days the bone goes through a stage of rapid mineralization, where 70% of the full matrix mineral content is achieved in a few days (*primary mineralization*). Later on, the mineral content increases very slowly to reach full mineralization within years (*secondary mineralization*) [55, 56].

As a consequence of these processes, bone is composed of bone packets – also known as basic structural units (BSUs) – all of which have a different age and mineral content (Figure 12.3). The BSUs generate a characteristic mineralization pattern, sometimes referred to as the bone mineralization density distribution (BMDD) [57], which reflects the bone turnover status and the kinetics of mineralization in an individual [52].

The BMDD can best be measured and quantified using a backscattered electron method (quantitative backscattered electron imaging; qBEI), as described elsewhere [58] and recently reviewed [57]. In contrast to the noninvasive and widely used technique of dual X-ray absorptometry (DXA), which provides an estimate of the total amount of mineral in a scanned area of bone (BMD), the measurement of BMDD requires bone biopsies to be taken. However, the BMDD can also be determined using undecalcified resin-embedded bone blocks, as are prepared for histological examinations. The physical principle of the technique is based on a quantification of the intensity of electrons that are backscattered from a polished bone surface and yield a signal which is proportional to the local concentration of mineral (calcium). Thus, the resulting backscattered electron image visualizes regions of low and high mineral content in dark and light gray, respectively (Figure 12.3b). A suitable calibration of gray levels allows the deduction of frequency distributions of the Ca-concentrations that occur in the scanned bone area (BMDD) with a spatial resolution of 1–4 μm and a sensitivity of 0.17 wt% Ca (Figure 12.3c). The BMDD curve visualizes potential differences in mineralization status of bone between individuals with a high sensitivity (see Figure 12.3c).

With this technique at hand, it has become possible to study the mineral distribution in bone as well as its disease-related changes. The trabecular bone of normal (healthy) adults was found to exhibit minimal variations in BMDD between different skeletal sites, and due to other biological factors such as age, gender or ethnicity. Hence, the BMDD of adult trabecular bone may reflect an evolutionary optimum in bone matrix mineralization as a result of the bone cells' activity and mechanical loading, which most likely represents a compromise between optimum stiffness (which increases with mineral content) and toughness (which decreases with mineral content) of the bone material [2, 3]. It follows that deviations from the normal BMDD, as are observed for example in osteoporosis of post-menopausal women, are most likely of mechanical relevance.

12.4.1 Osteoporosis

Osteoporosis is a disease of enormous socioeconomic impact that is characterized by increased bone fragility [33, 59]. Such fragility is generally associated with an abnormal loss in bone volume, a deterioration in the quality of the bone microarchitecture, an increased bone turnover rate, and also a shift of BMDD towards a lower mineralization density (Figure 12.3c). Interestingly, basic treatment with Ca and vitamin D can have a beneficial effect on bone matrix mineralization and shift the BMDD curve back towards the normal peak position [60, 61]. Additional treatment with antiresorptive agents (e.g. bisphophonates such as alendronate, risedronate or zolendronate) results in a further increase in mineralization, as well as in a higher homogeneity of mineralization within three years of treatment [60, 62, 63]. A prolonged treatment with bisphosphonates over five and 10 years, for example, seemed to restore the BMDD to normal [64]. The treatment effect on BMDD can be explained by a reduction of the remodeling rate, together with a restoration of sufficiently high levels of Ca and vitamin D, thus allowing a more complete mineralization of the BSUs. A combined analysis of bone density (BMD), measured using DXA, and of BMDD as determined by qBEI, revealed that an 8% increase in BMD by bisphosphonate plus Ca/vitamin D treatment was due to a 5% contribution in the improvement of matrix mineralization and a slight (3%) increase in bone volume [61]. The beneficial effect on bone volume might indicate that the therapy also positively affects the negative net balance between bone formation and resorption, as characteristically is the case for osteoporosis. Both effects likely contribute to the sustained anti-fracture efficacy of about 30–50% provided by this anti-resorptive treatment.

Another therapeutic approach in the treatment of osteoporosis is to stimulate bone formation (anabolic treatment). Such an effect on bone can, in principle, be achieved by using sodium fluoride (NaF), which has been used at a daily dose level of 60 mg in several European countries, although it failed to exhibit any anti-fracture efficacy [65]. Interestingly, the BMDD showed a shift to a higher mineralization density with fluoride treatment, and the bone matrix was partly hypermineralized. Changes in the nanocomposite structure of bone (as described in Section 12.5) are most likely responsible for this [66]. Another anabolic agent, parathyroid hormone (PTH), when provided intermittently for a limited period of about 18 months, proved to be successful and has now been approved for the treatment of osteoporosis worldwide. The anabolic effect is clearly reflected in the changes of the BMDD [55], which shows a slight shift to a lower mineralization density and a remarkable broadening of the distribution peak, indicating an increased formation of new bone matrix. This therapy has proven especially useful when the bone loss is already severe. In order to preserve the bone mass gained in such anabolic treatment, possible combinations with anti-resorptive treatments are presently under investigation [67]. It is expected that this might also be beneficial to the mineralization status of the bone, as the prolonged time for secondary mineralization during the anti-resorptive treatment would also normalize the BMDD.

In summary, the BMD in a healthy bone matrix seems to approach an optimum which represents the best compromise between stiffness requiring a high mineral content and toughness, which decreases with mineral content. Biopsies may be used to assess the status of mineral concentration for individual patients at risk of bone fractures in a number of diseases.

12.5 Examples of Disorders Affecting the Structure of Bone Material

As discussed above, the mechanical performance of bone tissue depends on all levels of hierarchy, and several diseases are characterized by modifications at the nanostructural level. In this section, we will detail three examples: (i) *osteogenesis imperfecta*, which is based on mutations of the collagen gene; (ii) *pycnodysostosis*, which originates from a mutation of the cathepsin K gene; and (iii) *fluorosis*, which is caused by higher doses of fluoride. Whilst all three conditions are characterized by a modification at the nanoscale either of the organic matrix or of the mineral particles, adaptation processes during bone remodeling [3] may lead to a partial compensation of the original defect, sometimes at higher hierarchical levels. This means that the modification of bone structure may 'spread' over different hierarchical levels, making it more difficult to pin down the actual origin of the defect.

12.5.1 Osteogenesis Imperfecta

Osteogenesis imperfecta (OI) is a genetic disease that generally affects the collagen gene and leads to brittle bone with different degrees of severity [68–70]. The origin of the brittleness of the tissue is not fully understood, but must be linked to a mutation of the collagen molecule and the resultant changes in tissue quality. Generally, OI also leads to a reduced bone mass and cortical thickness [70] which additionally increases bone fragility. It has been shown that anti-resorptive treatment of affected children with the bisphosphonate pamidronate leads to an increase in cortical thickness and to a concomitant reduction of fracture incidence [70]. At the nanostructural level, an increased mineral content was found in the bone matrix [71, 72], which leads to increased stiffness and hardness of the bone tissue [73]. However, the significance of bone fragility for this increased mineralization is not yet fully clear, as it is not affected by bisphosphonate treatment [73].

More detailed information on the bone matrix nanostructure and the disease-related changes of its properties were obtained in a mouse model of OI [74, 76–82]. This model, which is known as osteogenesis imperfecta murine (oim), is characterized by an absence of the α_2 procollagen molecule, leading to the formation of collagen α_1 homotrimers. The mechanical properties of bone tissue were found to be altered, with a reduced failure load [83] and toughness [82] in oim compared to controls. The mineral content was, however, increased in oim [75], leading to a stiffer matrix [75, 78] (see Figure 12.4). In agreement with observations in humans, this

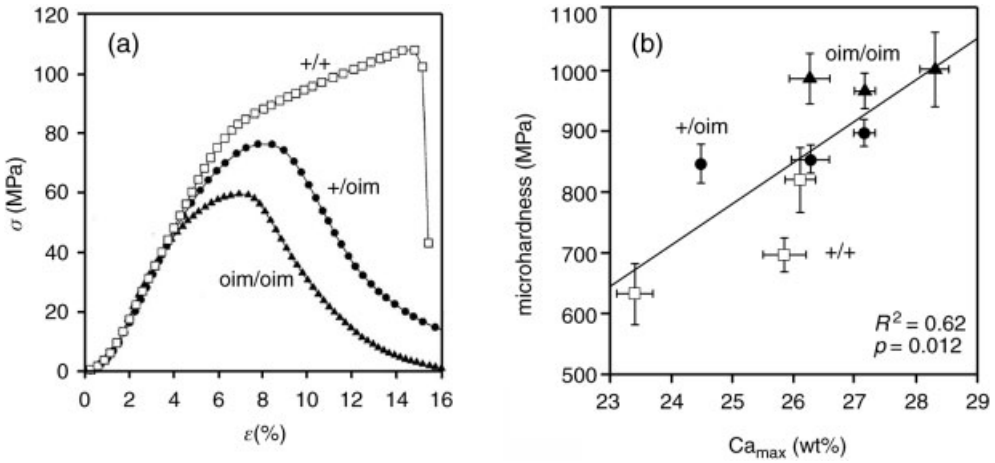


Figure 12.4 Collagen and mineral properties in the osteogenesis imperfecta murine (oim) model. Homozygote oim/oim (full triangles) and heterozygote + /oim (full circles) are compared to normal littermates + / + (open squares). (a) Stress–strain curve of collagen from the mouse tail (from Ref. [74]). The strength of collagen (the maximum stress σ) is reduced by half from + / + to oim; (b) Mineral content (Ca_{max}) and microhardness of bone tissue (from Ref. [75]). Both parameters increase from + / + to oim/oim.

increased tissue mineralization was preserved in treatment with bisphosphonates [84]. The increased brittleness of the tissue is most likely due to a weakness of the collagen-matrix, associated with an increase in mineralization. Indeed, the collagen fibrils seem to break at only half the load in oim homozygotes [74] (see Figure 12.4). The reason for this inherent weakness of collagen might be a modified crosslink pattern in the fibrils [81], due to the fact that normal crosslinks between α_2 and α_1 chains cannot form due to an absence of α_2.

12.5.2 Pycnodysostosis

Pycnodysostosis is an extremely rare human genetic syndrome characterized by an increased bone mass (osteosclerosis), short stature and high bone fragility. Despite very small numbers of cases (about 100), pycnodysostosis was best known for its suggested affliction of the French painter, Henri de Toulouse-Lautrec.

The disorder is caused by a mutation in the gene encoding for cathepsin K, a key enzyme of osteoclastic degradation [85]. Indeed, patients affected by the disease [86], as well as mice mutants lacking cathepsin K activity [87–89], have differentiated osteoclasts that are able to demineralize the bone matrix but not to degrade the remaining organic matrix. As a consequence, the loss of cathepsin K activity leads to a fundamental defect of bone resorption and, subsequently, to an increase in bone mass. The bone resorption is not completely inhibited, however, and can occur to a limited extent via an alternative pathway. The degradation of any unmineralized

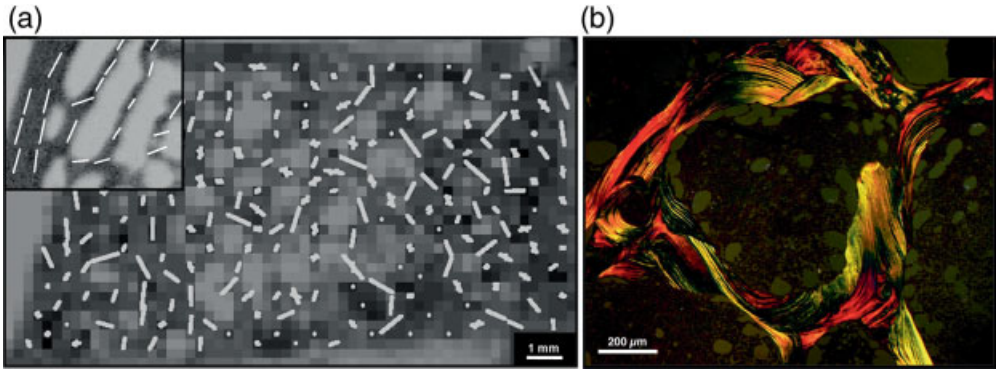


Figure 12.5 Disturbed lamellar organization in pycnodysostosis [92]. (a) The orientation of mineral particles in a biopsy is much less aligned than in normal bone (see inset (i)). The orientation and the length of the white bars indicate the direction and the local degree of alignment of the elongated plate-like mineral nanoparticles in bone, as measured with scanning small angle X-ray scattering; (b) Disturbed organization of the lamellar architecture of trabeculae, as revealed by polarized light microscopy.

collagen left by the dysfunctional osteoclasts may also be possible by the action of matrix-metalloproteinases synthesized by the bone lining-cells, which are members of the osteoblastic lineage [90]. It appears that these two pathways are not equivalent, however, as the lack of cathepsin K activity leads not only to a disturbed bone resorption [91] but also to a decreased bone formation activity [92, 93]. Bone tissue analyses of two affected patients also revealed defects at the nanostructural level, with mineral crystals being increased in size and reflecting a less-remodeled 'older' bone tissue. Moreover, the trabecular architecture appeared to be severely disturbed, with an unusually large variability in the orientation of the mineral particles and a highly disturbed lamellar organization, with the main collagen fibrils not oriented in the principal stress direction (see Figure 12.5). Thus, the absence of functional cathepsin K activity has a profound effect on bone quality at the nanoscale, and leads most likely to an observed increase in bone fragility.

12.5.3 Fluorosis

Fluoride has an anabolic effect on bone and is known to increase cancellous bone mass. During the 1980s and 1990s this led to fluoride being considered as a potential treatment of osteoporosis [94, 95], although clinical trials failed to confirm the anticipated anti-fracture efficacy [65, 96]. One reason for this is that fluoride clearly not only stimulates osteoblasts to form new bone but also has a direct effect on bone material quality. Studies involving small-angle X-ray scattering (SAXS) and back-scattered electron imaging [66, 97, 98] revealed that bone formed under the influence of fluoride has a quite different microscopic structure (see Figure 12.6). Moreover, the collagen–mineral nanocomposite was seen to be massively disturbed. Indeed, the

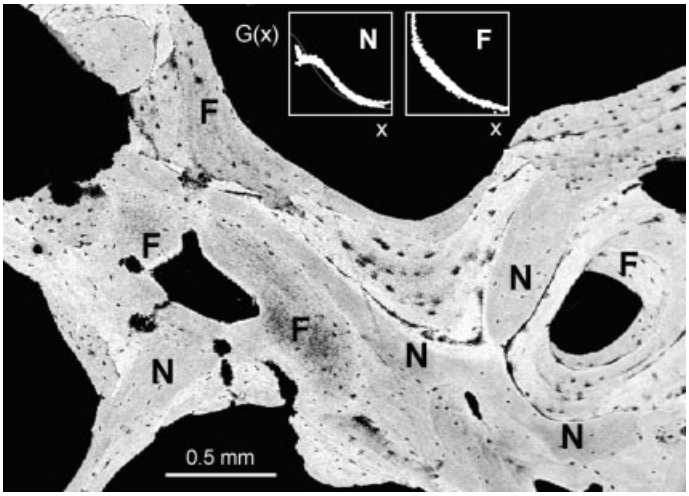


Figure 12.6 Bone biopsy after long-term fluoride treatment (3 years therapy; 50 mg NaF per day) [99]. 'N' indicates areas with normal bone and 'F' with fluorotic bone. The two insets at the top show the shape of the SAXS curves G(x) in normal and fluorotic bone [66, 98], indicating a severe modification of bone mineral nanoparticles during treatment.

strongly modified SAXS signal from bone areas newly formed under the influence of fluoride revealed the presence of mineral crystals much larger than in normal bone (Figure 12.6). This implied that the collagen and mineral in fluorotic bone did not form a well-organized nanocomposite, but that the large mineral crystals simply coexisted with the collagen fibrils. The result was a bone material of lower quality that would most likely be more brittle than usual. The images in Figure 12.6, which show a bone biopsy of a patient treated with sodium fluoride, also indicate that old bone with a normal structure coexists with newly formed fluorotic bone material. Due to a constant bone turnover, the old normal bone is gradually replaced by new bone with a fluorotic structure. This gradually compensates the positive mechanical effect of the bone mass increase and finally leads to a deterioration of bone stability against fracture [66, 97, 98].

12.6 Conclusions

Fractures – the clinical endpoint of disorders affecting the structure of bone material – are associated with increased morbidity, mortality and high socioeconomic costs [100, 101]. Today, due to an increased life expectancy for the general population, the incidence of fractures is also increasing, and the assessment of fracture risk and identification of those individuals who might benefit from the prevention and treatment of skeletal disorders represent major challenges in modern medicine.

New analyses of epidemiologic data provide strong evidence for the view that all (or better still, the overwhelming majority of) fractures – regardless of when they occur and of the level of trauma that precipitates them – may be based upon bone fragility [59], thus focusing all aspects of pathophysiology, diagnosis and treatment of skeletal diseases to the central question of mechanical competence and bone fragility. Following Robert Marcus' thoughts on ". . . the nature of osteoporosis" [102], 'bone fragility' might be defined most appropriately from the pathophysiological point of view as ". . . the consequence of a stochastic process, that is, multiple genetic, physical, hormonal and nutritional factors acting alone or in concert to diminish skeletal integrity."

Based on the fact that skeletal integrity is determined by the outstanding mechanical properties of bone at all hierarchical levels of its structure and organization [2], it becomes increasingly evident that a simple diagnostic parameter such as lumbar spine or hip BMD [103–105], although frequently used as a noninvasive diagnostic tool in clinical routine, does not have the diagnostic power to reflect the complex pathophysiological mechanisms that determine bone fragility. Thus, the availability of new diagnostic tools developed by materials scientists, coupled with a possible combinatorial approach using different methods to define the material qualities of bone from the micrometer to the nanometer scale, should introduce a renaissance of bone biopsies as diagnostic tools in clinical osteology. For example, the BMDD of trabecular human bone (as described above) was shown to be evolutionarily optimized within relative small variations (ca. 3%), independently of different skeletal regions for healthy adults aged between 25 and 95 years. Until now, no differences have been identified for BMDD-derived parameters with regards to gender or ethnicity. As shown in several examples, deviations from the normal BMDD seem to be associated with skeletal disorders, and in many examples indicate 'bone fragility' [57]. BMDD can be determined by using qBEI on a transiliac biopsy, as routinely occurs for histomorphometry, and combined with a variety of techniques based on spectroscopy, light scattering or biomechanical testing [2, 57].

When investigating the treatment of post-menopausal osteoporosis with the anti-resorptives alendronate [62] or risedronate [60] and the anabolic intermittent PTH [55], slight – but significant – deviations in BMDD indicated a lower mineralization for all placebo groups, and this was confirmed for idiopathic osteoporosis in pre-menopausal women. An example of learning from a materials science perspectives was that of fluorosis, and the fluoride treatment of post-menopausal osteoporosis [66, 97]. Yet, despite sodium fluoride being used widely to treat post-menopausal osteoporosis, no anti-fracture efficacy was reported. Rather, the bone quality revealed extensive and pathologic mineralization at both micro- and nanoscale, leading to a more brittle material with increased fragility.

Two classical genetic bone diseases – pycnodysostosis [92] and OI [68, 70, 71, 74–76, 78, 84] – point to a genetically related diminution of skeletal integrity. In OI, which often is fatal, the primary pathology was shown as brittle bones, inefficient repair mechanisms and a high bone turnover, whereas in pycnodysostosis the effects were caused by nonfunctioning osteoclasts due to mutations of the essential enzyme cathepsin K [85]. However, an inability to optimize structure by bone remodeling

results in a sclerosing bone disease with high bone mass and fragility fractures due to a disorganized structure at several hierarchical levels.

In conclusion, a wealth of evidence has been accumulated during the past years supporting the concept that the study of bone micro- and nanostructures will not only improve our understanding of the mechanisms that underlie bone fragility but also help to identify the effects of treatments. Nanomedicine, and its application to bone research, will in time undoubtedly broaden our knowledge of pathopysiology and improve the diagnoses, prevention and treatment of bone diseases. The availability of new techniques to investigate bone biopsies will surely challenge clinical osteologists and bone pathologists in the near future.

References

1 Weiner, S. and Wagner, H.D. (1998) *Annual Review of Materials Science*, **28**, 271.

2 Fratzl, P., Gupta, H.S., Paschalis, E.P. and Roschger, P. (2004) *Journal of Materials Chemistry*, **14**, 2115.

3 Fratzl, P. and Weinkamer, R. (2007) *Progress in Materials Science*, **52**, 1263.

4 Canty, E.G. and Kadler, K.E. (2002) *Comparative Biochemistry and Physiology. Part A, Molecular & Integrative Physiology*, **133**, 979.

5 Kadler, K.E., Holmes, D.F., Trotter, J.A. and Chapman, J.A. (1996) *The Biochemical Journal*, **316**, 1.

6 Hodge, A.J. and Petruska, J.A. (1963) *Aspects of Protein Structure* (ed G.N. Ramachandran), Academic Press, New York, p. 289.

7 Rubin, M.A., Rubin, J. and Jasiuk, W. (2004) *Bone*, **35**, 11.

8 Fantner, G.E., Hassenkam, T., Kindt, J.H., Weaver, J.C., Birkedal, H., Pechenik, L., Cutroni, J.A., Cidade, G.A.G., Stucky, G.D., Morse, D.E. and Hansma, P.K. (2005) *Nature Materials*, **4**, 612.

9 Landis, W.J. (1996) *Connective Tissue Research*, **35**, 1.

10 Weiner, S. and Traub, W. (1992) *The FASEB Journal*, **6**, 879.

11 Hassenkam, T., Fantner, G.E., Cutroni, J.A., Weaver, J.C., Morse, D.E. and Hansma, P.K. (2004) *Bone*, **35**, 4.

12 Sodek, J., Ganss, B. and McKee, M.D. (2000) *Critical Reviews in Oral Biology and Medicine*, **11**, 279.

13 Landis, W.J., Hodgens, K.J., Song, M.J., Arena, J., Kiyonaga, S., Marko, M., Owen, C. and McEwen, B.F. (1996) *Journal of Structural Biology*, **117**, 24.

14 Traub, W., Arad, T. and Weiner, S. (1992) *Matrix (Stuttgart, Germany)*, **12**, 251.

15 Roschger, P., Grabner, B.M., Rinnerthaler, S., Tesch, W., Kneissel, M., Berzlanovich, A., Klaushofer, K. and Fratzl, P. (2001) *Journal of Structural Biology*, **136**, 126.

16 Fratzl, P., Fratzl-Zelman, N., Klaushofer, K., Vogl, G. and Koller, K. (1991) *Calcified Tissue International*, **48**, 407.

17 Glimcher, M.J. (1984) *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **304**, 479.

18 Grynpas, M., Bonar, L.C. and Glimcher, M.J. (1985) *Journal of Materials Science*, **19**, 723.

19 Posner, A.S. (1985) *Clinical Orthopaedics*, **200**, 87.

20 Fratzl, P., Groschner, M., Vogl, G., Plenk, H., Eschberger, J., Fratzl-Zelman, N., Koller, K. and Klaushofer, K. (1992) *Journal of Bone and Mineral Research*, **7**, 329.

21 Currey, J.D. (1999) *The Journal of Experimental Biology*, **202**, 3285.

22 Currey, J.D. (2003) *Journal of Bone and Mineral Research*, **18**, 591.

23 Taylor, D., Hazenberg, J.G. and Lee, T.C. (2007) *Nature Materials*, **6**, 263.

24 Peterlik, H., Roschger, P., Klaushofer, K. and Fratzl, P. (2006) *Nature Materials*, **5**, 52.

25 Zioupos, P. and Currey, J.D. (1994) *Journal of Materials Science*, **29**, 978.

26 Vashishth, D., Tanner, K.E. and Bonfield, W. (2003) *Journal of Biomechanics*, **36**, 121.

27 Liu, D.M., Weiner, S. and Wagner, H.D. (1999) *Journal of Biomechanics*, **32**, 647.

28 Nalla, R.K., Kruzic, J.J. and Ritchie, R.O. (2004) *Bone*, **34**, 790.

29 Nalla, R.K., Kinney, J.H. and Ritchie, R.O. (2003) *Nature Materials*, **2**, 164.

30 Nalla, R.K., Kruzic, J.J., Kinney, J.H. and Ritchie, R.O. (2005) *Biomaterials*, **26**, 217.

31 Viguet-Carrin, S., Garnero, P. and Delmas, P.D. (2006) *Osteoporosis International*, **17**, 319.

32 Chavassieux, P., Seeman, E. and Delmas, P.D. (2007) *Endocrine Reviews*, **28**, 151.

33 Seeman, E. and Delmas, P.D. (2006) *The New England Journal of Medicine*, **354**, 2250.

34 Landis, W.J. (1995) *Bone*, **16**, 533.

35 Zioupos, P. and Currey, J.D. (1998) *Bone*, **22**, 57.

36 Zioupos, P., Currey, J.D. and Hamer, A.J. (1999) *Journal of Biomedical Materials Research*, **45**, 108.

37 Zioupos, P. (2001) *Journal of Biomaterials Applications*, **15**, 187.

38 Wang, X.D., Bank, R.A., Te Koppele, J.M. and Agrawal, C.M. (2001) *Journal of Orthopaedic Research*, **19**, 1021.

39 Wang, X., Shen, X., Li, X. and Agrawal, C.M. (2002) *Bone*, **31**, 1.

40 Gupta, H.S., Seto, J., Wagermaier, W., Zaslansky, P., Boesecke, P. and Fratzl, P. (2006) *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 17741.

41 Jager, I. and Fratzl, P. (2000) *Biophysical Journal*, **79**, 1737.

42 Thompson, J.B., Kindt, J.H., Drake, B., Hansma, H.G., Morse, D.E. and Hansma, P.K. (2001) *Nature*, **414**, 773.

43 Fantner, G., Hassenkam, T., Kindt, J.H., Weaver, J.C., Birkedal, H., Pechenik, L., Cutroni, J.A., Cidade, G.A.G., Stucky, G.D., Morse, D.E. and Hansma, P.K. (2005) *Nature Materials*, **4**, 612.

44 Gupta, H.S., Fratzl, P., Kerschnitzki, M., Benecke, G., Wagermaier, W. and Kirchner, H.O.K. (2007) *Journal of the Royal Society Interface*, **4**, 277.

45 Gupta, H.S., Wagermaier, W., Zickler, G.A., Aroush, D.R.B., Funari, S.S., Roschger, P., Wagner, H.D. and Fratzl, P. (2005) *Nano Letters*, **5**, 2108.

46 Tai, K., Ulm, F.J. and Ortiz, C. (2006) *Nano Letters*, **6**, 2520.

47 Giraud-Guille, M.M. (1988) *Calcified Tissue International*, **42**, 167.

48 Weiner, S., Arad, T., Sabanay, I. and Traub, W. (1997) *Bone*, **20**, 509.

49 Gao, H.J., Ji, B.H., Jager, I.L., Arzt, E. and Fratzl, P. (2003) *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 5597.

50 Currey, J.D. (2001) *Calcified Tissue International*, **68**, 205.

51 Currey, J.D. (2002) *Bones - Structure and Mechanics*, Princeton University Press, Princeton.

52 Ruffoni, D., Fratzl, P., Roschger, P., Klaushofer, K. and Weinkamer, R. (2007) *Bone*, **40**, 1308.

53 Eriksen, E.F., Axelrod, D.W. and Melsen, F. (1994) *Bone histomorphometry*, Raven Press: New York.

54 Eriksen, E.F., Melsen, F., Sod, E., Barton, I. and Chines, A. (2002) *Bone*, **31**, 620.

55 Misof, B.M., Roschger, P., Cosman, F., Kurland, E.S., Tesch, W., Messmer, P., Dempster, D.W., Nieves, J., Shane, E., Fratzl, P., Klaushofer, K., Bilezikian, J. and Lindsay, R. 2003 *The Journal of Clinical Endocrinology and Metabolism*, **88**, 1150.

56 Boivin, G. and Meunier, P.J. (2003) *Osteoporosis International*, **14**, S22.

57 Roschger, P., Paschalis, E.P., Fratzl, P. and Klaushofer, K. (2008) *Bone*, **42**, 456.

58 Roschger, P., Fratzl, P., Eschberger, J. and Klaushofer, K. (1998) *Bone*, **23**, 319.

59 Mackey, D.C., Lui, L.Y., Cawthon, P.M., Bauer, D.C., Nevitt, M.C., Cauley, J.A., Hillier, T.A., Lewis, C.E., Barrett-Connor, E. and Cummings, S.R. (2007) *The Journal of the American Medical Association*, **298**, 2381.

60 Zoehrer, R., Roschger, P., Paschalis, E.P., Hofstaetter, J.G., Durchschlag, E., Fratzl, P., Phipps, R. and Klaushofer, K. (2006) *Journal of Bone and Mineral Research*, **21**, 1106.

61 Fratzl, P., Roschger, P., Fratzl-Zelman, N., Paschalis, E.P., Phipps, R. and Klaushofer, K. (2007) *Calcified Tissue International*, **81**, 73.

62 Roschger, P., Rinnerthaler, S., Yates, J., Rodan, G.A., Fratzl, P. and Klaushofer, K. (2001) *Bone*, **29**, 185.

63 Haas, M., Leko-Mohr, Z., Roschger, P., Kletzmayr, J., Schwarz, C., Mitterbauer, C., Steininger, R., Grampp, S., Klaushofer, K., Delling, G. and Oberbauer, R. (2003) *Kidney International*, **63**, 1130.

64 Roschger, P., Mair, G., Fratzl-Zelman, N., Fratzl, P., Kimmel, D., Klaushofer, K., LaMotta, A. and Lombardi, A. (2007) *Journal of Bone and Mineral Research*, **22**, S129.

65 Riggs, B.L., Hodgson, S.F., Ofallon, W.M., Chao, E.Y.S., Wahner, H.W., Muhs, J.M., Cedel, S.L. and Melton, L.J. (1990) *The New England Journal of Medicine*, **322**, 802.

66 Fratzl, P., Roschger, P., Eschberger, J., Abendroth, B. and Klaushofer, K. (1994) *Journal of Bone and Mineral Research*, **9**, 1541.

67 Finkelstein, J.S., Hayes, A., Hunzelman, J.L., Wyland, J.J., Lee, H. and Neer, R.M. (2003) *The New England Journal of Medicine*, **349**, 1216.

68 Prockop, D.J. (1992) *The New England Journal of Medicine*, **326**, 540.

69 Sillence, D.O., Senn, A. and Danks, M.D. (1979) *Journal of Medical Genetics*, **16**, 101.

70 Rauch, F. and Glorieux, F.H. (2004) *Lancet*, **363**, 1377.

71 Jones, S.J., Glorieux, F.H., Travers, R. and Boyde, A. (1999) *Calcified Tissue International*, **64**, 8.

72 Roschger, P., Fratzl-Zelman, N., Misof, B.M., Glorieux, F.H., Klaushofer, K. and Rauch, F. (2008) *Calcified Tissue International*, **82**, 263.

73 Weber, M., Roschger, P., Fratzl-Zelman, N., Schoberl, T., Rauch, F., Glorieux, F.H., Fratzl, P. and Klaushofer, K. (2006) *Bone*, **39**, 616.

74 Misof, K., Landis, W.J., Klaushofer, K. and Fratzl, P. (1997) *The Journal of Clinical Investigation*, **100**, 40.

75 Grabner, B., Landis, W.J., Roschger, P., Rinnerthaler, S., Peterlik, H., Klaushofer, K. and Fratzl, P. (2001) *Bone*, **29**, 453.

76 Fratzl, P., Paris, O., Klaushofer, K. and Landis, W.J. (1996) *The Journal of Clinical Investigation*, **97**, 396.

77 Camacho, N.P., Hou, L., Toledano, T.R., Ilg, W.A., Brayton, C.F., Raggio, C.L., Root, L. and Boskey, A.L. (1999) *Journal of Bone and Mineral Research*, **14**, 264.

78 Mehta, S.S., Antich, P.P. and Landis, W.J. (1999) *Connective Tissue Research*, **40**, 189.

79 Grabner, B., Landis, W.J., Roschger, P., Rinnerthaler, S., Peterlik, H., Klaushofer, K. and Fratzl, P. (2001) *Bone*, **29**, 453.

80 Miles, C.A., Sims, T.J., Camacho, N.P. and Bailey, A.J. (2002) *Journal of Molecular Biology*, **321**, 797.

81 Sims, T.J., Miles, C.A., Bailey, A.J. and Camacho, N.P. (2003) *Connective Tissue Research*, **44**, 202.

82 Miller, E., Delos, D., Baldini, T., Wright, T.M. and Pleshko Camacho, N. (2007) *Calcified Tissue International*, **81**, 206.

83 Camacho, N.P., Hou, L., Toledano, T.R., Ilg, W.A., Brayton, C.F., Raggio, C.L., Root, L. and Boskey, A.L. (1999) *Journal of Bone and Mineral Research*, **14**, 264.

84 Misof, B.M., Roschger, P., Baldini, T., Raggio, C.L., Zraick, V., Root, L., Boskey, A.L., Klaushofer, K., Fratzl, P. and Camacho, N.P. (2005) *Bone*, **36**, 150.

85 Gelb, B.D., Shi, G.P., Chapman, H.A. and Desnick, R.J. (1996) *Science*, **273**, 1236.

86 Hou, W.S., Bromme, D., Zhao, Y.M., Mehler, E., Dushey, C., Weinstein, H., Miranda, C.S., Fraga, C., Greig, F., Carey, J., Rimoin, D.L., Desnick, R.J. and Gelb, B.D. 1999 *The Journal of Clinical Investigation*, **103**, 731.

87 Saftig, P., Hunziker, E., Wehmeyer, O., Jones, S., Boyde, A., Rommerskirch, W., Moritz, J.D., Schu, P. and von Figura, K. (1998) *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 13453.

88 Gowen, M., Lazner, F., Dodds, R., Kapadia, R., Feild, J., Tavaria, M., Bertoncello, I., Drake, F., Zavarselk, S., Tellis, I., Hertzog, P., Debouck, C. and Kola, I. 1999 *Journal of Bone and Mineral Research*, **14**, 1654.

89 Li, C.Y., Jepsen, K.J., Majeska, R.J., Zhang, J., Ni, R.J., Gelb, B.D. and Schaffler, M.B. (2006) *Journal of Bone and Mineral Research*, **21**, 865.

90 Everts, V., Delaisse, J.M., Korper, W., Jansen, D.C., Tigchelaar-Gutter, W., Saftig, P. and Beertsen, W. (2002) *Journal of Bone and Mineral Research*, **17**, 77.

91 Chen, W., Yang, S.Y., Abe, Y., Li, M., Wang, Y.C., Shao, J.Z., Li, E. and Li, Y.P. (2007) *Human Molecular Genetics*, **16**, 410.

92 Fratzl-Zelman, N., Valenta, A., Roschger, P., Nader, A., Gelb, B.D., Fratzl, P. and Klaushofer, K. (2004) *The Journal of Clinical Endocrinology and Metabolism*, **89**, 1538.

93 Chavassieux, P., Asser Karsdal, M., Segovia-Silvestre, T., Neutzsky-Wulff, A.V., Chapurlat, R., Boivin, G. and Delmas, P.D. (2008) *Journal of Bone and Mineral Research*, **23**, 1076.

94 Kleerekoper, M. (1998) *Endocrinology and Metabolism Clinics of North America*, **27**, 441.

95 Kleerekoper, M. (1996) *Critical Reviews in Clinical Laboratory Sciences*, **33**, 139.

96 Rubin, M.R. and Bilezikian, J.P. (2003) *Endocrinology and Metabolism Clinics of North America*, **32**, 285.

97 Roschger, P., Fratzl, P., Klaushofer, K. and Rodan, G. (1997) *Bone*, **20**, 393.

98 Fratzl, P., Schreiber, S., Roschger, P., Lafage, M.H., Rodan, G. and Klaushofer, K. (1996) *Journal of Bone and Mineral Research*, **11**, 248.

99 Fratzl, P., Rinnerthaler, S., Roschger, P. and Klaushofer, K. (1998) *Osteologie*, **7**, 130.

100 Ray, N.F., Chan, J.K., Thamer, M. and Melton, L.J. 3rd (1997) *Journal of Bone and Mineral Research*, **12**, 24.

101 Melton, L.J. 3rd (1993) *Bone*, **14** (Suppl 1), S1.

102 Marcus, R. (1996) *The Journal of Clinical Endocrinology and Metabolism*, **81**, 1.

103 Cummings, S.R., Bates, D. and Black, D.M. (2002) *The Journal of the American Medical Association*, **288**, 1889.

104 Schuit, S.C., van der Klift, M., Weel, A.E., de Laet, C.E., Burger, H., Seeman, E., Hofman, A., Uitterlinden, A.G., van Leeuwen, J.P. and Pols, H.A. (2004) *Bone*, **34**, 195.

105 Miller, P.D. (2006) *Reviews in Endocrine and Metabolic Disorders*, **7**, 75.

13
Nanoengineered Systems for Tissue Engineering and Regeneration

Ali Khademhosseini, Bimal Rajalingam, Satoshi Jinno, and Robert Langer

13.1
Introduction

In recent years, tissue engineering has emerged as a potentially powerful approach for the treatment of a variety of diseases by merging the principles of life sciences and engineering to generate biological substitutes that restore, maintain and enhance human tissue function [1]. In a typical tissue engineering approach, cells are seeded within biodegradable scaffolds. Then as the scaffolds degrade, the cells deposit their own matrices and self-assemble into tissue-like structures. This reassembly and degradation process eventually results in the formation of three-dimensional (3-D) tissue structures.

Over the past few years, tissue engineering has generated much excitement for fabricating a renewable source of transplantable tissues. Advances in the field have resulted in the engineering of clinically usable skin substitutes. In addition, other engineered tissues such as cartilage and bone are at various stages of clinical trials. Research groups have also attempted to engineer nerve tissue, pancreas, bladder and other organs. However, despite success in clinical studies, the dream of 'off-the-shelf' organs has eluded scientists and clinicians alike. This can be attributed to our inability to direct the behavior of cells in a desired manner, as well as to generate 3-D tissues with sufficient complexity and structural integrity to perform the function of the native tissues. Other concerns include the transmission of infection through the implanted tissue-engineered substrate, the possibility of immune reaction against the implanted cells, the variability of engineered products, the introduction of genetically modified, unwanted and potentially harmful cells into the body as well as regulatory and ethical aspects.

In order to generate functional tissues it is important to mimic the biological microenvironment by developing approaches and tools that can control materials and cells at the nanoscale. This is because many structural elements such as the extracellular matrix (ECM), as well as biological processes such as receptor clustering, are at the nanoscale. Thus, the ability to engineer the cellular environment and tissue

Nanotechnology, Volume 5: Nanomedicine. Edited by Viola Vogel

ISBN: 978-3-527-31736-3

structure at the nanoscale is a potentially powerful approach for generating biomimetic tissues. These processes will be critical not only for generating tissue engineered constructs but also for engineering *in vitro* systems that can be used for various drug discovery and diagnostics applications.

Nanotechnology is an emerging field that is concerned with the design, synthesis, characterization and application of materials and devices that have a functional organization in at least one dimension on the nanometer scale, ranging from a few to about 100 nm [1]. Due to this ability to control features at small length scales, nanotechnology is becoming more commonly used in a number of biomedical endeavors ranging from drug delivery [2–5] to *in vivo* imaging [6].

In this chapter we will discuss the application of nanotechnology to tissue engineering as an enabling tool. Specifically, we will provide an overview of two different types of nanoengineered system that are used in tissue engineering. First, we will focus on various approaches that are used to generate nanoscale modifications to existing polymers and materials. These nanoengineered systems, such as nanopatterned substrates and electrospun scaffolds, provide structures that influence cell behavior and the subsequent tissue formation. Furthermore, we will discuss the use of other nanoscale structures such as controlled-release nanoparticles for tissue engineering. In the second part of the chapter we will discuss the use of nanotechnology for the synthesis of novel materials that behave differently as bulk compared to their nanoscale versions. Such materials include self-assembled materials, carbon nanotubes and quantum dots. Throughout the chapter we will discuss the use of nanomaterials for controlling the cellular microenvironment and for generating 3-D tissues. We will also detail the potential limitations and emerging topics of interest and challenges in this area of research. Clearly, the application of nanotechnology to tissue engineering and cell culture is an 'exploding' field, and hopefully in this chapter we will provide a glimpse into the various applications. Throughout the chapter, when applicable, the reader is directed to more extensive reviews to provide further detail regarding specific topics.

13.2
Nanomaterials Synthesized Using Top-Down Approaches

In 1959, Richard Feynman introduced the significant benefits associated with manipulating materials atom by atom. Since then, extensive research has been conducted in nanotechnology owing to the advances in technologies that enable the manipulation and characterization of nanoscale objects such as scanning tunneling microscopy (STM), atomic force microscopy (AFM) and other related technologies. Nanomaterials have generated interest in a variety of fields such as clinical medicine, defense, pharmaceuticals, aerospace, energy and biological research. Today, engineered nanomaterials are increasingly utilized for various tissue engineering applications due to their controllable and unique properties. Furthermore, nanostructures can aid in mimicking Nature, as many biological structures have features that are on the order of few to hundreds of nanometers. Nanoengineered materials

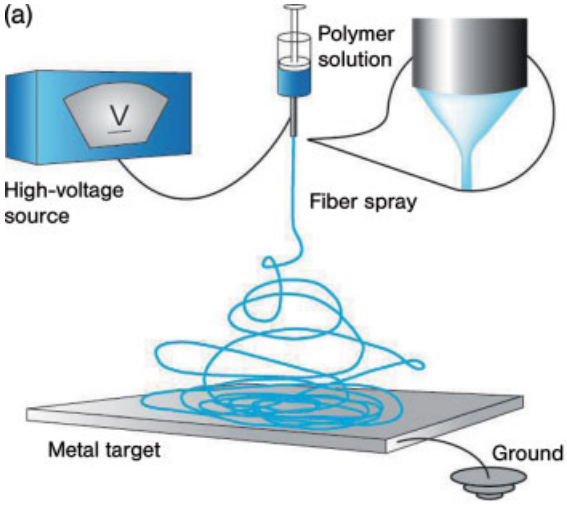
can be created by tailoring the properties of existing materials, for example by controlling the 3-D structure or surface roughness. In general, nanomaterials can be produced by either 'top-down' or 'bottom-up' approaches. Top-down approaches involve the miniaturization of materials to nano length scales, and have been enabled due to increased technological capability for miniaturizing materials either by using novel approaches or by making improvements to existing techniques. In this section, we will describe two techniques used to prepare nanomaterials using top-down approaches, namely the use of electrospinning and nanopatterned substrates to engineer cell behavior.

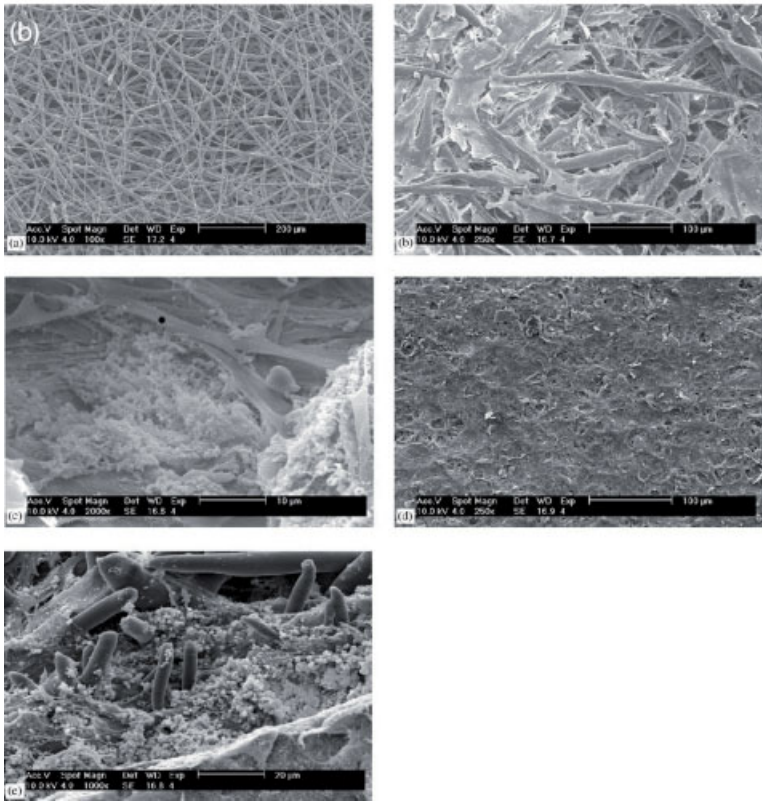
13.2.1 Electrospinning Nanofibers

Electrospinning is a technique that is used to generate nanofibrous scaffolds (Figure 13.1a). These scaffolds are highly porous (i.e. a large surface area-to-volume ratio), and thus mimic many of the properties of natural tissues and also provide cells with a pseudo 3-D environment [7]. Electrospinning is also relatively inexpensive and capable of producing nanofibers from a variety of biodegradable synthetic polymers, such as poly(lactic-*co*-glycolic acid) (PLGA) [8], polycaprolactone (PCL) [9] and poly(L-lactic acid) (PLLA) [10], as well as natural polymers such as collagen [11]. The nanoscale features of electrospun scaffolds promote cell proliferation and also guide cell growth. For example, aligned nanofibers have been shown to induce the growth and proliferation of cardiac cells into contractile spindle structures [12]. The seeding of neural stem cells (NSCs) on aligned PLLA nano/micro fibrous scaffolds has also resulted in neurite outgrowth along with the fiber direction for the aligned scaffolds [10]. Furthermore, fiber diameter and orientation have been shown to influence the cell morphology, while cell proliferation is not sensitive to the above-mentioned parameters. Goldstein and colleagues, while testing a range of diameters (0.14–3.6 μm) and angular standard deviations (31–60°), found that an increasing fiber diameter and degree of fiber orientation resulted in increased projected cell area and aspect ratio [8].

Electrospinning has been used to fabricate scaffolds for various tissues such as bone [9], cartilage [13–15] and cardiac muscle [16]. An example of bone tissue fabrication is shown in Figure 13.1b. These scaffolds have also been used to direct the differentiation of stem cells; for example, the coating of electrospun nanofibers composed of polyamide have been used to promote the proliferation and self-renewal of mouse embryonic stem cells (ESCs). These ESCs maintained their ability to differentiate into various lineages, which showed that such fibers could be used not only for *in vivo* applications as tissue engineered scaffolds but also as tools for *in vitro* cell culture. Electrospun nanofibers have also been shown to influence cell shape, actin cytoskeleton and matrix deposition, both *in vitro* [17, 18] and *in vivo* [19].

Despite its versatility, one disadvantage with electrospinning process is that the range of the resulting fibers is usually limited to the upper range of natural ECM fibers. In addition, it is difficult to control the complex architecture and intricate


(a)
Polymer solution
V
High-voltage source
Fiber spray
Metal target
Ground


(b)

cell–cell, cell–ECM and cell-soluble factors in the resulting scaffolds. This has prompted the use of other techniques to fabricate nanofabricated scaffolds. For example, nanofabricated PLLA scaffolds that supported the differentiation and outgrowth of NSCs have been fabricated by a liquid–liquid phase separation method [20]. Yet, despite these limitations, electrospinning is a powerful technology for the fabrication of 3-D nanoscaffolds.

13.2.2
Scaffolds with Nanogrooved Surfaces

Both, micro- and nanotextured substrates can significantly influence cell behavior such as adhesion, gene expression [15] and migration [16]. This is because the interaction of cells with a biomaterial results in the localization of focal adhesions, actin stress fibers and microtubules [21]. Focal contacts are involved in signal transduction pathways which in turn can regulate a wide array of cell function [22] (Figure 13.2). As nanofiber scaffolds have a larger surface area, they have more potential binding sites to cell membrane receptors, thus affecting the cell behavior in unique ways. It has also been observed that the cells display more *filopodia* when they come in contact with the nanoscale surfaces, presumably to sense the external topography [23]. Although the mechanism of cellular response to nanotopography is not entirely understood, it is suggested that an interaction of the cellular processes and interfacial forces results in peculiar cellular behavior [7]. Both, micro- and nanotextured substrates can be engineered either on tissue culture substrates or within 3-D tissue scaffolds. Within tissue engineering scaffolds, nanotextures provide physical cues to seeded cells and regulate the interaction of host cells with the scaffold. For example, surfaces that have desired roughness have been shown to increase osteoblast adhesion for orthopedic replacement/augmentation applications [24].

Nanotextures can be generated using a variety of techniques, depending on the material as well as the dimension and shapes of the desired structures. For example, features less than 100 nm may be produced by a range of techniques including chemical etching in metals [25], the embedding of carbon nanofibers in composite

Figure 13.1 (a) Schematic of the electrospinning apparatus. The set-up for electrospinning consists of a spinneret with a metallic needle, a syringe pump, a high-voltage power supply and a grounded collector. By using the electrospinning apparatus, uniform fibers with nanometer-scale diameters can be fabricated. These scaffolds are highly porous (large surface area-to-volume ratio), thus mimicking many of the properties of natural tissues and providing cells with a pseudo 3-D environment. The nanoscale features of electrospun scaffolds promote cell proliferation and guide cell growth; (b) Panel (a) shows an electrospun poly(ε-caprolactone) scaffold, prior to cell seeding. Panels (b) and (c) are images of mesenchymal stem cells (MSCs) seeded on the scaffold after 7 days of culture, with low and high magnification, respectively. Panels (d) and (e) show the MSCs after four weeks of culture, again with low and high magnification, respectively. After 7 days, osteoblast-like cells developed, indicating bone-like formation. (Adapted from Ref. [9].)

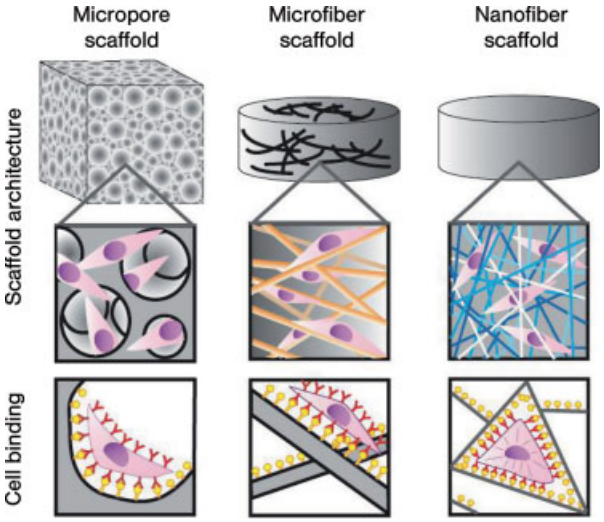


Figure 13.2 The influence of scaffold architecture on cell attachment and spreading. The attachment of cells on the micropore or microfiber scaffold is similar to that of cells cultured on flat surfaces. The larger surface area provided by the nanofiber scaffold results in many more binding sites to cell membrane receptors, thus influencing cell behavior and the subsequent tissue formation in unique ways. Focal contacts are involved in signal transduction pathways, which in turn can regulate a wide array of cell function. (Adapted from Ref. [7].)

materials [26], casting polymer replicas from ECM [27], or the embedding of constituent nanoparticles in materials ranging from metals to ceramics to composites [28–31]. Electron-beam lithography (EBL) technology has been used to fabricate well-defined nanostructures at sub-50 nm length scales [21, 32]. These tools can be used to form structures at the same length scales as the native ECM, and thus enable the systematic study of cell behavior (for a review, see Ref. [24]).

The shape of the nanostructures influences the cell behavior and phenotype. For example, nanogrooves [33–37] result in an alignment of cells parallel to the direction of the grooves [33, 38] as well as the alignment of actin, microtubules and other cytoskeletal elements [39–41]. Interestingly, both the pitch and depth of the grooves influence cell behavior. For example, typically the orientation increases with increased depth of the nanogrooves [42]. Another shape that has been shown to influence cell behavior is the natural roughness of tissues. For example, endothelial cells that were cultured on the ECM-textured replicas spread faster and had an appearance more like the cells in their native arteries than did cells grown on nontextured surfaces [27, 43]. Fibroblasts cultured on nanopatterned ε-PCL surfaces were also less spread compared to those on a planar substrate [44]. Furthermore, human mesenchymal stem cells (MSCs) and ESCs align on nanofabricated substrates and differentiate in a specific manner [45, 46]. Therefore, by controlling the nanotopography of tissue engineering scaffolds inductive signals can be delivered to enhance tissue formation and function.

13.3
Nanomaterials Synthesized using Bottom-Up Approaches

In this section we describe the synthesis and application of nanomaterials that are built by nanoscale assembly of molecules with properties that are often different from their individual components and their bulk material. These materials include self-assembled peptide hydrogels, quantum dots, carbon nanotubes and layer-by-layer deposited films.

13.3.1
Self-Assembled Peptide Scaffolds

A promising approach in tissue engineering is to use nanoengineered materials made from synthetic peptides or peptide amphiphiles (PA) that self-assemble. (An extensive review of this topic is provided in Chapter 14.) Self-assembled PA hydrogels can be generated by linking a carbon alkyl tail to functional peptides; these PA molecules then self-assemble based on hydrophobic interactions of the alkyl tail and thus form nanofibers that can form hydrogels either alone or by mixing with a cell suspension [47]. An example of a self-assembled peptide is shown in Figure 13.3. In addition, self-assembled peptide hydrogels made purely from peptides that self-assemble based on hydrophobic interactions have also been demonstrated [48–50]. In this approach, self-assembled beta-sheets are formed which can assemble into hydrogels. Both of these approaches have shown promising results in various tissue engineering, stem cell differentiation and cell culture applications [47, 51–54]. An example of this is the recent investigation of the proliferation and differentiation of MSCs in PA hydrogels. When rat MSCs were seeded into the PA nanofibers, with or without RGD, a larger number of cells attached to the PA nanofibers that contained RGD. Furthermore, upon examination of the osteogenic differentiation of MSCs, the alkaline phosphatase (ALP) activity and osteocalcin content increased for the PA nanofibers that contained RGD compared with those without RGD. In another study, the use of MSC-seeded hybrid scaffolds prepared from PAs and a collagen sponge reinforced with poly(glycolic acid) (PGA) fibers was examined to show increased osteogenic differentiation of MSCs and ectopic bone formation.

13.3.2
Layer-by-Layer Deposition of Nanomaterials

The ability to control the surface properties of biological interfaces is useful in various aspects of tissue engineering. One means of obtaining such controlled surfaces is by the layer-by-layer (LBL) deposition of the charged biopolymers. LBL deposition uses the electrostatic interaction between the surface and the polyelectrolyte solutions to generate films with nanoscale dimensions. LBL has been used extensively to control the cellular microenvironment *in vitro*. For example, we have generated patterned cellular cocultures using the LBL deposition of ionic biopolymers hyaluronic acid (HA), poly-L-lysine (PLL) [55] and collagen. In this approach, micropatterns of

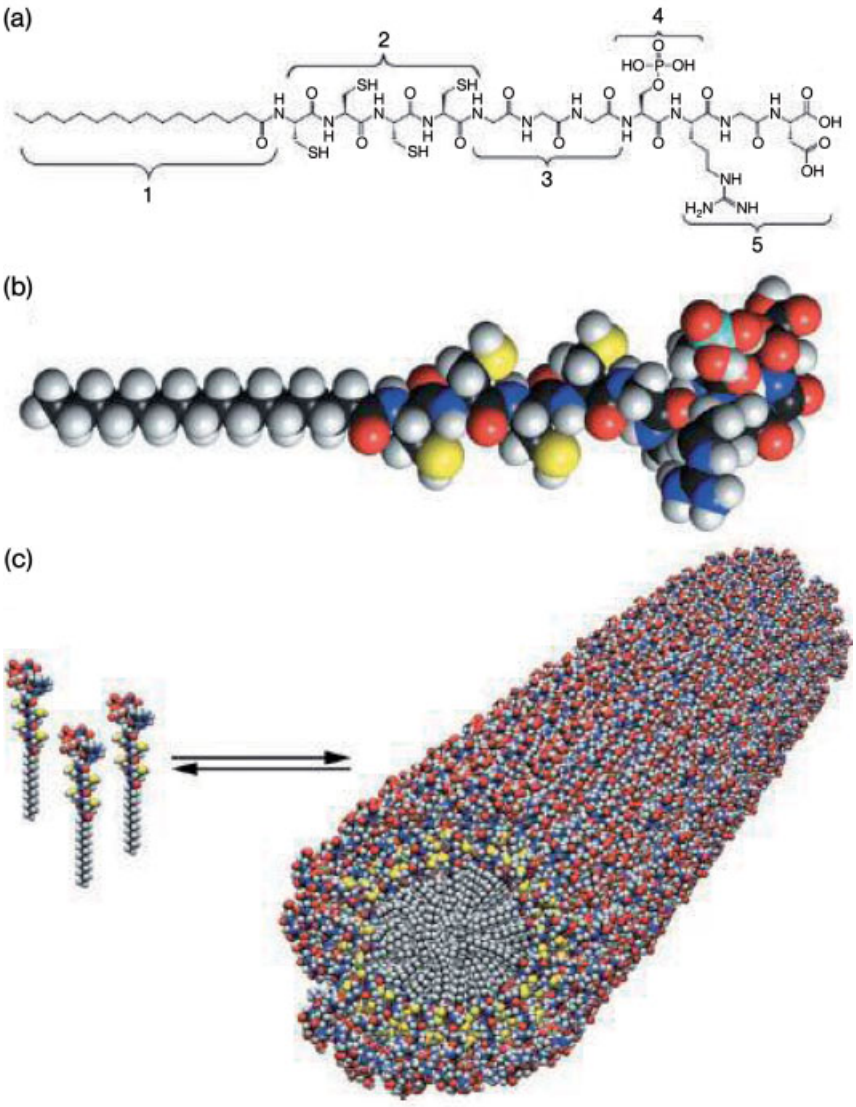


Figure 13.3 Self-assembled peptide scaffold. (a) The chemical structure of the peptide amphiphile. This is composed of a long alkyl tail (region 1), four consecutive cysteine residues (region 2), a flexible linker region of three glycine residues (region 3), a single phosphorylated serine residue (region 4) and cell adhesion ligand RGD (region 5); (b) The molecular model of (a); (c) These peptide amphiphiles are self-assembling into a cylindrical micelle. (Adapted from Ref. [47].)

nonbiofouling anionic HA were used to pattern cells on glass substrates. The subsequent adsorption of the cation PLL to HA pattern resulted in an adherent surface promoting the attachment of a second cell type. In order to minimize

toxicity, instead of PLL other positively charged molecules such as ECM components (i.e. collagen) have also been used (Figure 13.4) [56]. In a related experiment, cultured human endothelial cells have been patterned using LBL on a polyurethane surface. Here, it was observed that the cells did not spread on the negatively charged surface due to an electrostatic repulsion, whereas inversing the surface charge by adding positively charged collagen increased the cell spreading and proliferation. Thus, cell

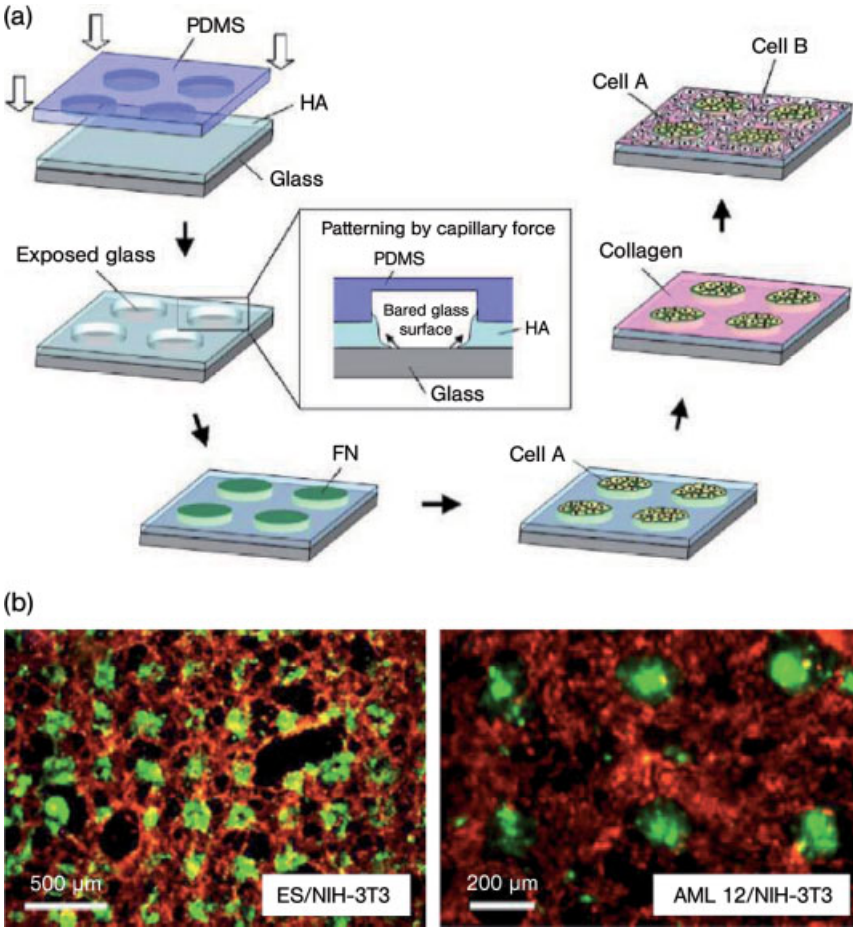


Figure 13.4 (a) The generation of cocultures using layer-by-layer deposition of the ECM materials, hyaluronic acid (HA) and collagen. A polydimethylsiloxane (PDMS) mold was placed on a glass slide coated with a thin layer of HA. Due to capillary forces, the HA in the exposed space of the PDMS mold receded, thus exposing the underlying glass surface. The exposed region of a glass substrate was coated with fibronectin (FN), where primary cells (cell A) could be selectively adhered. Subsequently, the PDMS mold was removed and collagen layered on the HA surface to make the surface adherent to secondary cells (cell B); (b) The fluorescent images of the patterned coculture after three days of culture, generated by a layer-by-layer coculture approach. The ESCs and hepatocytes (AML 12) were cocultured with the fibroblasts (NIH-3T3).

attachment on multilayer thin films may depend on the charge of the terminal polyion layer [57].

The number of layers in the LBL films may play a role in the cell attachment behavior. It was reported that increasing the number of layers of titanium oxide nanoparticle thin films increased the surface roughness, cell attachment and the rate of cell spreading. Although this may be due to the increased surface roughness, it also demonstrates the potential of this technology in controlling cell-surface properties [58]. Furthermore, the LBL assembly of PLL and dextran sulfate could be used to increase the rate of fibronectin deposition and the subsequent cell adhesion relative to the control substrates [59].

The LBL deposition of materials has been used in a variety of tissue engineering applications. For example, the LBL assembly of HgTe has been used to fabricate a hybrid device where the absorption of light by quantum nanoparticles stimulates neural cells by a sequence of photochemical and charge-transfer reactions [60]. These devices may be of potential use in tissue engineering applications in which it is desired to stimulate nerve cells using external cues. The LBL assemblies of nanoparticles have also been explored as a means of protecting arteries damaged during revascularization procedures. It has been reported that the deposition of self-assembled nanocoatings comprising alternating depositions of HA and chitosan onto aortic porcine arteries, led to a significant inhibition of the growth of thrombus on the damaged arterial surfaces. Clearly, this technique has the potential for clinical application to protect damaged arteries and to prevent subsequent restenosis [61]. Therefore, by properly choosing the LBL materials it is possible to modify the surface properties of materials for tissue engineering as well as for biosensing [62, 63] and drug delivery applications [64].

Mironov and colleagues have used the LBL technique for organ printing, with precise control over the spatial position of the deposited cell [65]. LBL deposition can also be used for the fabrication of immunosensors [66], islet cell encapsulation [67] and polyelectrolyte capsules for drug release [68].

13.3.3 Carbon Nanotubes

Carbon nanotubes are nanomaterials with unique mechanical and chemical properties. They have been used for cell tracking, for the delivery of desired molecules to cells, and as components of tissue engineering scaffolds [69]. Carbon nanotubes, depending on the number of carbon walls, can range from 1.5 to 30 nm in diameter and may be hundreds of nanometers in length.

Within tissue engineering scaffolds carbon nanotubes can be used to modify the mechanical and chemical properties. Furthermore, carbon nanotubes can be functionalized with biomolecules to signal the surrounding cells, or they may be electrically stimulated due to their high electrical conductivity to excite tissues such as muscle cells and nerve cells. One potentially powerful method of integrating carbon nanotubes into tissue engineering is by generating composite materials which comprise a biocompatible material such as collagen with embedded, single-

walled carbon nanotubes. As an example, smooth muscle cells (SMCs) have been encapsulated within collagen–carbon nanotube composite matrices with high cell viability (>85%) for at least 7 days [70]. Single-walled carbon nanotubes can also be used for culturing excitable tissues such as neuronal and muscle cells [71]. It has also been suggested that the growth of the neuronal circuits in carbon nanotubes might result in a significant increase in the network activity and an increase in neural transmission, perhaps due to the high electric conductivity of carbon nanotubes [72]. Furthermore, the electrical stimulation of osteoblasts cultured in nanocomposites comprising PLA and carbon nanotubes increased their cell proliferation and the deposition of calcium after 21 days. These data show that the use of novel current-conducting nanocomposites would be valuable for enhancing the function of the osteoblasts, and also provide useful avenues in bone tissue engineering [73].

Carbon nanotubes have also been used to delivery pharmaceutical drugs [74–76], genetic material [77–79] and biomolecules such as proteins [80, 81] to various cell types. For example, carbon nanotubes have been used to deliver Amphotericin B to fungal-infected cells. Here, the Amphotericin B was found to be bonded covalently to carbon nanotubes and was uptaken by mammalian cells, without significant toxicity, while maintaining its antifungal activity [82]. Thus, carbon nanotubes may be used for the delivery of antibiotics to specific cells.

The influence of carbon nanotubes on cells varies, depending on the type and their surface properties. For example, it has been reported that rat osteosarcoma (ROS) 17/2.8 cells, when cultured on carbon nanotubes carrying a neutral electric charge, proliferated to a greater extent than did other control cells [83]. Chemically modifying the surface of carbon nanotubes can also be used to enhance their cytocompatibility. For example, carbon nanotubes have been coated with bioactive 4-hydroxynonenal in order to culture embryonic rat brain neurons that promote neuronal branching compared to unmodified carbon nanotubes [84]. Despite such promise, however, the cytotoxicity of carbon nanotubes remains unclear. It is well known that various properties such as surface modifications and size greatly influence the potential toxicity of these structures. For example, long carbon nanotubes have been shown to generate a greater degree of inflammation in rats than shorter carbon nanotubes ($\sim$200 nm), which suggests that the smaller particles may be engulfed more easily by macrophages [85]. Other studies have also shown that carbon nanotubes may not only inhibit cell growth [86] but also induce pulmonary injury in the mouse model [87], when sequential exposure to carbon nanotubes and bacteria enhanced pulmonary inflammation and infectivity. Thus, more extensive and systematic studies must be conducted to ensure that the use of these nanomaterials in tissue engineering does not result in long-term toxicity.

13.3.4
MRI Contrast Agents

Nanotechnology may also permit the high-resolution imaging of tissue-engineered constructs. Specifically, the use of imaging contrast agents in magnetic resonance imaging (MRI) can be used to track cells *in vivo* and visualize constructs [88–90].

Although MRI contrast agents take many forms, nanoparticle systems have emerged in recent years as one of the most promising as nanoparticles not only provide an enormous surface area but can also be functionalized with targeting ligands and magnetic labels. Moreover, their smaller size provides an easy permeability across the blood capillaries.

Iron oxide nanoparticles have shown great promise for use in MRI to track cells, because they can be uptaken without compromising cell viability and are relatively safe. A wide variety of iron oxide-based nanoparticles have been developed that differ in hydrodynamic particle size and surface-coating material (dextran, starch, albumin, silicones) [91]. In general, these particles are categorized based on their diameter into superparamagnetic iron oxides (SPIOs) (50–500 nm) and ultrasmall superparamagnetic iron oxides (USPIOs) (<50 nm), with the size dictating their physico-chemical and pharmacokinetic properties. It has also been shown that clearance of the iron oxide nanoparticles in the rat liver depends on the outer coating [92].

Iron oxide nanoparticles have been used for imaging various organs, including the gastrointestinal tract, liver, spleen and lymph nodes. Furthermore, smaller-sized particles can also be used for angiography and perfusion imaging in myocardial and neurological diseases. Iron oxide particles can be coated with various molecules to increase their circulation and targeting. For example, dextran-coated iron oxide nanoparticles have been used for labeling cells, while anionic magnetic nanoparticles can be used to target positively charged tissues by using electrostatic interactions [93] (Figure 13.5). In addition, iron oxide nanoparticles can be used to track cells *in vivo* after transplantation. For example, MSCs and other mammalian cells labeled with SPIO nanoparticles were used to track cells in both experimental and clinical settings [94, 95]. Furthermore, green fluorescent protein (GFP^+) ESCs that were

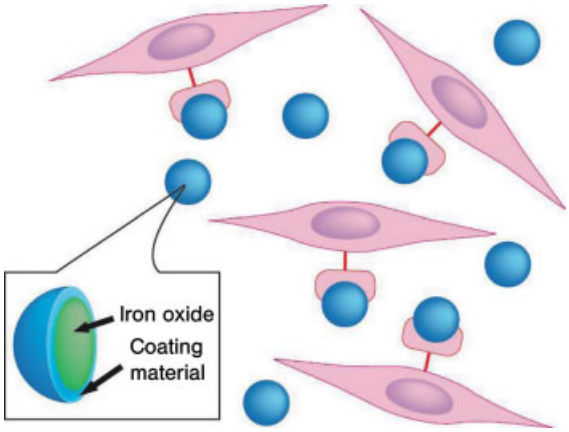


Figure 13.5 Schematic of the magnetic resonance molecular imaging. The tracking of magnetic nanoparticles to cancer cells is based on their static and dynamic magnetism, along with an ability to impart cell-specific functionality. The biocompatibility of coating materials (i.e., biscarboxyl-terminated poly(ethylene glycol)) allows for *in vivo* applications in animals and humans. Therefore, iron oxide nanoparticles show great promise for use in MRI to track cells *in vivo* and to visualize constructs, without compromising cell viability.

labeled with dextran-coated iron oxide nanoparticles and implanted into the brains of rats with brain stroke showed that the cells could be tracked for at least three weeks [96]. The *in vivo* tracking of iron oxide nanoparticle-labeled rat bone marrow MSCs and mouse ESCs and human $CD34^+$ hematopoietic progenitor cells in rats with a cortical or spinal cord lesion has also shown that cells may remain visible in the lesion for at least 50 days [97, 98]. Taken together, these and other results [99] indicate that magnetic nanoparticles are well-suited for the noninvasive analysis of cell migration, engraftment and morphological differentiation at high spatial and temporal resolution.

In order to target desired cells or to modify the rate of cellular uptake, nanoparticles may be engineered with specific molecules on their surfaces. For example, in order to increase their internalization, NSCs and $CD34^+$ bone marrow cells were labeled with superparamagnetic nanoparticles that were conjugated with short HIV-Tat peptides. This increased the internalization of the particles by the cells, without affecting their viability, differentiation or proliferation. The localization and retrieval of cell populations *in vivo* enabled a detailed analysis of specific stem cell and organ interactions that were critical for advancing the therapeutic use of stem cells [100]. In addition to iron oxide nanoparticles, other types of nanoparticle have also been used for tissue imaging, notably with applications in tissue engineering. As an example, fluoroscein isothiocyanate (FITC) -conjugated mesoporous silica nanoparticles (MSNs) have been used to label human bone marrow MSCs and 3T3-L1 cells. The FITC-MSNs were efficiently internalized into MSCs and 3T3-L1 cells, even with short-term incubation (2–4 h), without affecting cell viability [101]. Thus, it seems that nanoparticles can be used potentially not only to track cells but also to image tissues which may be useful for the noninvasive imaging of tissue-engineered constructs.

13.3.5
Quantum Dots

Nanoscale probes can also be used in tissue engineering applications for the study of various biological processes, as well as for real-time cell detection and tracking. Fluorescent dyes, which traditionally have been used to image cells and tissues, have several drawbacks including photobleaching and a lack of long-term stability. Quantum dots (QDs) are nanoparticles that have several advantages over conventional fluorophores for imaging, including tunable properties and a resistance to photobleaching [6]. QDs are semiconductor nanostructures that confine the motion of conduction band electrons, valence band holes or excitons in all three spatial directions. The band gap energy of the QD is the energy difference between the valence band and the conduction band. For nanoscale semiconductor particles such as QDs, the bandgap is dependent on the size of the nanocrystal, which results in a size-dependent variation in emission. A single light source can also be used for the simultaneous excitation of a spectrum of emission wavelength, which makes the method useful for multicolor, multiplexed biological detection and imaging applications.

QDs can be used for the ultrasensitive imaging of molecular targets in deep tissue and living animals (Figure 13.6). Here, they are used as specific markers for cellular structures [102, 103] and molecules [104], for monitoring physiological events in live cells [105–107], for measuring cell motility [108], and for monitoring RNA delivery and tracking cells [109] *in vivo*. As an example, QDs have been used for locating multiple distinct endogenous proteins within cells, thus determining the precise protein distribution in a high-throughput manner [110]. Peptide ligand-conjugated QDs have also been used for imaging G-protein-coupling receptors in both whole-cells and as single-molecules [111]. Cellular events such as the transport of lipids and proteins across membranes have also been tracked using QDs with molecular resolution in live cells [112]. Furthermore, QDs conjugated to immunoglobulin G (IgG) and streptavidin have been used to label the breast cancer marker Her2 on the surface of fixed and live cancer cells [113].

QDs have significant potential in analyzing the mechanisms of cell growth, apoptosis, cell–cell interactions, cell differentiations and inflammatory responses. For example, QDs have been used to study the signaling pathways of mast cells during an inflammatory response [114], as well as to quantify changes in organelle morphology during apoptosis [115]. In addition, the photostability and biocompatibility of QDs make them the preferred agents for the long-term tracking of live cells [116]. QDs are internalized into cells by endocytosis [117], by receptor-mediated uptake [118], by peptide-mediated transportation [119, 120] or microinjection [121]. An example of this was recently demonstrated in studies in which ligand-conjugated QDs were used to monitor antigen binding, entry and trafficking in dendritic cells [122]. QDs, when conjugated to a transporter protein, have also been used to label malignant and nonmalignant hematological cells and to track cell division, thus enabling lineage tracking [109].

Despite the remarkable potential for the application of QDs in clinical medicine, their toxicity and long-term adverse effects are still not clearly understood. The metabolism, excretion and toxicity of QDs may depend on multiple factors such as size, charge, concentration and outer coating bioactivity, as well as their oxidative, photolytic and mechanical stabilities and other unknown factors [123]. Importantly, these issues must be addressed before QDs can be used for *in vivo* applications in humans.

13.4 Future Directions

During a relatively short period of time, nanomaterials have spawned a number of new approaches to address important challenges in tissue engineering. These challenges range from understanding the mechanisms of stem cell differentiation to generating functional vasculature within tissue engineering constructs. Despite these advancements, further investigations are required to analyze the true potential and clinical viability of these technologies. Our current lack of knowledge regarding the long-term toxicity of many nanoengineered materials represents a

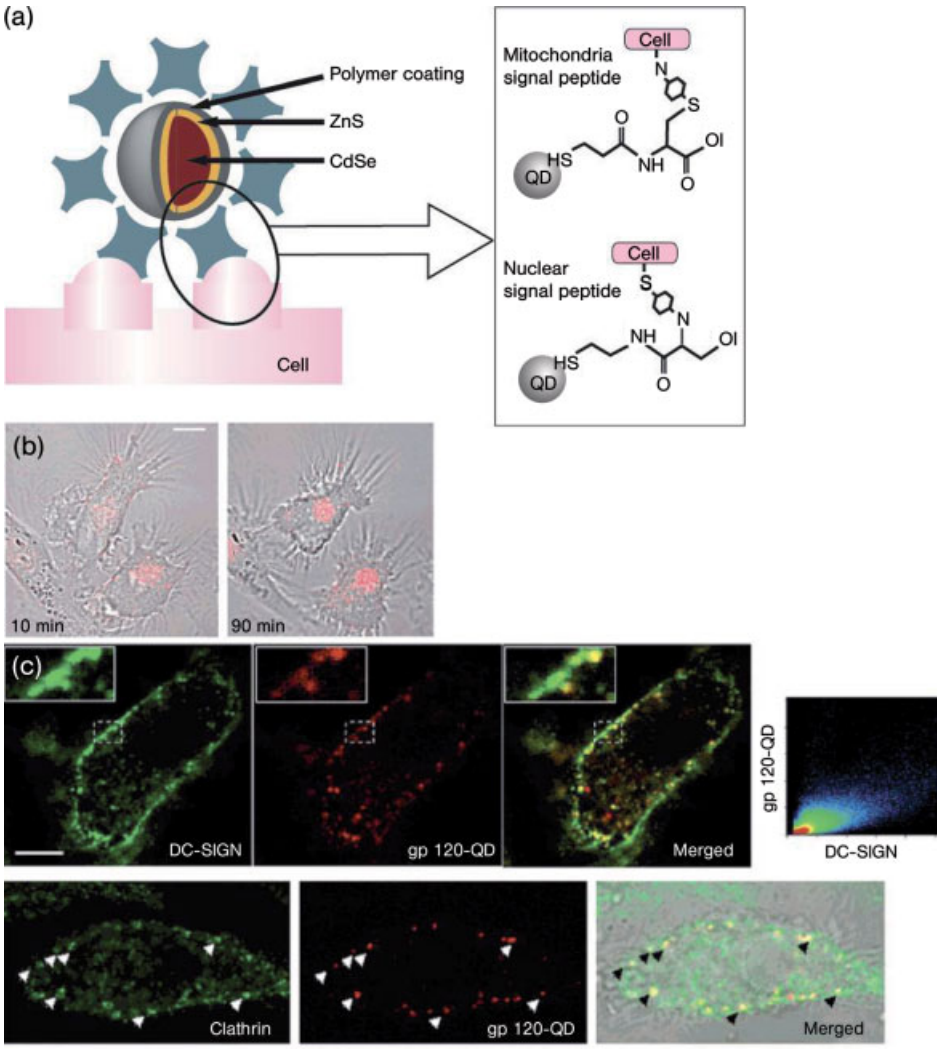


Figure 13.6 (a) Schematic of quantum dots (QDs). QDs are typically made from nanocrystals of a semiconductor material (CdSe), which has been coated with an additional semiconductor shell (ZnS) to improve the material's optical properties. This material is coated with a polymer shell that allows the materials to be conjugated to biological molecules and to retain their optical properties. These nanocrystals have been coupled to various biomolecules directly or indirectly. The inset at the right shows a schematic of peptide-conjugated QDs for organelle targeting and imaging. The amino acid-coated QDs are conjugated with target peptides by coupling. QDs can reveal the transduction of proteins and peptides into specific subcellular compartments as a powerful tool for studying intracellular analysis *in vitro* and even *in vivo*. (Adapted from Ref. [124].); (b) Ligand-conjugated QDs internalized by dendritic cells (DCs) via their specific binding protein (DC-SIGN). Ligand-coated QDs bind to DC-SIGN and are endocytosed into DCs; (c) DCs were incubated with the HIV-1 envelope glycoprotein gp120-QDs (red). After washing of unbound QDs, DCs were fixed and labeled with DC-SIGN marker (green). Data were obtained using confocal microscopy. The right-hand panel shows a 2-D histogram of DC-SIGN signal versus gp120-QDs signal. This result indicated that the small amount of dispersion leads to high colocalization. (Adapted from Ref. [122].)

critical barrier for their use in humans. Traditionally, tissue engineers have favored materials that have a long history of medical application (i.e. FDA approved), although many such materials have limitations to be overcome, perhaps through rational design enabled by nanotechnology. Thus, there remains a clear need to develop nanoengineered materials capable of addressing the various challenges of tissue engineering. In addition, systematic toxicity studies must be conducted in order to fully optimize and characterize not only the function but also the long-term behavior of nanomaterials *in vivo*. Yet, many of the traditional methods used to analyze previous generations of biomaterials do not apply to nanoscale materials, and a clear paradigm shift is required in these analytical and standardization procedures. This will range from how we study material–cell interactions *in vitro* and *in vivo*, to the standardization requirements of regulatory bodies such as the FDA. Clearly, these modifications will require extensive discussion amongst the scientists, the patients, the general public, the clinicians and the regulatory officers.

13.5 Conclusions

Today, nanotechnology offers a wide variety of tools in tissue engineering, biomedical imaging, biosensing, diagnostics and drug delivery. Nanotechnology-based applications are valuable in the research and development of viable substitutes that may restore, maintain or even improve the function of human tissues. Today, these materials have not only opened up novel applications but have also addressed a number of limitations associated with traditional approaches and materials. Nonetheless, much research is required to further demonstrate the long-term stability and clinical utility of nanoengineered materials.

Acknowledgments

The authors greatly appreciate the helpful discussions with Drs Hossein Hosseinkhani and Hossein Baharvand. They also acknowledge generous funding from the Draper laboratory, the CIMIT, NIH and Coulter Foundation.

References

1 Khademhosseini, A. and Langer, R. (2006) Nanobiotechnology for Tissue Engineering and Drug Delivery. *Chemical Engineering Progress*, **102**, 38–42.

2 Sengupta, S. and Sasisekharan, R. (2007) Exploiting nanotechnology to target cancer. *British Journal of Cancer*, **96**, 1315–1319.

3 Bisht, S., Feldmann, G., Soni, S., Ravi, R., Karikar, C., Maitra, A. and Maitra, A. (2007) Polymeric nanoparticle-encapsulated curcumin (nanocurcumin):

a novel strategy for human cancer therapy. *Journal of Nanobiotechnology*, **5**, 3.

4 Allen, T.M. and Cullis, P.R. (2004) Drug delivery systems: entering the mainstream. *Science*, **303**, 1818–1822.

5 Pfeifer, B.A., Burdick, J.A., Little, S.R. and Langer, R. (2005) Poly(ester-anhydride): poly(beta-amino ester) micro- and nanospheres: DNA encapsulation and cellular transfection. *International Journal of Pharmaceutics*, **304**, 210–219.

6 Alivisatos, A.P., Gu, W. and Larabell, C. (2005) Quantum dots as cellular probes. *Annual Review of Biomedical Engineering*, **7**, 55–76.

7 Stevens, M.M. and George, J.H. (2005) Exploring and engineering the cell surface interface. *Science*, **310**, 1135–1138.

8 Bashur, C.A., Dahlgren, L.A. and Goldstein, A.S. (2006) Effect of fiber diameter and orientation on fibroblast morphology and proliferation on electrospun poly(D,L-lactic-*co*-glycolic acid) meshes. *Biomaterials*, **27**, 5681–5688.

9 Yoshimoto, H., Shin, Y.M., Terai, H. and Vacanti, J.P. (2003) A biodegradable nanofiber scaffold by electrospinning and its potential for bone tissue engineering. *Biomaterials*, **24**, 2077–2082.

10 Yang, F., Murugan, R., Wang, S. and Ramakrishna, S. (2005) Electrospinning of nano/micro scale poly(L-lactic acid) aligned fibers and their potential in neural tissue engineering. *Biomaterials*, **26**, 2603–2610.

11 Matthews, J.A., Wnek, G.E., Simpson, D.G. and Bowlin, G.L. (2002) Electrospinning of collagen nanofibers. *Biomacromolecules*, **3**, 232–238.

12 Xu, C.Y., Inai, R., Kotaki, M. and Ramakrishna, S. (2004) Aligned biodegradable nanofibrous structure: a potential scaffold for blood vessel engineering. *Biomaterials*, **25**, 877–886.

13 Fertala, A., Han, W.B. and Ko, F.K. (2001) Mapping critical sites in collagen II for rational design of gene-engineered proteins for cell-supporting materials. *Journal of Biomedical Materials Research*, **57**, 48–58.

14 Li, W.J., Laurencin, C.T., Caterson, E.J., Tuan, R.S. and Ko, F.K. (2002) Electrospun nanofibrous structure: a novel scaffold for tissue engineering. *Journal of Biomedical Materials Research*, **60**, 613–621.

15 Li, W.J., Cooper, J.A. Jr, Mauck, R.L. and Tuan, R.S. (2006) Fabrication and characterization of six electrospun poly (alpha-hydroxy ester)-based fibrous scaffolds for tissue engineering applications. *Acta Biomaterialia*, **2**, 377–385.

16 Zong, X., Bien, H., Chung, C.-Y., Yin, L., Fang, D., Hsiao, B.S., Chu, B. and Entcheva, E. (2005) Electrospun fine-textured scaffolds for heart tissue constructs. *Biomaterials*, **26**, 5330–5338.

17 Nur, E.K.A., Ahmed, I., Kamal, J., Schindler, M. and Meiners, S. (2005) Three dimensional nanofibrillar surfaces induce activation of Rac. *Biochemical and Biophysical Research Communications*, **331**, 428–434.

18 Schindler, M., Ahmed, I., Kamal, J., Nur-E-Kamal, A., Grafe, T.H., Chung, H.Y. and Meiners, S. (2005) A synthetic nanofibrillar matrix promotes *in vivo*-like organization and morphogenesis for cells in culture. *Biomaterials*, **26**, 5624–5631.

19 Nur, E.K.A., Ahmed, I., Kamal, J., Schindler, M. and Meiners, S. (2006) Three-dimensional nanofibrillar surfaces promote self-renewal in mouse embryonic stem cells. *Stem Cells (Dayton, Ohio)*, **24**, 426–433.

20 Yang, F., Murugan, R., Ramakrishna, S., Wang, X., Ma, Y.-X. and Wang, S. (2004) Fabrication of nano-structured porous PLLA scaffold intended for nerve tissue engineering. *Biomaterials*, **25**, 1891–1900.

21 Dalby, M.J., Marshall, G.E., Johnstone, H.J., Affrossman, S. and Riehle, M.O. (2002) Interactions of human blood and tissue cell types with 95-nm-high

nanotopography. *IEEE Transactions on NanoBioscience*, **1**, 18–23.

22 Mrksich, M., Chen, C.S., Xia, Y., Dike, L.E., Ingber, D.E. and Whitesides, G.M. (1996) Controlling cell attachment on contoured surfaces with self-assembled, monolayers of alkanethiolates on gold. *Proceedings of the National Academy of Sciences of the United States of America*, **93**, 10775–10778.

23 Curtis, A.S., Gadegaard, N., Dalby, M.J., Riehle, M.O., Wilkinson, C.D. and Aitchison, G. (2004) Cells react to nanoscale order and symmetry in their surroundings. *IEEE Transactions on NanoBioscience*, **3**, 61–65.

24 Sato, M. and Webster, T.J. (2004) Nanobiotechnology: implications for the future of nanotechnology in orthopedic applications. *Expert Review of Medical Devices*, **1**, 105–114.

25 de Oliveira, P.T. and Nanci, A. (2004) Nanotexturing of titanium-based surfaces upregulates expression of bone sialoprotein and osteopontin by cultured osteogenic cells. *Biomaterials*, **25**, 403–413.

26 Price, R.L., Waid, M.C., Haberstroh, K.M. and Webster, T.J. (2003) Selective bone cell adhesion on formulations containing carbon nanofibers. *Biomaterials*, **24**, 1877–1887.

27 Goodman, S.L., Sims, P.A. and Albrecht, R.M. (1996) Three-dimensional extracellular matrix textured biomaterials. *Biomaterials*, **17**, 2087–2095.

28 Webster, T.J., Siegel, R.W. and Bizios, R. (1999) Osteoblast adhesion on nanophase ceramics. *Biomaterials*, **20**, 1221–1227.

29 Webster, T.J. (2001) *Nanostructured Materials* (ed. J.Y. Ying), Academy Press, New York, pp. 126–166.

30 Webster, T.J., Ergun, C., Doremus, R.H., Siegel, R.W. and Bizios, R. (2000) Enhanced functions of osteoblasts on nanophase ceramics. *Biomaterials*, **21**, 1803–1810.

31 Webster, T.J., Ergun, C., Doremus, R.H., Siegel, R.W. and Bizios, R. (2001) Enhanced osteoclast-like cell functions on nanophase ceramics. *Biomaterials*, **22**, 1327–1333.

32 Suh, K.Y., Khademhosseini, A., Eng, G. and Langer, R. (2004) Single nanocrystal arrays on patterned poly(ethylene glycol) copolymer microstructures using selective wetting and drying. *Langmuir*, **20**, 6080–6084.

33 Chou, L., Firth, J.D., Uitto, V.J. and Brunette, D.M. (1998) Effects of titanium substratum and grooved surface topography on metalloproteinase-2 expression in human fibroblasts. *Journal of Biomedical Materials Research*, **39**, 437–445.

34 Rajnicek, A., Britland, S. and McCaig, C. (1997) Contact guidance of CNS neurites on grooved quartz: influence of groove dimensions, neuronal age and cell type. *Journal of Cell Science*, **110** (Pt 23), 2905–2913.

35 Meyle, J., Wolburg, H. and von Recum, A.F. (1993) Surface micromorphology and cellular interactions. *Journal of Biomaterials Applications*, **7**, 362–374.

36 van Kooten, T.G. and von Recum, A.F. (1999) Cell adhesion to textured silicone surfaces: the influence of time of adhesion and texture on focal contact and fibronectin fibril formation. *Tissue Engineering*, **5**, 223–240.

37 Wojciak-Stothard, B., Curtis, A., Monaghan, W., MacDonald, K. and Wilkinson, C. (1996) Guidance and activation of murine macrophages by nanometric scale topography. *Experimental Cell Research*, **223**, 426–435.

38 Meyle, J., Gultig, K. and Nisch, W. (1995) Variation in contact guidance by human cells on a microstructured surface. *Journal of Biomedical Materials Research*, **29**, 81–88.

39 Wojciak-Stothard, B., Curtis, A.S., Monaghan, W., McGrath, M., Sommer, I. and Wilkinson, C.D. (1995) Role of the cytoskeleton in the reaction of fibroblasts to multiple grooved substrata. *Cell Motility and the Cytoskeleton*, **31**, 147–158.

40 Wojciak-Stothard, B., Madeja, Z., Korohoda, W., Curtis, A. and Wilkinson, C. (1995) Activation of macrophage-like cells by multiple grooved substrata. Topographical control of cell behaviour. *Cell Biology International*, **19**, 485–490.

41 Oakley, C. and Brunette, D.M. (1995) Response of single, pairs, and clusters of epithelial cells to substratum topography. *Biochemistry and Cell Biology*, **73**, 473–489.

42 Webb, A., Clark, P., Skepper, J., Compston, A. and Wood, A. (1995) Guidance of oligodendrocytes and their progenitors by substratum topography. *Journal of Cell Science*, **108** (Pt 8), 2747–2760.

43 Flemming, R.G., Murphy, C.J., Abrams, G.A., Goodman, S.L. and Nealey, P.F. (1999) Effects of synthetic micro- and nano-structured surfaces on cell behavior. *Biomaterials*, **20**, 573–588.

44 Gallagher, J.O., McGhee, K.F., Wilkinson, C.D. and Riehle, M.O. (2002) Interaction of animal cells with ordered nanotopography. *IEEE Transactions on NanoBioscience*, **1**, 24–28.

45 Yim, E.K. and Leong, K.W. (2005) Significance of synthetic nanostructures in dictating cellular response. *Nanomedicine*, **1**, 10–21.

46 Yim, E.K., Wen, J. and Leong, K.W. (2006) Enhanced extracellular matrix production and differentiation of human embryonic germ cell derivatives in biodegradable poly(epsilon-caprolactone-*co*-ethyl ethylene phosphate) scaffold. *Acta Biomaterialia*, **2**, 365–376.

47 Hartgerink, J.D., Beniash, E. and Stupp, S.I. (2001) Self-assembly and mineralization of peptide-amphiphile nanofibers. *Science*, **294**, 1684–1688.

48 Zhang, S. (2002) Emerging biological materials through molecular self-assembly. *Biotechnology Advances*, **20**, 321–339.

49 Zhang, S. (2003) Fabrication of novel biomaterials through molecular self-assembly. *Nature Biotechnology*, **21**, 1171–1178.

50 Zhang, S., Marini, D.M., Hwang, W. and Santoso, S. (2002) Design of nanostructured biological materials through self-assembly of peptides and proteins. *Current Opinion in Chemical Biology*, **6**, 865–871.

51 Hosseinkhani, H., Hosseinkhani, M., Tian, F., Kobayashi, H. and Tabata, Y. (2006) Ectopic bone formation in collagen sponge self-assembled peptide-amphiphile nanofibers hybrid scaffold in a perfusion culture bioreactor. *Biomaterials*, **27**, 5089–5098.

52 Hosseinkhani, H., Hosseinkhani, M., Tian, F., Kobayashi, H. and Tabata, Y. (2006) Osteogenic differentiation of mesenchymal stem cells in self-assembled peptide-amphiphile nanofibers. *Biomaterials*, **27**, 4079–4086.

53 Hartgerink, J.D., Beniash, E. and Stupp, S.I. (2002) Peptide-amphiphile nanofibers: a versatile scaffold for the preparation, of self-assembling materials. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 5133–5138.

54 Niece, K.L., Hartgerink, J.D., Donners, J.J. and Stupp, S.I. (2003) Self-assembly combining two bioactive peptide-amphiphile molecules into nanofibers by electrostatic attraction. *Journal of the American Chemical Society*, **125**, 7146–7147.

55 Khademhosseini, A., Suh, K.Y., Yang, J.M., Eng, G., Yeh, J., Levenberg, S. and Langer, R. (2004) Layer-by-layer deposition of hyaluronic acid and poly-l-lysine for patterned cell co-cultures. *Biomaterials*, **25**, 3583–3592.

56 Fukuda, J., Khademhosseini, A., Yeh, J., Eng, G., Cheng, J., Farokhzad, O.C. and Langer, R. (2006) Micropatterned cell co-cultures using layer-by-layer deposition of extracellular matrix components. *Biomaterials*, **27**, 1479–1486.

57 Zhu, Y. and Sun, Y. (2004) The influence of polyelectrolyte charges of polyurethane membrane surface on the growth of

human endothelial cells. *Colloids and Surfaces. B, Biointerfaces*, **36**, 49–55.

58 Kommireddy, D.S., Sriram, S.M., Lvov, Y.M. and Mills, D.K. (2006) Stem cell attachment to layer-by-layer assembled TiO_2 nanoparticle thin films. *Biomaterials*, **27**, 4296–4303.

59 Wittmer, C.R., Phelps, J.A., Saltzman, W.M. and Van Tassel, P.R. (2007) Fibronectin terminated multilayer films: protein adsorption and cell attachment studies. *Biomaterials*, **28**, 851–860.

60 Pappas, T.C., Wickramanyake, W.M., Jan, E., Motamedi, M., Brodwick, M. and Kotov, N.A. (2007) Nanoscale engineering of a cellular interface with semiconductor nanoparticle films for photoelectric stimulation of neurons. *Nano Letters*, **7**, 513–519.

61 Thierry, B., Winnik, F.M., Merhi, Y. and Tabrizian, M. (2003) Nanocoatings onto arteries via layer-by-layer deposition: toward the *in vivo* repair of damaged blood vessels. *Journal of the American Chemical Society*, **125**, 7494–7495.

62 Zhao, L., Liu, H. and Hu, N. (2006) Assembly of layer-by-layer films of heme proteins and single-walled carbon nanotubes: electrochemistry and electrocatalysis. *Analytical and Bioanalytical Chemistry*, **384**, 414–422.

63 Wu, B.Y., Hou, S.H., Yin, F., Li, J., Zhao, Z.X., Huang, J.D. and Chen, Q. (2007) Amperometric glucose biosensor based on layer-by-layer assembly of multilayer films composed of chitosan, gold nanoparticles and glucose oxidase modified Pt electrode. *Biosensors and Bioelectronics*, **22**, 838–844.

64 Fan, Y.F., Wang, Y.N., Fan, Y.G. and Ma, J.B. (2006) Preparation of insulin nanoparticles and their encapsulation with biodegradable polyelectrolytes via the layer-by-layer adsorption. *International Journal of Pharmaceutics*, **324**, 158–167.

65 Mironov, V., Kasyanov, V., Drake, C. and Markwald, R.R. (2008) Organ printing: promises and challenges. *Regenerative Medicine*, **3**, 93–103.

66 Pastorino, L., Soumetz, F.C. and Ruggiero, C. (2007) Nanostructured thin films for the development of piezoelectric immunosensors. *29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS 2007), Volume 1*, pp. 2257–2260.

67 Teramura, Y., Kaneda, Y. and Iwata, H. (2007) Islet-encapsulation in ultra-thin layer-by-layer membranes of poly(vinyl alcohol) anchored to poly(ethylene glycol)-lipids in the cell membrane. *Biomaterials*, **28**, 4818–4825.

68 De Geest, B.G., Sanders, N.N., Sukhorukov, G.B., Demeester, J. and De Smedt, S.C. (2007) Release mechanisms for polyelectrolyte capsules. *Chemical Society Reviews*, **36**, 636–649.

69 Harrison, B.S. and Atala, A. (2007) Carbon nanotube applications for tissue engineering. *Biomaterials*, **28**, 344–353.

70 MacDonald, R.A., Laurenzi, B.F., Viswanathan, G., Ajayan, P.M. and Stegemann, J.P. (2005) Collagen-carbon nanotube composite materials as scaffolds in tissue, engineering. *Journal of Biomedical Materials Research Part A*, **74**, 489–496.

71 Liopo, A.V., Stewart, M.P., Hudson, J., Tour, J.M. and Pappas, T.C. (2006) Biocompatibility of native and functionalized single-walled carbon nanotubes for neuronal interface. *Journal of Nanoscience and Nanotechnology*, **6**, 1365–1374.

72 Lovat, V., Pantarotto, D., Lagostena, L., Cacciari, B., Grandolfo, M., Righi, M., Spalluto, G., Prato, M. and Ballerini, L. (2005) Carbon nanotube substrates boost neuronal electrical signaling. *Nano Letters*, **5**, 1107–1110.

73 Supronowicz, P.R., Ajayan, P.M., Ullmann, K.R., Arulanandam, B.P., Metzger, D.W. and Bizios, R. (2002) Novel current-conducting composite substrates for exposing osteoblasts to alternating

current stimulation. *Journal of Biomedical Materials Research*, **59**, 499–506.

74 Bianco, A., Kostarelos, K. and Prato, M. (2005) Applications of carbon nanotubes in drug delivery. *Current Opinion in Chemical Biology*, **9**, 674–679.

75 Venkatesan, N., Yoshimitsu, J., Ito, Y., Shibata, N. and Takada, K. (2005) Liquid filled nanoparticles as a drug delivery tool for protein therapeutics. *Biomaterials*, **26**, 7154–7163.

76 LaVan, D.A., McGuire, T. and Langer, R. (2003) Small-scale systems for *in vivo* drug delivery. *Nature Biotechnology*, **21**, 1184–1191.

77 Singh, R., Pantarotto, D., McCarthy, D., Chaloin, O., Hoebeke, J., Partidos, C.D., Briand, J.P., Prato, M., Bianco, A. and Kostarelos, K. (2005) Binding and condensation of plasmid DNA onto functionalized carbon nanotubes: toward the construction of nanotube-based gene delivery vectors. *Journal of the American Chemical Society*, **127**, 4388–4396.

78 Dobson, J. (2006) Gene therapy progress and prospects: magnetic nanoparticle-based gene delivery. *Gene Therapy*, **13**, 283–287.

79 Gao, L., Nie, L., Wang, T., Qin, Y., Guo, Z., Yang, D. and Yan, X. (2006) Carbon nanotube delivery of the GFP gene into mammalian cells. *Chembiochem: A European Journal of Chemical Biology*, **7**, 239–242.

80 Shi Kam, N.W., Jessop, T.C., Wender, P.A. and Dai, H. (2004) Nanotube molecular transporters: internalization of carbon nanotube-protein conjugates into mammalian cells. *Journal of the American Chemical Society*, **126**, 6850–6851.

81 Kam, N.W. and Dai, H. (2005) Carbon nanotubes as intracellular protein transporters: generality and biological functionality. *Journal of the American Chemical Society*, **127**, 6021–6026.

82 Wu, W., Wieckowski, S., Pastorin, G., Benincasa, M., Klumpp, C., Briand, J.P., Gennaro, R., Prato, M. and Bianco, A. (2005) Targeted delivery of amphotericin B to cells by using functionalized carbon nanotubes. *Angewandte Chemie - International Edition in English*, **44**, 6358–6362.

83 Zanello, L.P., Zhao, B., Hu, H. and Haddon, R.C. (2006) Bone cell proliferation on carbon nanotubes. *Nano Letters*, **6**, 562–567.

84 Mattson, M.P., Haddon, R.C. and Rao, A.M. (2000) Molecular functionalization of carbon nanotubes and use as substrates for neuronal growth. *Journal of Molecular Neuroscience*, **14**, 175–182.

85 Sato, Y., Yokoyama, A., Shibata, K., Akimoto, Y., Ogino, S., Nodasaka, Y., Kohgo, T., Tamura, K., Akasaka, T., Uo, M., Motomiya, K., Jeyadevan, B., Ishiguro, M., Hatakeyama, R., Watari, F. and Tohji, K. (2005) Influence of length on cytotoxicity of multi-walled carbon nanotubes against human acute monocytic leukemia cell line THP-1 in vitro and subcutaneous tissue of rats *in vivo*. *Molecular BioSystems*, **1**, 176–182.

86 Raja, P.M., Connolley, J., Ganesan, G.P., Ci, L., Ajayan, P.M., Nalamasu, O. and Thompson, D.M. (2007) Impact of carbon nanotube exposure, dosage and aggregation on smooth muscle cells. *Toxicology Letters*, **169**, 51–63.

87 Chou, C.C., Hsiao, H.Y., Hong, Q.S., Chen, C.H., Peng, Y.W., Chen, H.W. and Yang, P.C. (2008) Single-walled carbon nanotubes can induce pulmonary injury in mouse model. *Nano Letters*, **8**, 437–445.

88 Frangioni, J.V. and Hajjar, R.J. (2004) *In vivo* tracking of stem cells for clinical trials in cardiovascular disease. *Circulation*, **110**, 3378–3383.

89 Shapiro, E.M., Sharer, K., Skrtic, S. and Koretsky, A.P. (2006) *In vivo* detection of single cells by MRI. *Magnetic Resonance in Medicine*, **55**, 242–249.

90 Shapiro, E.M., Gonzalez-Perez, O., Manuel Garcia-Verdugo, J., Alvarez-Buylla, A. and Koretsky, A.P. (2006) Magnetic resonance imaging of the migration of neuronal precursors

generated in the adult rodent brain. *NeuroImage*, **32**, 1150–1157.

91 Weissleder, R., Elizondo, G., Wittenberg, J., Rabito, C.A., Bengele, H.H. and Josephson, L. (1990) Ultrasmall superparamagnetic iron oxide: characterization of a new class of contrast agents for MR imaging. *Radiology*, **175**, 489–493.

92 Briley-Saebo, K.C., Johansson, L.O., Hustvedt, S.O., Haldorsen, A.G., Bjornerud, A., Fayad, Z.A. and Ahlstrom, H.K. (2006) Clearance of iron oxide particles in rat liver: effect of hydrated particle size and coating material on liver metabolism. *Investigative Radiology*, **41**, 560–571.

93 Wilhelm, C., Billotey, C., Roger, J., Pons, J.N., Bacri, J.C. and Gazeau, F. (2003) Intracellular uptake of anionic superparamagnetic nanoparticles as a function of their surface coating. *Biomaterials*, **24**, 1001–1011.

94 Frank, J.A., Miller, B.R., Arbab, A.S., Zywicke, H.A., Jordan, E.K., Lewis, B.K., Bryant, L.H. Jr. and Bulte, J.W. (2003) Clinically applicable labeling of mammalian and stem cells by combining superparamagnetic iron oxides and transfection agents. *Radiology*, **228**, 480–487.

95 Frank, J.A., Anderson, S.A., Kalsih, H., Jordan, E.K., Lewis, B.K., Yocum, G.T. and Arbab, A.S. (2004) Methods for magnetically labeling stem and other cells for detection by *in vivo* magnetic resonance imaging. *Cytotherapy*, **6**, 621–625.

96 Hoehn, M., Kustermann, E., Blunk, J., Wiedermann, D., Trapp, T., Wecker, S., Focking, M., Arnold, H., Hescheler, J., Fleischmann, B.K., Schwindt, W. and Buhrle, C. (2002) Monitoring of implanted stem cell migration *in vivo*: a highly resolved, *in vivo* magnetic resonance imaging investigation of experimental stroke in rat. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 16267–16272.

97 Sykova, E. and Jendelova, P. (2005) Magnetic resonance tracking of implanted adult and embryonic stem cells in injured brain and spinal cord. *Annals of the New York Academy of Sciences*, **1049**, 146–160.

98 Stroh, A., Faber, C., Neuberger, T., Lorenz, P., Sieland, K., Jakob, P.M., Webb, A., Pilgrimm, H., Schober, R., Pohl, E.E. and Zimmer, C. (2005) *In vivo* detection limits of magnetically labeled embryonic stem cells in the rat brain using high-field (17.6 T) magnetic resonance imaging. *NeuroImage*, **24**, 635–645.

99 Zhu, M., Zhong, Y.M., Li, Y.H., Sun, A.M. and Jin, B. (2005) Congenital aortic arch anomalies: diagnosis using contrast enhanced magnetic resonance angiography. *Chinese Medical Journal*, **118**, 1751–1753.

100 Lewin, M., Carlesso, N., Tung, C.H., Tang, X.W., Cory, D., Scadden, D.T. and Weissleder, R. (2000) Tat peptide-derivatized magnetic nanoparticles allow *in vivo* tracking and recovery of progenitor cells. *Nature Biotechnology*, **18**, 410–414.

101 Huang, D.M., Hung, Y., Ko, B.S., Hsu, S.C., Chen, W.H., Chien, C.L., Tsai, C.P., Kuo, C.T., Kang, J.C., Yang, C.S., Mou, C.Y. and Chen, Y.C. (2005) Highly efficient cellular labeling of mesoporous nanoparticles in, human mesenchymal stem cells: implication for stem cell tracking. *The FASEB Journal*, **19**, 2014–2016.

102 Bouzigues, C., Levi, S., Triller, A. and Dahan, M. (2007) Single quantum dot tracking of membrane receptors. *Methods in Molecular Biology (Clifton, NJ)*, **374**, 81–92.

103 Courty, S., Luccardini, C., Bellaiche, Y., Cappello, G. and Dahan, M. (2006) Tracking individual kinesin motors in living cells using single quantum-dot imaging. *Nano Letters*, **6**, 1491–1495.

104 Courty, S., Bouzigues, C., Luccardini, C., Ehrensperger, M.V., Bonneau, S. and Dahan, M. (2006) Tracking individual proteins in living cells using single

quantum dot imaging. *Methods in Enzymology*, **414**, 211–228.

105 Foster, K.A., Galeffi, F., Gerich, F.J., Turner, D.A. and Muller, M. (2006) Optical and pharmacological tools to investigate the role of mitochondria during oxidative stress and neurodegeneration. *Progress in Neurobiology*, **79**, 136–171.

106 Dahan, M., Levi, S., Luccardini, C., Rostaing, P., Riveau, B. and Triller, A. (2003) Diffusion dynamics of glycine receptors revealed by single-quantum dot tracking. *Science*, **302**, 442–445.

107 Schwartz, M.P., Derfus, A.M., Alvarez, S.D., Bhatia, S.N. and Sailor, M.J. (2006) The smart Petri dish: a nanostructured photonic crystal for real-time monitoring of living cells. *Langmuir*, **22**, 7084–7090.

108 Gu, W., Pellegrino, T., Parak, W.J., Boudreau, R., Le Gros, M.A., Gerion, D., Alivisatos, A.P. and Larabell, C.A. (2005) Quantum-dot-based cell motility assay. *Science's STKE: Signal Transduction Knowledge Environment*, **2005**, l5.

109 Garon, E.B., Marcu, L., Luong, Q., Tcherniantchouk, O., Crooks, G.M. and Koeffler, H.P. (2007) Quantum dot labeling and tracking of human leukemic, bone marrow and cord blood cells. *Leukemia Research*, **31**, 643–651.

110 Giepmans, B.N., Deerinck, T.J., Smarr, B.L., Jones, Y.Z. and Ellisman, M.H. (2005) Correlated light and electron microscopic imaging of multiple endogenous proteins using quantum dots. *Nature Methods*, **2**, 743–749.

111 Zhou, M., Nakatani, E., Gronenberg, L.S., Tokimoto, T., Wirth, M.J., Hruby, V.J., Roberts, A., Lynch, R.M. and Ghosh, I. (2007) Peptide-labeled quantum dots for imaging GPCRs in whole cells and as single molecules. *Bioconjugate Chemistry*, **18**, 323–332.

112 Bannai, H., Levi, S., Schweizer, C., Dahan, M. and Triller, A. (2006) Imaging the lateral diffusion of membrane molecules with quantum dots. *Nature Protocols*, **1**, 2628–2634.

113 Wu, X., Liu, H., Liu, J., Haley, K.N., Treadway, J.A., Larson, J.P., Ge, N., Peale, F. and Bruchez, M.P. (2003) Immunofluorescent labeling of cancer marker Her2 and other cellular targets with semiconductor quantum dots. *Nature Biotechnology*, **21**, 41–46.

114 Hernandez-Sanchez, B.A., Boyle, T.J., Lambert, T.N., Daniel-Taylor, S.D., Oliver, J.M., Wilson, B.S., Lidke, D.S. and Andrews, N.L. (2006) Synthesizing biofunctionalized nanoparticles to image cell signaling pathways. *IEEE Transactions on NanoBioscience*, **5**, 222–230.

115 Funnell, W.R. and Maysinger, D. (2006) Three-dimensional reconstruction of cell nuclei, internalized quantum dots and sites of lipid peroxidation. *Journal of Nanobiotechnology*, **4**, 10.

116 Jaiswal, J.K. and Simon, S.M. (2007) Optical monitoring of single cells using quantum dots. *Methods in Molecular Biology (Clifton, NJ)*, **374**, 93–104.

117 Jaiswal, J.K., Mattoussi, H., Mauro, J.M. and Simon, S.M. (2003) Long-term multiple color imaging of live cells using quantum dot bioconjugates. *Nature Biotechnology*, **21**, 47–51.

118 Chan, W.C. and Nie, S. (1998) Quantum dot bioconjugates for ultrasensitive nonisotopic detection. *Science*, **281**, 2016–2018.

119 Mattheakis, L.C., Dias, J.M., Choi, Y.J., Gong, J., Bruchez, M.P., Liu, J. and Wang, E. (2004) Optical coding of mammalian cells using semiconductor quantum dots. *Analytical Biochemistry*, **327**, 200–208.

120 Lagerholm, B.C. (2007) Peptide-mediated intracellular delivery of quantum dots. *Methods in Molecular Biology (Clifton, NJ)*, **374**, 105–112.

121 Dubertret, B., Skourides, P., Norris, D.J., Noireaux, V., Brivanlou, A.H. and Libchaber, A. (2002) *In vivo* imaging of quantum dots encapsulated in phospholipid micelles. *Science*, **298**, 1759–1762.

122 Cambi, A., Lidke, D.S., Arndt-Jovin, D.J., Figdor, C.G. and Jovin, T.M. (2007) Ligand-conjugated quantum dots monitor antigen uptake and processing by dendritic cells. *Nano Letters*, **7**, 970–977.

123 Hardman, R. (2006) A toxicologic review of quantum dots: toxicity depends on physicochemical and environmental factors. *Environmental Health Perspectives*, **114**, 165–172.

124 Hoshino, A., Fujioka, K., Oku, T., Nakamura, S., Suga, M., Yamaguchi, Y., Suzuki, K., Yasuhara, M. and Yamamoto, K. (2004) Quantum dots targeted to the assigned organelle in living cells. *Microbiology and Immunology*, **48**, 985–994.

14
Self-Assembling Peptide-Based Nanostructures for Regenerative Medicine

Ramille M. Capito, Alvaro Mata, and Samuel I. Stupp

14.1
Introduction

The goal of regenerative medicine is to develop therapies that can promote the growth of tissues and organs in need of repair as a result of trauma, disease or congenital defects. For most of the patient population this means regeneration of our bodies in adulthood, although there are also many critical pediatric needs in regenerative medicine. One specific target that would deeply impact the human condition is regeneration of the central nervous system (CNS). This would bring a higher quality of life to individuals paralyzed as a result of spinal cord injury, brought into serious dysfunction by stroke, afflicted with Parkinson's and Alzheimer's diseases, or those blind as a result of macular degeneration or retinitis pigmentosa. Another area that would benefit from regenerative medicine is heart disease, which continues to be one of the most dominant sources of premature death in humans. Here, the potential to regenerate myocardium would have a great impact on clinical outcomes. Many additional important targets exist. The regeneration of insulin-producing pancreatic β cells would bring a higher quality of life to individuals suffering from diabetes. Damage to cartilage – a critical tissue in correct joint function – is an enormous source of pain and compromised agility for many individuals, especially in societies that value a physically active life style for as long as possible. Other musculoskeletal tissues such as bone, intervertebral disc, tendon, meniscus and ligament all remain major therapeutic challenges in regenerative medicine. Another emerging target that could have an enormous impact is the regeneration of teeth, as this would prevent the need for dentures and other dental implants. All of these important targets in regenerative medicine would not only raise the quality of life for many individuals worldwide, but they would also have, for obvious reasons, a significant economic impact.

The development of effective regenerative medicine strategies generally includes the use of cells, soluble regulators (e.g. growth factors or genes) and

Nanotechnology, Volume 5: Nanomedicine. Edited by Viola Vogel

ISBN: 978-3-527-31736-3

scaffold technologies. In their natural environment, mammalian cells live surrounded by a form of solid or fluid matrix composed of structural protein fibers (i.e. collagen and elastin), adhesive proteins (i.e. fibronectin and laminin), soluble proteins (i.e. growth factors) and other biopolymers (i.e. polysaccharides), all of which have specific inter-related roles in the structure and normal function of the extracellular matrix (ECM). The creation of biomimetic artificial matrices represents a common theme in designing materials for regenerative medicine therapies, and stems from the idea that providing a more natural three-dimensional (3-D) environment can preserve cell viability and encourage cell differentiation and matrix synthesis. The nanoscale design of biomaterials, with particular attention to dimension, shape, internal structure and surface chemistry, may more effectively emulate the very sophisticated architecture and signaling machinery of the natural ECM for improved regeneration.

Strategies utilizing self-assembled supramolecular aggregates, macromolecules and even inorganic particles could be used to design a signaling machinery *de novo* that initiates regeneration events which do not occur naturally in mammalian biology. Self-assembly – a bioinspired phenomenon which involves the spontaneous association of disordered components into well-defined and functionally organized structures [1] – can play a major role in creating sophisticated and biomimetic biomaterials for regenerative therapies [2–7]. In molecular systems, self-assembly implies that molecules are programmed by design to organize spontaneously into supramolecular structures held together through noncovalent interactions, such as electrostatic or ionic interactions, hydrogen bonding, hydrophobic interactions and van der Waals interactions. Large collections of these relatively weak bonds compared to covalent bonds can result in very stable structures.

The first fundamental reason for a link between self-assembly and regenerative medicine is the potential to create multifunctional artificial forms of an ECM starting with liquids. Such liquids could contain dissolved molecules or pre-assembled nanostructures, and they could then be introduced by injection at a specific site or targeted through the circulation. Following self-assembly, a solid matrix could mechanically support cells and also signal them for survival, proliferation, differentiation or migration. Alternatively, the self-assembled solid matrix could be designed to recruit specific types of cells in order to promote a regenerative biological event, or serve as cell delivery vehicles by localizing them in 3-D environments within tissues and organs. These self-assembling molecules could also be used to modify the surfaces of solid implants in order to render them bioactive [1, 8, 9]. The 'bottom-up' approach that is possible using self-assembly can permit the creation of an architecture that multiplexes signals or tunes their concentration per unit area. This versatility makes self-assembling systems ideal for creating optimal materials for regenerative medicine therapies.

In this chapter, we focus on the use of self-assembling nanostructures – in particular, peptide-based molecules – which are currently being developed for regenerative medicine applications. Although many of these technologies are relatively new, much very promising biological data – both *in vitro* and *in vivo* – demonstrating

the promise of these systems in regenerative medicine is already available. Two currently important research areas in this field include:

- The development of self-assembling injectable bioactive scaffolds that have the ability to mimic the natural 3-D ECM of cells.
- The nanoscale surface modification of surfaces or 3-D tissue engineering scaffolds using self-assembling molecules to create bioactive implants and devices.

14.2 Self-Assembling Synthetic Peptide Scaffolds

Peptides are among the most useful building blocks for creating self-assembled structures at the nanoscale; they possess the biocompatibility and chemical versatility that are found in proteins, yet they are more chemically and thermally stable [10]. They can also be easily synthesized on a large scale by using conventional chemical techniques, and designed to contain functional bioactive sequences. A variety of short peptide molecules have been shown to self-assemble into a wide range of supramolecular structures including nanofibers, nanotubes, nanospheres and nanotapes. Some self-assembling nanostructures have been used successfully to generate injectable scaffolds with an extremely high water content and architectural features that mimic the natural structure of the ECM. These self-assembling scaffolds show great potential as 3-D environments for cell culture and regenerative medicine applications, and also as vehicles for drug, gene or protein delivery.

14.2.1 β-Sheet Peptides

Aggeli and colleagues demonstrated that the biological peptide β-sheet motif can be used to design oligopeptides that self-assemble into semi-flexible β-sheet nanotapes [11]. Depending on the intrinsic chirality of the peptides and concentration, these nanostructures can further assemble into twisted ribbons (double tapes), fibrils (twisted stacks of ribbons) and fibers (fibrils entwined edge-to-edge) (Figure 14.1a) [12]. The assembly process is principally driven by hydrogen bonding along the peptide backbone and interactions between specific side chains [13]. At sufficiently high peptide concentrations, these structures can become entangled to form gels, the viscoelastic properties of which can be altered by controlling the pH, by applying a physical (shear) stress, or by altering the peptide concentration. As in other peptide self-assembling systems, the hierarchical assembly can be altered by the addition and position of charged amino acids within the peptide sequence that is highly controlled by changes in pH [12] (Figure 14.1b and c). It has also been shown that mixing aqueous solutions of cationic and anionic peptides that have complementary charged side chains and a propensity to form antiparallel β-sheets, results in the spontaneous self-assembly of fibrillar networks and hydrogels that are robust to variations in pH and peptide concentration [14].

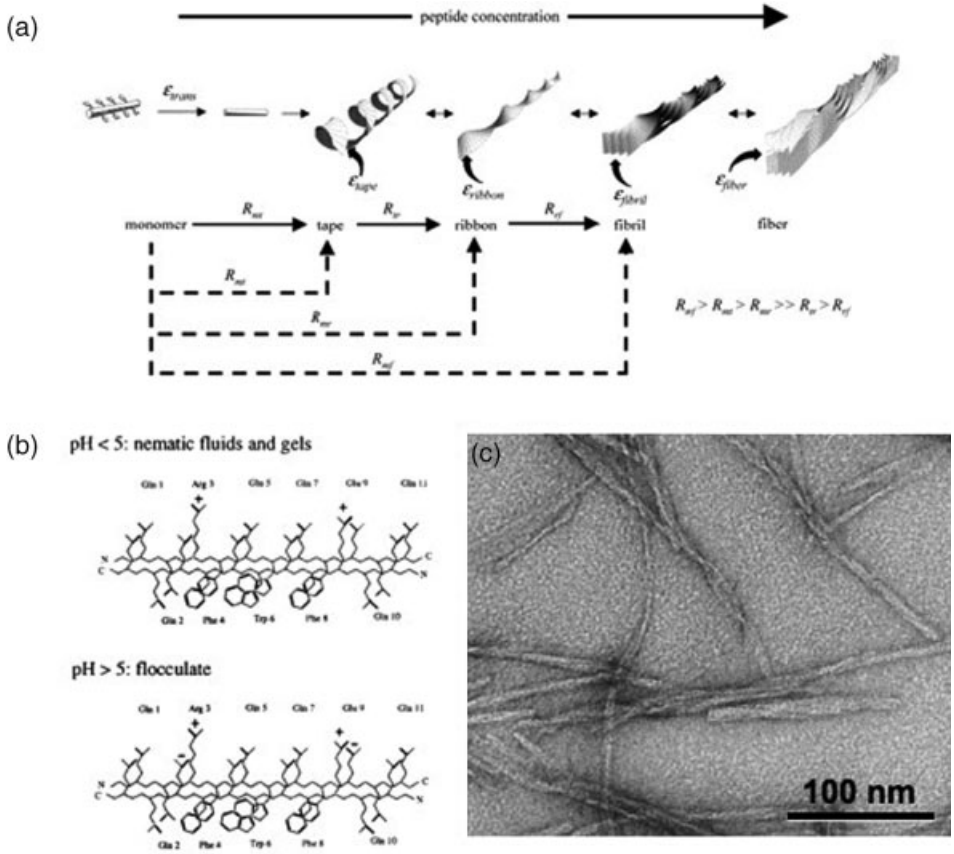


Figure 14.1 (a) The global equilibrium configurations obtained by the hierarchical self-assembly of β-sheet-forming peptides. The set of energy parameters ε_j correspond to the free energy differences per peptide molecule between successive structures. The 'critical' peptide concentrations at which each new configuration begins to appear is determined by the ε_j. The R_{ij} are the conversion rates both between and to the various configurations. The process depicted by solid arrows represents the dissolution route of lyophilized solid at constant pH, while the dashed arrows represent the direct and simultaneous conversion of monomer to tapes, ribbons, fibrils and fibers when the respective critical concentrations governing their self-assembly are instantaneously switched by pH change to values above the absolute peptide concentration in solution; (b) Electrostatic charge distribution on P11-2 dimer in an antiparallel β-sheet tape-like substructure: top, pH < 5; bottom, pH > 5; (c) Transmission electron microscopy image of a gel. (Reproduced with permission from Ref. [12]; © 2003, American Chemical Society.)

These β-sheet peptide nanostructures have been studied for the treatment of enamel decay [13]. *In vitro*, extracted human premolar teeth (containing caries-like lesions) were exposed to several cycles of demineralizing (acidic conditions) and remineralizing solutions (neutral pH conditions). Application of the self-assembling peptides to the defects significantly decreased demineralization during exposure

to acid and increased remineralization at neutral pH, resulting in a net mineral gain of the lesions compared to untreated controls [13]. Furthermore, when the peptide gels were incubated for one week in mineralizing solutions, *de novo* nucleation of hydroxyapatite by the nanostructures was observed [13].

In another application, these peptides were evaluated as an alternative injectable joint lubricant to hyaluronic acid (HA) for the treatment of osteoarthritis [15]. A β-sheet peptide designed to have molecular, mesoscopic and rheological properties that most closely resembled HA, performed similarly to HA in healthy static and dynamic friction tests, but not as well in friction tests with damaged cartilage [15]. The optimization of these peptides may result in a new alternative viscosupplementation treatment for degenerative osteoarthritis.

14.2.2 β-Hairpin Peptides

Another peptide design that exploits β-sheet nanostructure formation into a hydrogel network is composed of strands of alternating hydrophilic and hydrophobic residues flanking an intermittent tetrapetide [16–21] (Figure 14.2a and b). These peptides are designed so that they are fully dissolved in aqueous solutions in random coil conformations. Under specific stimuli, the molecules can be triggered to fold into a β-hairpin conformation that undergoes rapid self-assembly into a β-sheet-rich, highly crosslinked hydrogel. The molecular folding event – where one face of the β-hairpin structure is lined with hydrophobic valines and the other with hydrophilic lysines – is governed by the arrangement of polar and nonpolar residues within the sequence. Subsequent self-assembly of the individual hairpins occurs by hydrogen bonding between distinct hairpins and hydrophobic association of the hydrophobic faces. One such peptide was designed to self-assemble under specific pH conditions [16]. Under basic conditions, the peptide intramolecularly folds into the hairpin structure and forms a hydrogel. Unfolding of the hairpins and dissociation of the hydrogel structure can be triggered when the pH is subsequently lowered below the pK_a of the lysine side chains, where unfolding is a result of the intrastrand charge repulsion between the lysine residues [16]. Rheological studies indicate that these β-hairpin hydrogels are both rigid and shear-thinning; however, the mechanical strength is quickly regained after cessation of shear [16] (Figure 14.2a).

These gels can also be triggered to self-assemble when the charged amino acid residues within the sequence are screened by ions [22]. If a positively charged side chain of lysine is replaced by a negatively charged side chain of glutamic acid, the overall peptide charge is decreased and the peptide can be more easily screened, resulting in a much faster self-assembly [21]. The kinetics of hydrogelation were found to be significant for the homogeneous distribution of encapsulated cells within these types of self-assembling gels [21] (Figure 14.2c). Thermally reversible, self-assembling peptides were also synthesized by replacing specific valine residues with threonines to render the peptides less hydrophobic [17]. At ambient temperature and slightly basic pH, the peptide is unfolded; however, upon heating the peptide

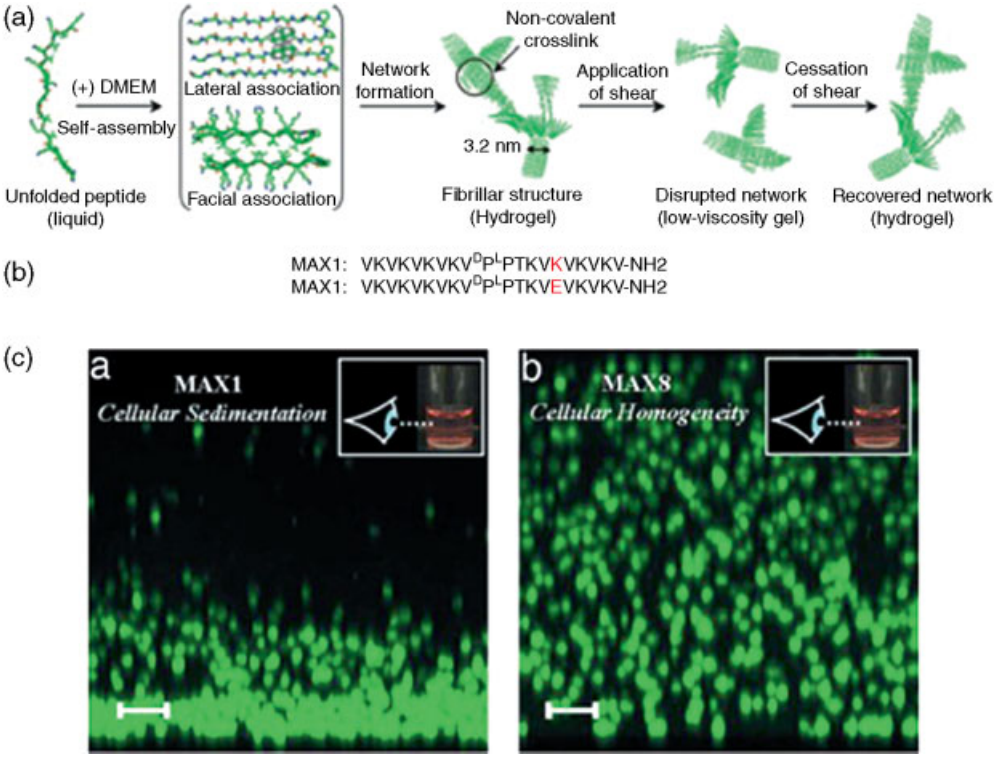


Figure 14.2 Self-assembly, shear-thinning and self-healing mechanism allowing rapid formation of β-hairpin hydrogels that can be subsequently syringe-delivered. (a) Addition of Dulbecco's modified Eagle medium (DMEM; pH 7.4, 37 °C) to a buffered solution of unfolded peptide induces the formation of a β-hairpin structure that undergoes lateral and facial self-assembly affording a rigid hydrogel with a fibrillar supramolecular structure. Subsequent application of shear stress disrupts the noncovalently stabilized network, leading to the conversion of hydrogel to a low-viscosity gel. Upon cessation of shear stress, the network structure recovers, converting the liquid back to a rigid hydrogel; (b) Peptide sequences of MAX8 and MAX1; (c) Encapsulation of mesenchymal C3H10t1/2 stem cells in 0.5 wt% MAX1 and MAX8 hydrogels. Shown are LSCM z-stack images (viewed along the *y*-axis) showing the incorporation of cells into a MAX1 gel leading to cell sedimentation (panel a) and into a MAX8 gel resulting in cellular homogeneity (panel b). Cells are prelabeled with cell tracker green to aid visualization (scale bars = 100 μm). (Reproduced with permission from Ref. [21]; © 2007, National Academy of Sciences.)

folds and assembles via hydrophobic collapse as the temperature dehydrates the nonpolar residues within the peptide [17]. Yet another rendition of this peptide self-assembles via light activation [20]. In this design, a photocaged peptide is incorporated within the peptide sequence, with β-hairpin folding and subsequent hydrogelation occurring only when the photocage is released upon irradiation of the sample [20].

In vitro studies have shown that these types of β-hairpin hydrogels can support the survival, adhesion and migration of NIH 3T3 fibroblasts [20–22], and can be used to

encapsulate mesenchymal stem cells (MSCs) and hepatocytes [21]. Another study also showed that these gels have an inherent antibacterial activity, with selective toxicity to bacterial cells versus mammalian cells [23].

14.2.3 Block Copolypeptides

Deming and colleagues have developed diblock copolypeptide amphiphiles containing charged and hydrophobic segments that self-assemble into rigid hydrogels and can remain mechanically stable even at high temperatures (up to 90 °C) [24, 25] (Figure 14.3). These hydrogels were also found to recover rapidly after an applied stress, attributed to the relatively low molecular mass of the copolypetides, enabling rapid molecular organization. Their amphiphilic characteristics, architecture (diblock versus triblock) and block secondary structure (e.g. α-helix, β-strand or random coil) were found to play important roles in the gelation, rheological and morphological properties of the hydrogel [24–26]. One type of block copolypeptide consists of a hydrophobic block of poly-L-lysine and a shorter hydrophobic block of poly-L-leucine [26]. The helical secondary structure of the poly-L-leucine blocks was shown to be instrumental for gelation, while the hydrophilic polyelectrolyte segments helped to stabilize the twisted fibril assemblies by forming a corona around the hydrophobic core [26] (Figure 14.3).

In vitro studies using mouse fibroblasts revealed that, at concentrations below gelation, lysine-containing diblocks were cytotoxic to the cells, whereas glutamic

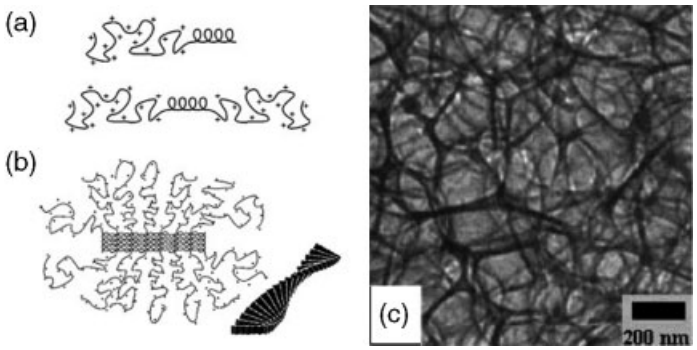


Figure 14.3 (a) Diblock (top) and triblock (bottom) copolypeptide architectures. The hydrophobic leucine block exhibits α-helical secondary structure, and the charged polyelectrolyte block has a stretched-coil configuration; (b) Packing of amphiphilic diblock copolypeptide molecule fibrils, the cross-section being shown in detail and the inset schematically depicting how the cross-sectional layers assemble into twisted fibers (for clarity, only the helices are drawn). (Reproduced with permission from Ref. [26]; © 2004, American Chemical Society.); (c) Cryogenic transmission electron microscopy image of 5.0 wt% $K_{180}(LV)_{20}$ showing the interconnected membrane, cellular nanostructure of gel matrix (dark) surrounded/filled by vitreous water (light). (Reproduced with permission from Ref. [25]; © 2002, American Chemical Society.)

acid-containing peptides were not cytotoxic [27]. In gel form, however, both lysine and glutamic acid-based diblocks were noncytotoxic, although the scaffolds did not support cell attachment or proliferation. This demonstrates how molecular design and charge can significantly affect the cytoxicity and biological activity of the resulting self-assembled material. Future research is directed towards covalently incorporating bioactive sites within these hydrogels in order to increase cellular attachment and enhance the biological response [27].

14.2.4 Ionic Self-Complementary Peptides

Another class of self-assembling peptide molecules developed by Zhang *et al*. was designed to include alternating positive and negative amino acid repeats within the peptide sequence [28, 29]. These oligopeptides associate to form stable fibrillar nanostructures in aqueous solution through β-sheet formation, due to their hydrophilic and hydrophobic surfaces and complementary ionic bonding between the oppositely charged residues. Upon addition of monovalent cations or physiological media, they form hydrogels composed of interwoven nanofibers (Figure 14.4a) [29, 30]. Studies have shown that oligopeptide length [31] and side-chain hydrophobicity [32] were important variables that affected the self-assembly and the resulting gel properties.

Several *in vitro* and *in vivo* studies have been conducted investigating the ability of these scaffolds to support cell attachment [29], survival, proliferation and differentiation for neural [30, 33–35], blood vessel [36–38], myocardial [39–42], liver [43], cartilage [44] and bone tissue regeneration [45, 46]. For the treatment of neural defects, primary mouse neuron cells encapsulated within the hydrogels were able to attach to the nanofiber matrix and showed extensive neurite outgrowth [30]. Furthermore, peptides implanted *in vivo* did not elicit a measurable immune response or tissue inflammation [30]. In a hamster model, the peptide scaffolds were shown to regenerate axons and reconnect target tissues in a severed optic tract that resulted in the restoration of visual function [33]. Likewise, the peptide scaffolds caused an effective promotion of cell migration, blood vessel growth and axonal elongation when implanted with neural progenitor cells and Schwann cells in the transected dorsal column of the rat spinal cord [47] (Figure 14.4b).

For the treatment of myocardial infarction, Davis *et al*. injected peptide scaffolds into rat myocardium and observed the recruitment of endothelial progenitor cells and vascular smooth muscle cells into the injection site that appeared to form functional vascular structures [40]. Biotinylated nanofibers were subsequently used to deliver insulin-like growth factor 1 (IGF-1) *in vivo* over prolonged periods (28 days) and were shown to significantly improve systolic function after myocardial infarction when delivered with transplanted cardiomyocytes [39]. Other *in vivo* studies, which delivered platelet-derived growth factor (PDGF)-BB with the self-assembling nanofibers in a rat myocardial infarct model, showed decreased cardiomyocyte death, reduced infarct size and a long-term improvement in cardiac performance after infarction, without systemic toxicity [41, 42].

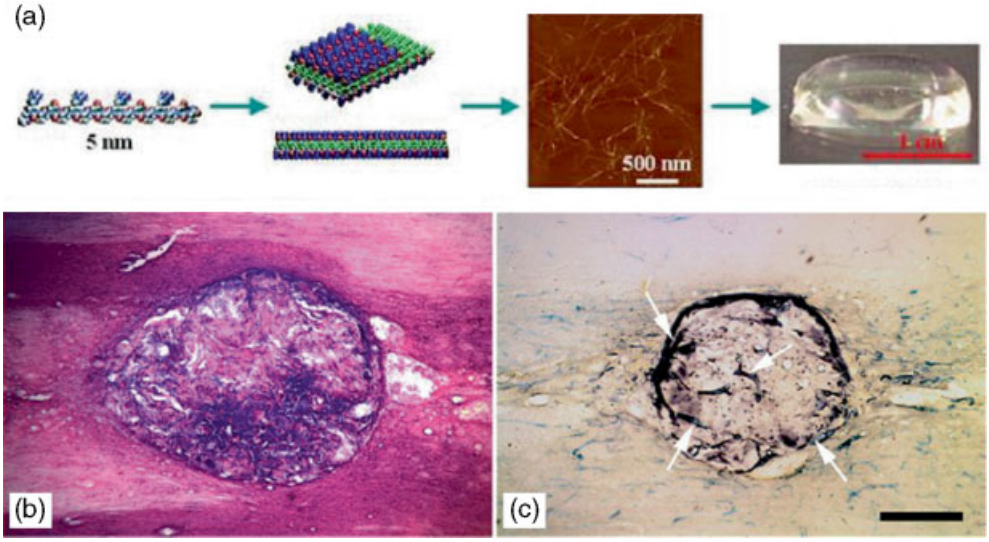


Figure 14.4 (a) Ionic self-complementary peptide consisting of 16 amino acids, ~5 nm in size, with an alternating polar and nonpolar pattern. These peptides form stable β-strand and β-sheet structures the side chains of which partition into two sides, one polar and the other nonpolar. These undergo self-assembly to form nanofibers with the nonpolar residues inside (green) and + (blue) and − (red) charged residues forming complementary ionic interactions, like a checkerboard. These nanofibers form interwoven matrices that further form a scaffold hydrogel with very high water content (>99.5%). (Reproduced with permission from Zhang, S. (2003) Fabrication of novel biomaterials through molecular self-assembly. *Nature Biotechnology*, **21**, 1171–8; © Wiley-VCH Verlag GmbH & Co. KGaA.); (b, c) Implantation of precultured peptide gels into the injured spinal cord of GFP-transgenic rats. (b) Hematoxylin and eosin staining showed a high level of integration between the implants and host, although in most cases a few small cysts were found near the implants; (c) Alkaline phosphatase histochemistry staining showed that blood vessels grew into the implants (arrows). Scale bar = 500 μm. (Reproduced with permission from Ref. [34]; © 2007, Elsevier Limited.)

14.2.5
Fmoc Peptides

A more recently developed class of self-assembling peptides that uses fluorenylmethyloxycarbonyl (Fmoc) -protected di- and tri-peptides have been shown to form highly tunable hydrogel structures (Figure 14.5a). The formation of these gels can be achieved either by pH adjustment [48] (Figure 14.5b) or by a reverse-hydrolysis enzyme action [49] (Figure 14.5c). Assembly occurs via hydrogen bonding in β-sheet arrangement and by π–π stacking of the fluorenyl rings that also stabilize the system [48] (Figure 14.5b). A number of sheets then twist together to form nanotubes (Figure 14.5d).

The results of *in vitro* studies indicated that these hydrogels can support chondrocyte survival and proliferation in both two and three dimensions [48]. It was also observed that cell morphology varied according to the nature of the molecular structure [48].

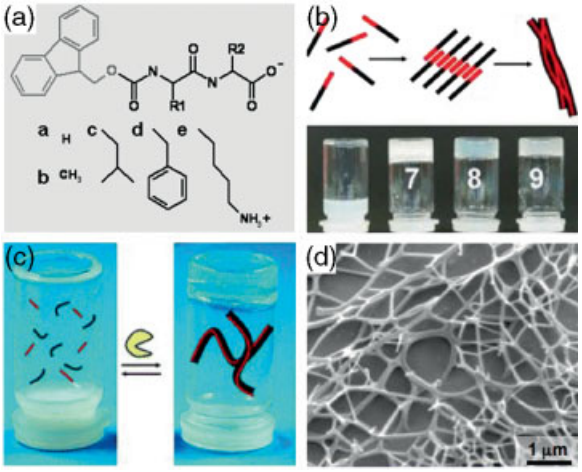


Figure 14.5 (a) Molecular structure of Fmoc–dipeptides. The R groups are the amino acids Gly (a), Ala (b), Leu (c), Phe (d) and Lys (e); (b) Proposed self-assembly mechanism (top): Fmoc groups stack through π–π interactions, and the resulting molecular stacks further assemble to form nanofibers. Self-supporting gels can be formed by manipulation of pH or by reverse-hydrolysis enzyme action (c); (d) Cryogenic scanning electron microscopy image of nanofibrous material obtained by self-assembly. (Panels (a)–(c) reproduced with permission from Ref. [49]; © 2006, American Chemical Society; panel (d) reproduced with permission from Ref. [48]; © 2007, The Biochemical Society.)

14.2.6
Peptide Amphiphiles

Peptide amphiphiles (PAs) are self-assembling molecules that also use hydrophobic and hydrophilic elements to drive self-assembly. There are different types of peptide amphiphiles that can assemble into a variety of nanostructures such as spherical micelles, fibrils, tubes or ribbons [50]. One unique PA design, which forms high-aspect ratio cylindrical nanofibers, has been exclusively studied during the past decade for regenerative medicine applications. These molecules are particularly distinguished from the other peptide systems described above, in that their amphiphilic nature derives from the incorporation of a hydrophilic head group and a hydrophobic alkyl tail, as opposed to molecules consisting of all amino acid residues with resultant hydrophilic and hydrophobic faces.

Stupp and colleagues have developed a family of amphiphilic molecules that can self-assemble from aqueous media into 3-D matrices composed of supramolecular nanofibers [4–6, 9, 51–59]. These molecules consist of a hydrophilic peptide segment which is bound covalently to a highly hydrophobic alkyl tail found in ordinary lipid molecules. The alkyl tail can be located at either the C or N terminus [51], and can also be constructed to contain branched structures [55]. In Stupp *et al.*'s specific design, the peptide region contains a β-sheet-forming peptide domain close to the hydrophobic segment and a bioactive peptide sequence (Figure 14.6a). Upon changes in

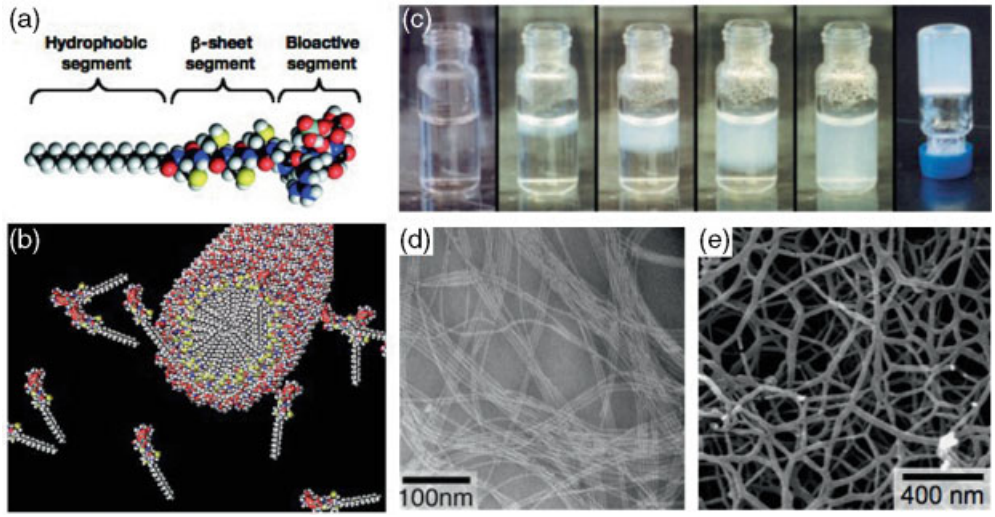


Figure 14.6 (a) General structure of PA molecules; (b) Illustrated self-assembly of PA molecules into nanofibers with hydrophobic cores; (c) Time sequence of pH-controlled PA self-assembly. From left to right: PA molecule dissolved in water at a concentration of 0.5% by weight at pH 8 is exposed to HCl vapor. As the acid diffuses into the solution a gel phase is formed, which self-supports upon inversion; (d) Transmission electron microscopy image of PA nanofibers. (Reproduced with permission from Ref. [5]; © 2002, National Academy of Sciences.); (e) Scanning electron microscopy image of PA nanofiber network. (Reproduced with permission from Ref. [7]; © 2005, Materials Research Society.)

pH or the addition of multivalent ions, the structure of these molecules drives their assembly into cylindrical nanofibers through hydrogen bonding into β-sheets and hydrophobic collapse of alkyl tails away from the aqueous environment to create nanofibers with a hydrophobic core (Figure 14.6b). This cylindrical nanostructure allows the presentation of high densities of bioactive epitopes at the surface of the nanofibers [6], whereas, if peptides were assembled into twisted sheets or tubes, this type of orientational biological signaling would not be possible [7]. These systems can also be used to craft nanofibers containing two or more PA molecules that can effectively coassemble, thus offering the possibility of multiplexing different biological signals within a single nanofiber [56].

The presence of a net charge in the peptide sequence ensures that the molecules or small β-sheet aggregates remain dissolved in water, inhibiting self-assembly through coulombic repulsion. Self-assembly and gelation is subsequently triggered when the charged amino acid residues are electrostatically screened or neutralized by pH adjustment, or by the addition of ions (Figure 14.6c–e). The growth of nanofibers can therefore be controlled by changing the pH or raising the concentration of screening electrolytes in the aqueous medium [7]. Growth and bundling of the nanofibers eventually lead to gelation of the PA solution. *In vivo*, ion concentrations present in physiological fluids can be sufficient to induce the formation of PA nanostructures [54]. Thus, a minimally invasive procedure could be designed

with these systems through a simple injection of the PA solution that spontaneously self-assembles into a bioactive scaffold at the desired site. Over time, the small molecules composing the nanofibers should biodegrade into amino acids and lipids, thus minimizing the potential problems of toxicity or immune response [54].

The results of both *in vitro* and *in vivo* studies have shown that these PA systems can serve as an effective analogue of the ECM by successfully supporting cell survival and attachment [60, 61], mediating cell differentiation [6] and promoting regeneration *in vivo* [57]. In efforts to address neural tissue regeneration for the repair of a spinal cord injury or treatment of extensive dysfunction as a result of stroke or Parkinson's disease, Stupp and colleagues have designed PAs to display the pentapeptide epitope isoleucine-lysine-valine-alanine-valine (IKVAV). This particular peptide sequence is found in the protein laminin, and has been shown to promote neurite sprouting and to direct neurite growth [6]. When neural progenitor cells were encapsulated within this PA nanofiber network, the cells more rapidly differentiated into neurons compared to using the protein laminin or the soluble peptide (Figure 14.7a and b). The PA scaffold was also found to discourage the development of astrocytes, a type of cell in the CNS which is responsible for the formation of glial scars that prevent regeneration and recovery after spinal cord injury. In this same study, the

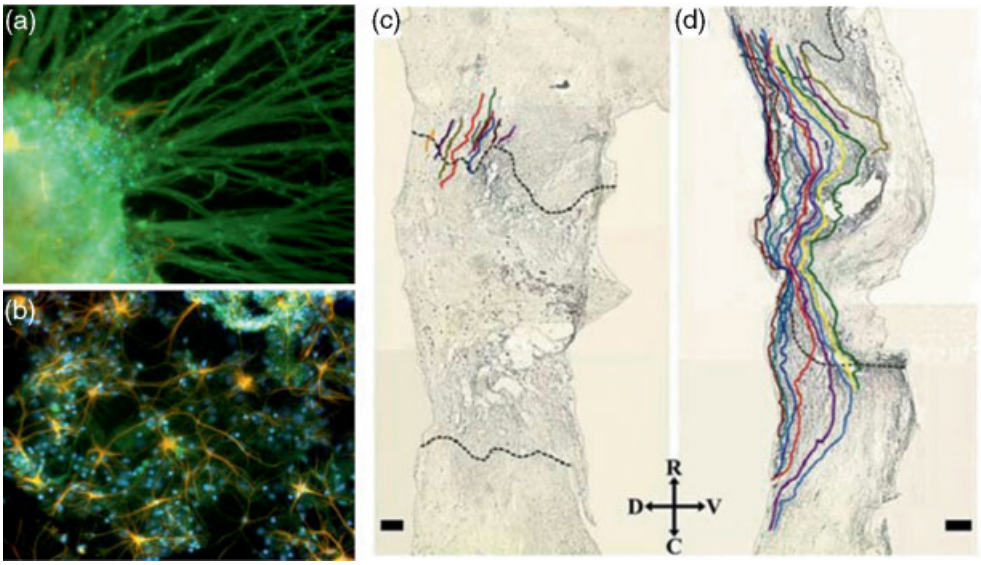


Figure 14.7 (a) Immunocytochemistry of a neuroprogenitor cell neurosphere encapsulated in an IKVAV-PA nanofiber network at 7 days, showing a large extent of neurite outgrowth; (b) Neural progenitor cells cultured on laminin-coated coverslips at 7 days. The prevalence of astrocytes is apparent. (Reproduced with permission from Ref. [6]; © 2004, American Association for the Advancement of Science.); (c, d) Representative Neurolucida tracings of labeled descending motor fibers within a distance of 500 µm rostral of the lesion in vehicle-injected (c) and IKVAV PA-injected (d) animals. The dotted lines demarcate the borders of the lesion. Scale bars = 100 µm. (Reproduced with permission from Ref. [62]; © 2008, Society for Neuroscience.)

density of epitopes displayed on the nanofibers proved to be a significant variable in the ability to induce rapid and selective differentiation of cells encapsulated within the PA gels. Furthermore, *in vivo* studies in which this self-assembling neural nanofiber scaffold was injected within a spinal cord injury in a rat model showed better functional improvement and axonal elongation through the injury site compared to controls [62] (Figure 14.7c and d).

Another PA molecule was designed to self-assemble upon the addition of heparin, a biopolymer that binds to angiogenic growth factors [57]. The resultant nanofibers displayed heparin chains on the periphery, which orient proteins on the surface for cell signaling. In an *in vivo* rabbit corneal model, the heparin-binding PA nanostructures, administered with only nanogram quantities of angiogenic growth factors, was sufficient to stimulate extensive neovascularization compared to controls (Figure 14.8). When using the same PA system, Kapadia *et al.*, was also able to create self-assembling nitric oxide (NO)-releasing nanofiber gels for the prevention of neointimal hyperplasia [63]. Using a rat carotid artery model, the group showed that the NO-releasing PA gels significantly reduced neointimal hyperplasia, inhibited inflammation and stimulated re-endothelialization compared to controls.

The value of this nanotechnology lies in its self-assembly code, which yields nanofibers that can be designed to have a great diversity of bioactive signals [64, 65]. For example, PAs have been successfully synthesized to contain binding groups for growth factors by phage display technology [51]. The inclusion of growth factor binding domains enables a greater retention of incorporated growth factors within the scaffold, or even the 'capture' of desired endogenous growth factors localized

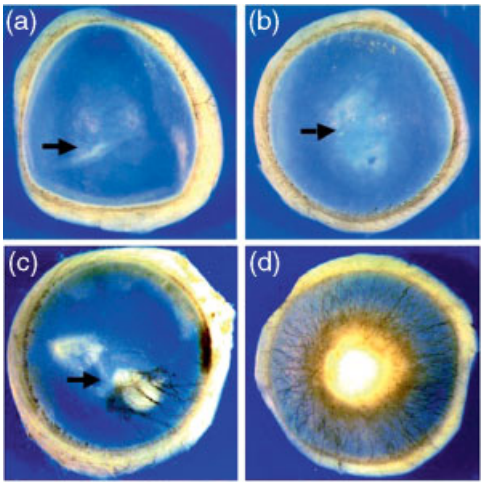


Figure 14.8 *In vivo* angiogenesis assay in a rat cornea 10 days after the placement of various materials at the site indicated by the black arrow. Growth factors alone (a) and heparin with growth factors (b) showed little to no neovascularization. Collagen, heparin and growth factors (c) showed some neovascularization; (d) Heparin-nucleated PA nanofiber networks with growth factors showed extensive neovascularization. (Reproduced with permission from Ref. [57]; © 2006, American Chemical Society.)

at the implant site, thus eliminating the need for exogenous growth factor supplementation altogether. Hartgerink *et al.* synthesized PAs with a combination of biofunctional groups including a cell-mediated enzyme-sensitive site, a calcium-binding site and a cell-adhesive ligand [66]. The incorporation of an enzyme-specific cleavage site allows cell-mediated proteolytic degradation of the scaffold for cell-controlled migration and matrix remodeling. *In vitro* studies demonstrated that these PA scaffolds do degrade in the presence of proteases, and that the morphology of cells encapsulated within the nanofiber scaffolds was dependent on the density of the cell-adhesive ligand, with more elongated cells observed in gels with a higher adhesive ligand density [66].

To date, hundreds of peptide amphiphile nanofibers designs are known, including those that nucleate hydroxyapatite with some of the crystallographic features found in bone [4], increase the survival of cultured islets for the treatment of diabetes, bind to various growth factors [51], contain integrin-binding sequences [61], incorporate contrast agents for fate mapping of PA nanostructures [52], and have pro-apoptotic sequences for cancer therapy, among many others. Research investigating the development of hybrid materials using these versatile PA systems is also emerging. For example, PA nanofibers were integrated within titanium foams as a means to promote bone ingrowth or bone adhesion for improved orthopedic implant fixation (Figure 14.9) [1]. Preliminary *in vivo* results implanting these PA–Ti hybrids within bone defects in a rat femur demonstrated *de novo* bone formation around and inside the implant, vascularization in the vicinity of the implant, and no cytotoxic response [1]. Another type of hybrid system developed by Hartgerink *et al.* includes hydrogels that contain a mixture of PA and phospholipid (Figure 14.10) [67]. The phospholipid inclusions within the PA nanostructure were found to modulate

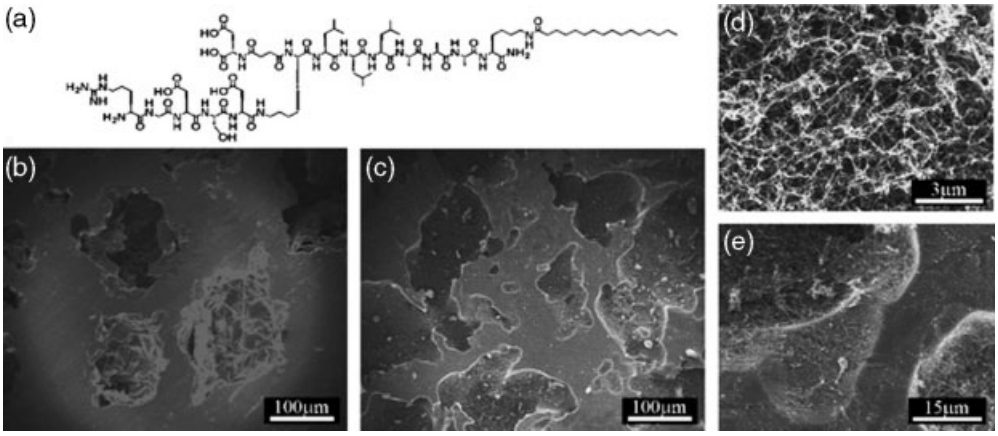


Figure 14.9 (a) Chemical structure of the peptide amphiphile (PA) used to infiltrate and fill the pores of the Ti–6Al–4V foam. Scanning electron microscopy (SEM) images of (b) the bare Ti–6Al–4V foam; (c) Ti–6Al–4V foam filled with PA gel; (d) higher magnification of the self-assembled PA nanofibers forming a 3-D matrix within the pores; (e) Higher magnification of the PA coating the Ti–6Al–4V foam surface and filling the pores. (Reproduced with permission from Ref. [1]; © 2008, Elsevier Limited.)

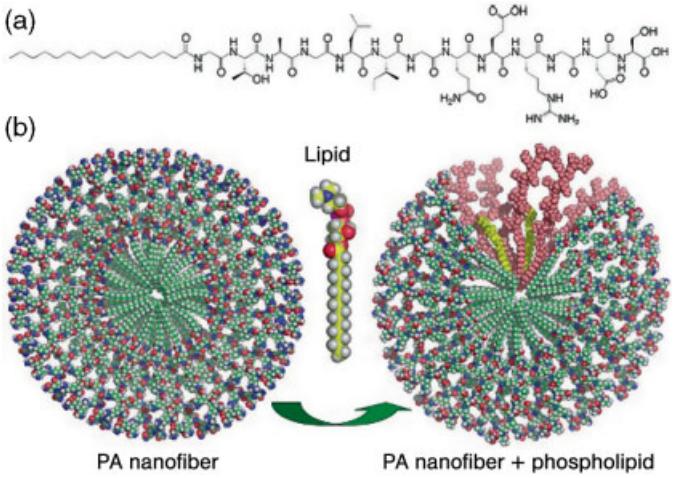


Figure 14.10 (a) Chemical structure of the PA and (b) cross-section of a PA fiber and a PA fiber containing 6.25 mol% of lipid (yellow). Highlighted in pink are the PA molecules situated adjacent to the lipid molecules. (Reproduced with permission from Ref. [67]; © 2006, American Chemical Society.)

the peptide secondary structure as well as the mechanical properties of the hydrogel, with little change in the nanostructure. This composite system enables the optimization of mechanical and chemical properties of the hydrogel by simple adjustment of the PA to phospholipid ratios [67].

The ability to access new mechanisms to control self-assembly across the scales, and not just at the nanostructure level, offers new possibilities for regenerative therapies as bioactive functions can be extended by design into microscopic – and even macroscopic – dimensions. One system involves the self-assembly of hierarchically ordered materials at an aqueous interface resulting from the interaction between small, charged self-assembling PA molecules and oppositely charged high-molar mass biopolymers [68]. A PA–polymer sac can be formed instantly by injecting the polymer directly into the PA solution (Figure 14.11a). The interfacial interaction between the two aqueous liquids allows the formation of relatively robust membranes with tailorable size and shape (Figure 14.11b), self-sealing and suturable sacs (Figure 14.11c–f), as well as continuous strings (Figure 14.11g). The membrane structure grows to macroscopic dimensions with a high degree of hierarchical order across the scales. Studies have demonstrated that the sac membrane is permeable to large proteins, and therefore can be successfully used to encapsulate cells (Figure 14.11h). *In vitro* studies of mouse pancreatic islets (Figure 14.11i) and human MSCs (Figure 14.11j) cultured within the sacs showed that these structures can support cell survival and can be effective 3-D environments for cell differentiation. The unique structural and physical characteristics of these novel systems offer significant potential in cell therapies, drug diagnostics and regenerative medicine applications.

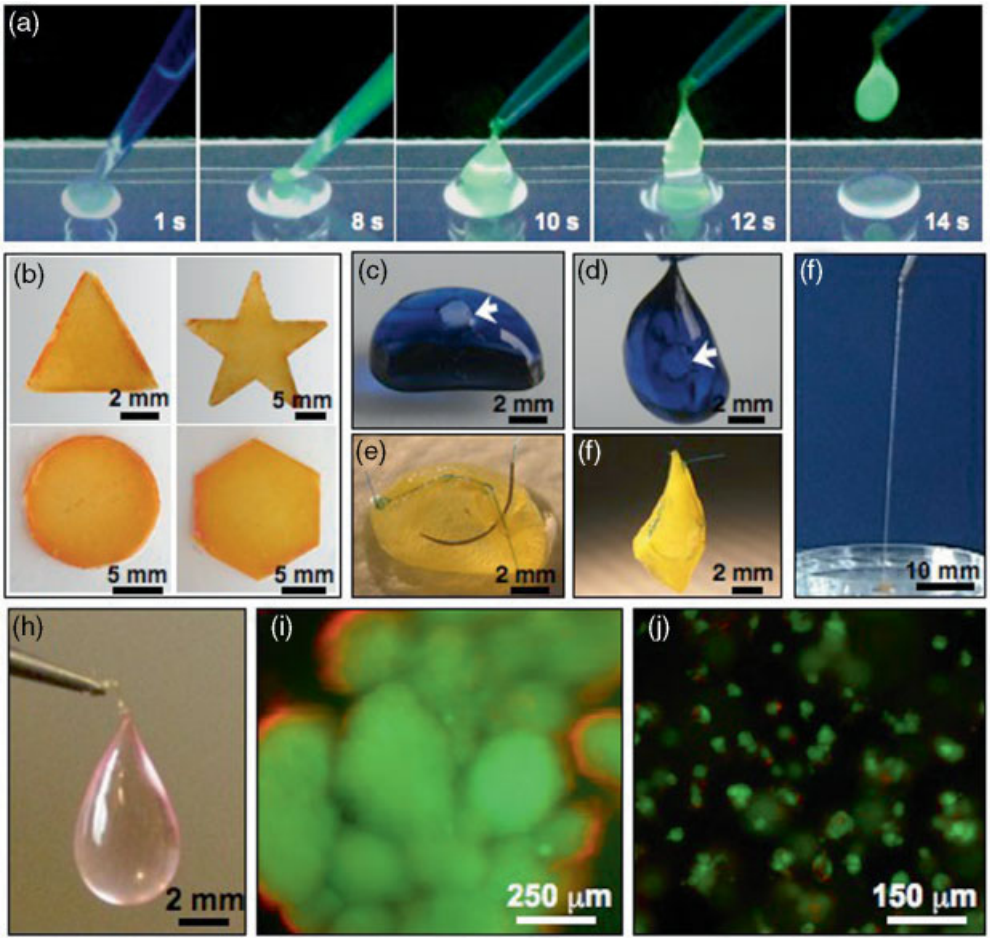


Figure 14.11 (a) Time-lapse photography of sac formation. A sample of a charged biopolymer solution is injected into an oppositely charged peptide amphiphile (PA) solution. The self-assembled sac is formed instantly; (b) PA-polymer membranes of different shapes created by interfacing the large- and small-molecule solutions in a very shallow template (~1 mm thick); (c) Hierarchically ordered sac formed with polydiacetylene PA containing a macroscopic defect within the membrane (arrow); (d) Sac in (c) after the defect is repaired and the sac resealed by triggering additional self-assembly with a drop of PA (arrow). Sacs are robust enough to withstand suturing (e) and can hold their weight without further tearing of the membrane (f); (g) Continuous string pulled from the interface between the PA and polymer solutions; (h) A sac encapsulating cells (sac is a pink color from cell media). Live/dead assay of (i) mouse pancreatic islets and (j) human mesenchymal stem cells (hMSCs) cultured within the sacs (green cells are live, red cells are dead). The islets remained viable up to a week and the hMSCs up to a month in sac culture. (Reproduced with permission from Ref. [68]; © 2008, American Association for the Advancement of Science.)

14.3
Self-Assembling Systems for Surface Modification

Implantable materials are the essence of today's regenerative medicine. The ability to control these materials at the nanoscale has moved them from simple inert materials to biocompatible and bioactive materials [69]. The surface of a biomaterial is particularly important in regenerative medicine as it is the first point of contact with the body. Whether it is presenting a biomimetic atmosphere, disguising a foreign body, or activating specific biological processes, the surface of an implant plays a crucial role and can determine its success or failure. One key advantage of self-assembly is the possibility to modify and tailor surfaces to elicit a specific biological response. In the following, we discuss the use of self-assembly to modify the properties of surfaces and 3-D structures.

14.3.1
Coatings on Surfaces

Within the scope of regenerative medicine, the molecular self-assembly of peptides represents a promising tool to modify the surfaces of medical implants or regenerative scaffolds. This technique facilitates the presentation of bioactive surface chemistries in a controlled, ordered fashion to mimic those of natural extracellular matrices. While the bioactivity of surfaces can be highly modulated by the presentation of specific ECM proteins, the effectiveness of this approach depends greatly on the appropriate conformation of the protein to expose the bioactive epitopes. Self-assembling materials offer the possibility to incorporate small peptide sequences as part of the self-assembling molecule. This approach avoids complications associated with intact proteins, such as undesirable protein folding and immune reactions, and also increases the specificity and efficiency of the bioactive epitope [70]. The use of small peptide sequences such as RGDS for cell adhesion [61] or IKVAV for neuronal differentiation [6], in combination with the capacity to self-assemble molecules in unique and specific conformations, makes this a powerful tool to modify and functionalize surfaces of materials used in regenerative medicine.

Self-assembled monolayers (SAMs) are single layers of molecules that react with and spontaneously order on solid surfaces. SAMs of alkanethiols on gold have been used extensively to study peptide and protein adsorption on surfaces [71, 72], as well as their effect on cell behavior [73, 74]. Recently, the modification of traditional SAM techniques has increased the level of surface chemistry manipulation and complexity that can be achieved. For example, the introduction of soft lithographic techniques such as microcontact printing has facilitated the use and significantly increased the potential of peptide-containing SAMs [70, 75]. This approach has been used to create self-assembled surface patterns to control and study a variety of cell behaviors such as cell adhesion, growth and apoptosis (Figure 14.12a) [76, 77]. Another approach that takes advantage of SAMs, and has been used in combination with soft lithographic techniques, consists of developing dynamically controlled and regulated surfaces,

which offer a unique opportunity to recreate and study dynamic biological processes. These types of surface can be achieved by controlling SAMs through different switching mechanisms (i.e. electrical, electrochemical, photochemical, thermal, and mechanical transduction [78, 79]) that organize specific ligands and peptides. By using these techniques, SAMs of peptides such as EG3- and RGD-terminated peptides have been used to study dynamic mechanisms controlling the adhesion and migration of bovine capillary endothelial cells [80] and fibroblasts [81], respectively (Figure 14.12b). Another modification of traditional SAM patterning takes advantage of dip-pen nanolithography (DPN), which uses atomic force microscopy (AFM) tips dipped into alkanethiol inks to transfer molecules by capillary force on

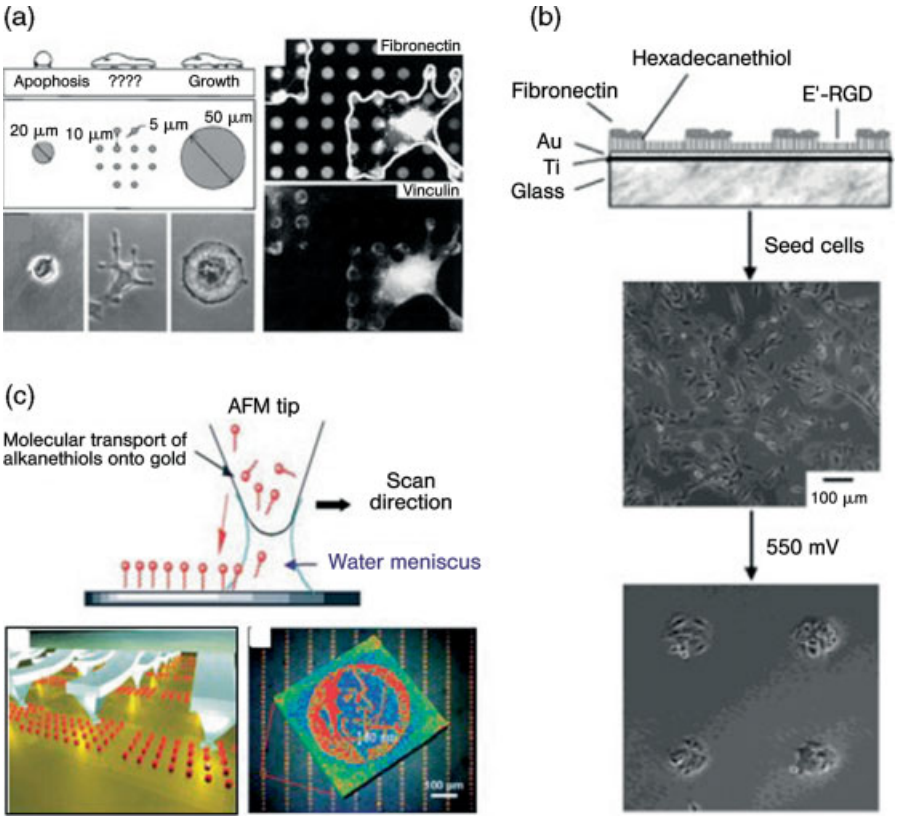


Figure 14.12 Approaches to create complex self-assembled monolayers (SAMs) including: (a) Micro-contact printing to create adhesive patterns of SAMs to study cell mechanisms such as growth and apoptosis. (Reproduced with permission from Ref. [76]; © 1997, American Association for the Advancement of Science.); (b) Patterns of SAMs that can be electroactively controlled and regulated to study cell adhesion and migration. (Reproduced with permission from Ref. [81]; © 2003, The American Chemical Society.); (c) Schematic of SAMs generated through dip-pen nanolithography (DPN). (Reproduced with permission from Ref. [82]; © 2007, Wiley-VCH Verlag GmbH & Co. KGaA.)

the gold surface (Figure 14.12c) [82, 83]. A major advantage of this technique is that it can create patterns of SAMs down to 15 nm in lateral dimension, significantly surpassing the resolution of soft lithographic techniques [82]. These types of studies not only provide reproducible tools to engineer biomimetic cell microenvironments, but also offer great promise for a deeper understanding of cell behaviors that could then be applied to the design of materials and implants in regenerative medicine [69].

While SAMs rely on individual molecules or peptides to create single-layer coatings, more complex self-assembled structures such as tubes or fibers are also being used as surface modifiers. One such example is a class of organic self-assembled fibers, referred to as helical rosette nanotubes (HRNs), that have been used to coat and functionalize bone prosthetic biomaterials (Figure 14.13). This approach was recently used to coat titanium surfaces, and caused a significant enhancement of osteoblast adhesion *in vitro* [84]. These molecules self-assemble from a single bicyclic block resulting from the complementary hydrogen-bonding arrays of both guanine (G) and cytosine (C). This C/G motif serves as the building block that self-assembles in water to form a six-membered supermacrocycle (rosette) maintained by 18 H-bonds. Subsequent assembly of these rosettes forms hollow nanotubes that are 1.1 nm in diameter and up to millimeters in length [85]. The outer surface of the G/C motif could further be modified to present specific physical and chemical properties.

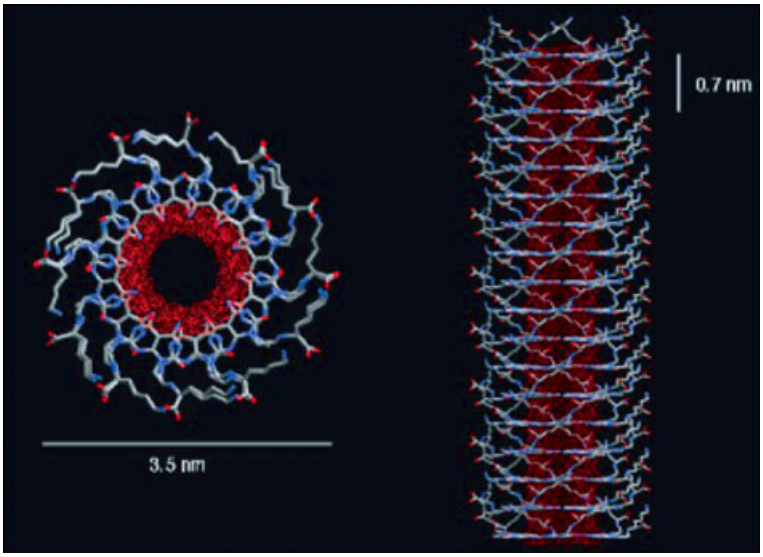


Figure 14.13 Illustration depicting the molecular structure of helical rosette nanotubes (HRNs) used for coating titanium surfaces with potential use in functionalizing the surface of bone prosthetic biomaterials. (Reproduced with permission from Ref. [84]; © 2005, Elsevier Limited.)

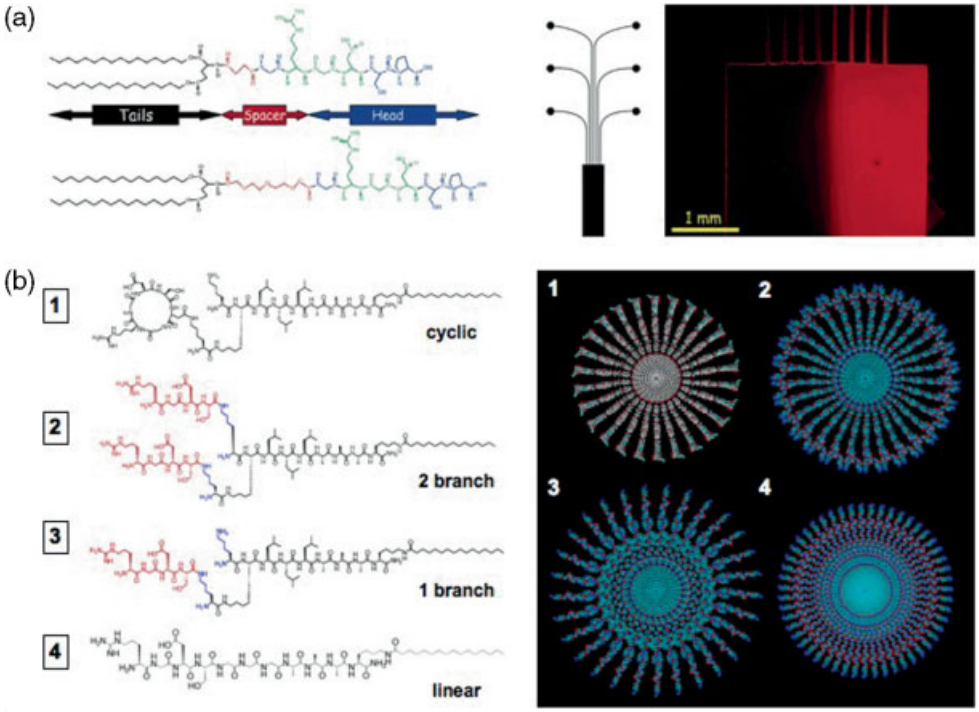


Figure 14.14 RGD-containing peptide amphiphile (PA) molecules used to control and modulate surface cell adhesion. (a) Surface-patterning techniques using microfluidic devices. (Reproduced with permission from Ref. [87]; © 2007, American Chemical Society.); (b) Optimum epitope presentation through specific molecular architectures. (Reproduced with permission from Ref. [61]; © 2007, Elsevier Limited.)

The self-assembly of PAs has also been used to functionalize two-dimensional (2-D) surfaces. As mentioned previously PAs are highly bioactive and biocompatible materials that have been used to develop 3-D scaffolds for tissue engineering and regenerative medicine. The diversity of design and robustness of these molecules has permitted a wide range of approaches for their use as surface-functionalizing coatings. For example, PAs containing cell-adhesive and triple-helical or α-helical structural motifs have been used to influence adhesion and signal transduction of human melanoma cells *in vitro* [86]. By utilizing microfluidic systems, Stroumpoulis *et al.* generated self-assembling patterns of RGD-containing PAs that directed mouse fibroblast adhesion (Figure 14.14a) [87]. In an attempt to optimize the presentation of PA coatings, Storrie *et al.* investigated the effect of PA molecular architecture and epitope concentration on nanofiber self-assembly, epitope presentation and fibroblast recognition for cell adhesion and spreading on surfaces (Figure 14.14b) [61].

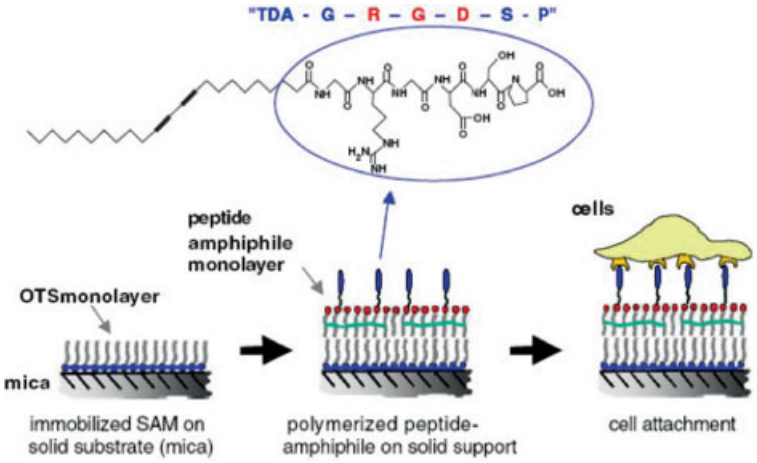


Figure 14.15 Peptide amphiphile (PA) molecules and self-assembling mechanism used to functionalize surfaces with improved molecular properties. A diacetylene-photosensitive segment in the hydrophobic tail promotes PA polymerization and subsequent monolayer stability. (Reproduced with permission from Ref. [89]; © 2006, Elsevier Limited.)

A number of strategies have been investigated to improve the chemical stability of PA coatings on implant surfaces. For example, Sargeant *et al.* has developed a method to covalently attach PA nanofibers to the surface of nickel–titanium (NiTi) shape memory alloys [88]. Here, the group used an RGDS containing PA and demonstrated its capacity to form robust PA coatings capable of promoting cell adhesion, proliferation and differentiation. This method significantly improves the potential of using PA materials for *in vivo* applications such as vascular stents, bone plates and artificial joints. Another approach used to improve PA stability was reported by Biesalski *et al.*, who developed a PA molecule comprising a diacetylene photosensitive segment, which promotes PA polymerization (Figure 14.15) [89]. This molecule was used to develop a stable polymerized monolayer of RGD-terminated PA molecules that significantly enhanced fibroblast adhesion.

The vast majority of investigations related to self-assembling peptides for surface modification has been dedicated to developing functional and bioactive surface chemistries to affect or elicit specific cell behaviors. However, in addition to surface chemistry, surface topography has also been shown to significantly affect cell and tissue behavior [90–92]. An ideal surface modification treatment for regenerative medicine would permit the fine-tuning of both surface chemistry and surface topography across different size scales. One approach to achieve this topographical/biochemical integration is to create SAMs of peptides on microfabricated surfaces comprising topographical features [93, 94]. The integration of microfabrication with molecularly designed self-assembling PAs also represents a unique opportunity to develop physical and biochemical environments with hierarchical

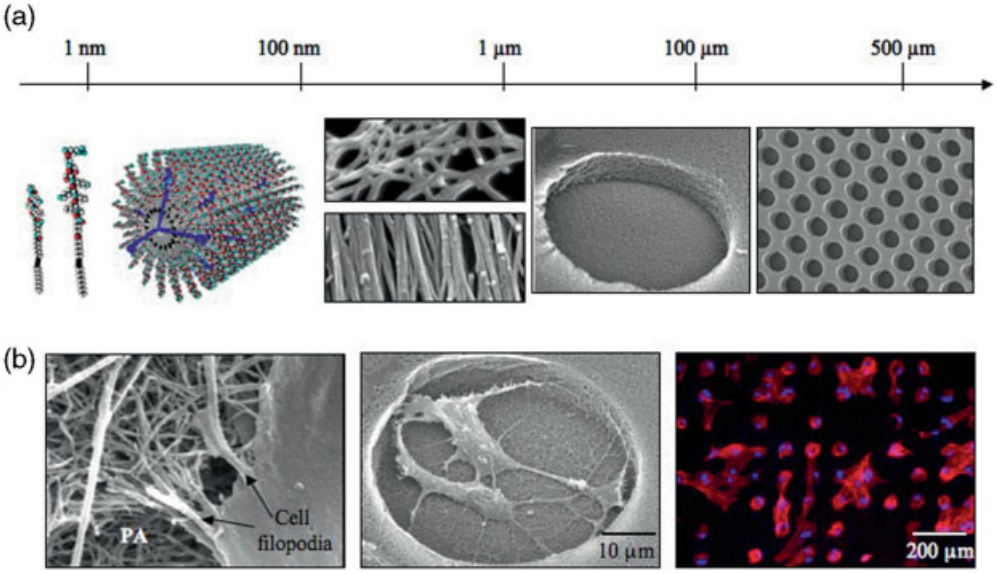


Figure 14.16 (a) Fabrication approach that combines bottom-up (self-assembling peptide-amphiphiles) with top-down (microfabrication) techniques to create biomimetic environments for stem cell manipulation. This method integrates precise topographical patterns and specific biochemical signals within a hierarchical structure that expands from the molecular to the macro scale; (b) Topographical patterns made from self-assembled PA nanofibers have demonstrated the capacity to guide the growth and differentiation of human mesenchymal stem cells (red = actin cytoskeleton, blue = cell nuclei). (with permission from *Soft Matter*, DOI:10.1039/b819002j).

organization (Figure 14.16a). With this approach, it may be possible to create biomimetic scaffolds made from PA molecules with the capacity to elicit specific cell behaviors using both topographical features and bioactive epitopes at different size scales (Figure 14.16b). Recent studies performed by Stupp *et al.* have demonstrated the capacity to promote osteoblastic differentiation of human MSCs by using topographical patterns made from self-assembled PA nanofibers. (*Soft Matter*, DOI:10.1039/b819002j).

14.3.2
Coatings on 3-D Scaffolds

Three-dimensional scaffolds prepared with synthetic materials such as poly(glygolic acid) (PGA) and poly (L-lactic acid) (PLLA) provide porous and biodegradable materials that have found extensive use in regenerative medicine applications [95]. The surface characteristics of these materials, however, do not have any specific or desired bioactivity. Therefore, self-assembling peptides may be used to functionalize surfaces to further improve bioactivity and tissue integration. For example, Harrington *et al.* used an RGDS-containing PA to self-assemble into well-defined

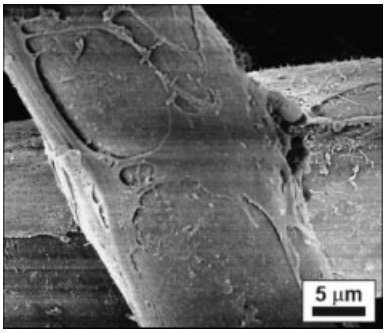


Figure 14.17 Osteoblasts growing on a fibrous poly(L-lactic acid) (PLLA) scaffold coated with molecules comprising cholesterol and lysine moieties. (Reproduced with permission from Ref. [8]; © 2004, Elsevier Limited.)

nanofibers on the surface of PGA porous scaffolds [95]. These RGDS-coated scaffolds significantly improved human bladder smooth muscle cell adhesion. Stendahl *et al.* self-assembled a triblock molecule comprising cholesterol and lysine moieties to coat and modify PLLA fiber scaffolds [8]. These amphiphilic molecules improved the adhesion and overall growth of osteoblastic cells (Figure 14.17). Another examples of a recent surface modification technique used within 3-D architectures includes studies conducted by Zhu *et al.*, who used poly(ethylenimine) (PEI) to activate the surface of poly(lactide) (PLA) scaffolds, which was subsequently modified with gelatin using electrostatic self-assembly [96, 97]. This treatment successfully promoted the growth of seeded osteoblasts.

14.4
Clinical Potential of Self-Assembling Systems

As discussed above, several preclinical studies have already shown great promise for the use of self-assembling biomaterials in regenerative medicine. Particularly, *in vivo* experiments using self-assembling peptide amphiphiles by Stupp and colleagues have shown that these bioactive matrices can be specifically designed to: (i) promote angiogenesis (rat cornea model); (ii) promote regeneration of axons in a spinal cord injury model (mouse and rat models), of cartilage (rabbit model), and of bone (rat model); (iii) promote recovery of cardiac function after infarct (mouse model); and (iv) show promise as treatments for Parkinson's disease (mouse models). It is expected that the self-assembly of supramolecular systems will, in time, lead to many effective regenerative medicine therapies providing an excellent platform to design for bioactivity, harmless degradation with appropriate half-lives after providing a function, and noninvasive methods for clinical delivery:

- Design for bioactivity: it is possible to engineer these peptide-based, self-assembling systems to include various combinations of amino acid sequences

that are bioactive and can enhance the regeneration process – that is, deliver growth factors, contain cell adhesion sequences, mimic the bioactivity of growth factors, and so on.

- Harmless degradation: peptide-based, self-assembling systems are capable of being degraded by enzymes in the body into basic amino acids that can be metabolized naturally, with appropriate half-lives after providing a function. Their degradation characteristics can be manipulated through molecular design. Ideally, such bioactive materials would provide a specific function (i.e. deliver growth factors, attract and adhere desired cell types, etc.) to enhance the regenerative process, and simultaneously degrade as the tissue starts to regenerate. Over time, what is left would be the completely regenerated tissue. The challenge would be to determine the 'appropriate' degradation rate for optimal regeneration.
- Noninvasive methods for clinical delivery: the ability of these peptide-based molecules to self-assemble spontaneously allows for the administration of materials through noninvasive methods. For example, a solution of the self-assembling molecules could be injected into the defect site, after which gelation in vivo could be triggered by ions within the body.

14.5 Conclusions

Today, research into the development of self-assembling biomaterials for regenerative medicine continues to expand–the main aim being to achieve real improvements in the quality of life for mankind. Without strategies for regeneration, genomic data and personalized medicine will not have the significant impact that is being promised. It is important that therapies for regenerative medicine must be not only highly effective and predictable, but also as noninvasive as possible, with the capacity to reach deep into problem areas of the heart, brain, skeleton, skin and other vital organs. It is for this reason that self-assembly at the nanoscale appears as the most sensible technological strategy, to signal and recruit the organism's own cells, or to manage the delivery of cell therapies to the correct sites after effective *in vitro* manipulation. The ability to design at both the nanoscale and macroscale will open the door to vast possibilities for biomaterials and regenerative medicine, with materials that can be designed to multiplex the required signals, can be delivered in a practical and optimal manner, and can reach targets across barriers via the blood circulation. In addition, molecular self-assembly on the surfaces of implants may enhance the bioactivity and predictable biocompatibility of metals, ceramics, composites and synthetic polymers. Self-assembly is at the root of structure versus function in biology and, in the context of regenerative medicine technology, is the 'ultimate inspiration from Nature'.

References

1 Sargeant, T.D., Guler, M.O., Oppenheimer, S.M., Mata, A., Satcher, R.L., Dunand, D.C. and Stupp, S.I. (2008) *Biomaterials*, **29**, 161.

2 Hwang, J.J., Iyer, S.N., Li, L.S., Claussen, R., Harrington, D.A. and Stupp, S.I. (2002) *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 9662.

3 Klok, H.A., Hwang, J.J., Iyer, S.N. and Stupp, S.I. (2002) *Macromolecules*, **35**, 746.

4 Hartgerink, J.D., Beniash, E. and Stupp, S.I. (2001) *Science*, **294**, 1684.

5 Hartgerink, J.D., Beniash, E. and Stupp, S.I. (2002) *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 5133.

6 Silva, G.A., Czeisler, C., Niece, K.L., Beniash, E., Harrington, D.A., Kessler, J.A. and Stupp, S.I. (2004) *Science*, **303**, 1352.

7 Stupp, S.I. (2005) *MRS Bulletin*, **30**, 546.

8 Stendahl, J.C., Li, L., Claussen, R.C. and Stupp, S.I. (2004) *Biomaterials*, **25**, 5847.

9 Harrington, D.A., Cheng, E.Y., Guler, M.O., Lee, L.K., Donovan, J.L., Claussen, R.C. and Stupp, S.I. (2006) *Journal of Biomedical Materials Research Part A*, **78**, 157.

10 Gazit, E. (2007) *Chemical Society Reviews*, **36**, 1263.

11 Aggeli, A., Bell, M., Boden, N., Keen, J.N., Knowles, P.F., McLeish, T.C., Pitkeathly, M. and Radford, S.E. (1997) *Nature*, **386**, 259.

12 Aggeli, A., Bell, M., Carrick, L.M., Fishwick, C.W., Harding, R., Mawer, P.J., Radford, S.E., Strong, A.E. and Boden, N. (2003) *Journal of the American Chemical Society*, **125**, 9619.

13 Kirkham, J., Firth, A., Vernals, D., Boden, N., Robinson, C., Shore, R.C., Brookes, S.J. and Aggeli, A. (2007) *Journal of Dental Research*, **86**, 426.

14 Aggeli, A., Bell, M., Boden, N., Carrick, L.M. and Strong, A.E. (2003) *Angewandte Chemie - International Edition in English*, **42**, 5603.

15 Bell, C.J., Carrick, L.M., Katta, J., Jin, Z., Ingham, E., Aggeli, A., Boden, N., Waigh, T.A. and Fisher, J. (2006) *Journal of Biomedical Materials Research Part A*, **78**, 236.

16 Schneider, J.P., Pochan, D.J., Ozbas, B., Rajagopal, K., Pakstis, L. and Kretsinger, J. (2002) *Journal of the American Chemical Society*, **124**, 15030.

17 Pochan, D.J., Schneider, J.P., Kretsinger, J., Ozbas, B., Rajagopal, K. and Haines, L. (2003) *Journal of the American Chemical Society*, **125**, 11802.

18 Ozbas, B., Rajagopal, K., Schneider, J.P. and Pochan, D.J. (2004) *Physical Review Letters*, **93**, 268106.

19 Ozbas, B., Rajagopal, K., Haines-Butterick, L., Schneider, J.P. and Pochan, D.J. (2007) *The Journal of Physical Chemistry B*, **111**, 13901.

20 Haines, L.A., Rajagopal, K., Ozbas, B., Salick, D.A., Pochan, D.J. and Schneider, J.P. (2005) *Journal of the American Chemical Society*, **127**, 17025.

21 Haines-Butterick, L., Rajagopal, K., Branco, M., Salick, D., Rughani, R., Pilarz, M., Lamm, M.S., Pochan, D.J. and Schneider, J.P. (2007) *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 7791.

22 Kretsinger, J.K., Haines, L.A., Ozbas, B., Pochan, D.J. and Schneider, J.P. (2005) *Biomaterials*, **26**, 5177.

23 Salick, D.A., Kretsinger, J.K., Pochan, D.J. and Schneider, J.P. (2007) *Journal of the American Chemical Society*, **129**, 14793.

24 Nowak, A.P., Breedveld, V., Pakstis, L., Ozbas, B., Pine, D.J., Pochan, D. and Deming, T.J. (2002) *Nature*, **417**, 424.

25 Pochan, D.J., Pakstis, L., Ozbas, B., Nowak, A.P. and Deming, T.J. (2002) *Macromolecules*, **35**, 5358.

26 Breedveld, V., Nowak, A.P., Sato, J., Deming, T.J. and Pine, D.J. (2004) *Macromolecules*, **37**, 3943.

27 Pakstis, L.M., Ozbas, B., Hales, K.D., Nowak, A.P., Deming, T.J. and Pochan, D. (2004) *Biomacromolecules*, **5**, 312.

28 Zhang, S., Holmes, T., Lockshin, C. and Rich, A. (1993) *Proceedings of the National Academy of Sciences of the United States of America*, **90**, 3334.

29 Zhang, S., Holmes, T.C., DiPersio, C.M., Hynes, R.O., Su, X. and Rich, A. (1995) *Biomaterials*, **16**, 1385.

30 Holmes, T.C., S., de Lacalle Su, X., Liu, G., Rich, A. and Zhang, S. (2000) *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 6728.

31 Caplan, M.R., Schwartzfarb, E.M., Zhang, S., Kamm, R.D. and Lauffenburger, D.A. (2002) *Journal of Biomaterials Science, Polymer Edition*, **13**, 225.

32 Caplan, M.R., Schwartzfarb, E.M., Zhang, S., Kamm, R.D. and Lauffenburger, D.A. (2002) *Biomaterials*, **23**, 219.

33 Ellis-Behnke, R.G., Liang, Y.X., You, S.W., Tay, D.K., Zhang, S., So, K.F. and Schneider, G.E. (2006) *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 5054.

34 Guo, J., Su, H., Zeng, Y., Liang, Y.X., Wong, W.M., Ellis-Behnke, R.G., So, K.F. and Wu, W. (2007) *Nanomedicine*, **3**, 311.

35 Semino, C.E., Kasahara, J., Hayashi, Y. and Zhang, S. (2004) *Tissue Engineering*, **10**, 643.

36 Genove, E., Shen, C., Zhang, S. and Semino, C.E. (2005) *Biomaterials*, **26**, 3341.

37 Narmoneva, D.A., Oni, O., Sieminski, A.L., Zhang, S., Gertler, J.P., Kamm, R.D. and Lee, R.T. (2005) *Biomaterials*, **26**, 4837.

38 Sieminski, A.L., Was, A.S., Kim, G., Gong, H. and Kamm, R.D. (2007) *Cell Biochemistry and Biophysics*, **49**, 73.

39 Davis, M.E., Hsieh, P.C., Takahashi, T., Song, Q., Zhang, S., Kamm, R.D., Grodzinsky, A.J., Anversa, P. and Lee, R.T. (2006) *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 8155.

40 Davis, M.E., Motion, J.P., Narmoneva, D.A., Takahashi, T., Hakuno, D., Kamm, R.D., Zhang, S. and Lee, R.T. (2005) *Circulation*, **111**, 442.

41 Hsieh, P.C., Davis, M.E., Gannon, J., MacGillivray, C. and Lee, R.T. (2006) *The Journal of Clinical Investigation*, **116**, 237.

42 Hsieh, P.C., MacGillivray, C., Gannon, J., Cruz, F.U. and Lee, R.T. (2006) *Circulation*, **114**, 637.

43 Semino, C.E., Merok, J.R., Crane, G.G., Panagiotakos, G. and Zhang, S. (2003) *Differentiation; Research in Biological Diversity*, **71**, 262.

44 Kisiday, J., Jin, M., Kurz, B., Hung, H., Semino, C., Zhang, S. and Grodzinsky, A.J. (2002) *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 9996.

45 Garreta, E., Gasset, D., Semino, C. and Borros, S. (2007) *Biomolecular Engineering*, **24**, 75.

46 Garreta, E., Genove, E., Borros, S. and Semino, C.E. (2006) *Tissue Engineering*, **12**, 2215.

47 Guo, J., Su, H., Zeng, Y., Liang, Y.X., Wong, W.M., Ellis-Behnke, R.G., So, K.F. and Wu, W. (2007) *Nanomedicine*, **3**, 311.

48 Jayawarna, V., Smith, A., Gough, J.E. and Ulijn, R.V. (2007) *Biochemical Society Transactions*, **35**, 535.

49 Toledano, S., Williams, R.J., Jayawarna, V. and Ulijn, R.V. (2006) *Journal of the American Chemical Society*, **128**, 1070.

50 Lowik, D.W. and van Hest, J.C. (2004) *Chemical Society Reviews*, **33**, 234.

51 Behanna, H.A., Donners, J.J., Gordon, A.C. and Stupp, S.I. (2005) *Journal of the American Chemical Society*, **127**, 1193.

52 Bull, S.R., Guler, M.O., Bras, R.E., Meade, T.J. and Stupp, S.I. (2005) *Nano Letters*, **5**, 1.

53 Claussen, R.C., Rabatic, B.M. and Stupp, S.I. (2003) *Journal of the American Chemical Society*, **125**, 12680.

54 Guler, M.O., Hsu, L., Soukasene, S., Harrington, D.A., Hulvat, J.F. and Stupp, S.I. (2006) *Biomacromolecules*, **7**, 1855.

55 Guler, M.O., Soukasene, S., Hulvat, J.F. and Stupp, S.I. (2005) *Nano Letters*, **5**, 249.

56 Niece, K.L., Hartgerink, J.D., Donners, J.J. and Stupp, S.I. (2003) *Journal of the American Chemical Society*, **125**, 7146.

57 Rajangam, K., Behanna, H.A., Hui, M.J., Han, X., Hulvat, J.F., Lomasney, J.W. and Stupp, S.I. (2006) *Nano Letters*, **6**, 2086.

58 Sone, E.D. and Stupp, S.I. (2004) *Journal of the American Chemical Society*, **126**, 12756.

59 Stendahl, J.C., Rao, M.S., Guler, M.O. and Stupp, S.I. (2006) *Advanced Functional Materials*, **16**, 499.

60 Beniash, E., Hartgerink, J.D., Storrie, H., Stendahl, J.C. and Stupp, S.I. (2005) *Acta Biomaterialia*, **1**, 387.

61 Storrie, H., Guler, M.O., Abu-Amara, S.N., Volberg, T., Rao, M., Geiger, B. and Stupp, S.I. (2007) *Biomaterials*, **28**, 4608.

62 Tysseling-Mattiace, V.M., Sahni, V., Niece, K.L., Birch, D., Czeisler, C., Fehlings, M.G., Stupp, S.I. and Kessler, J.A. (2008) *The Journal of Neuroscience*, **28**, 3814.

63 Kapadia, M.R., Chow, L.W., Tsihlis, N.D., Ahanchi, S.S., Eng, J.W., Murar, J., Martinez, J., Popowich, D.A., Jiang, Q., Hrabie, J.A., Saavedra, J.E., Keefer, L.K., Hulvat, J.F., Stupp, S.I. and Kibbe, M.R. 2008 *Journal of Vascular Surgery*, **47**, 173.

64 Jiang, H., Guler, M.O. and Stupp, S.I. (2007) *Soft Matter*, **3**, 454.

65 Palmer, L.C., Velichko, Y.S., Olvera De La Cruz, M. and Stupp, S.I. (2007) *Philosophical Transactions of the Royal Society of London. Series A: Mathematical and Physical Sciences*, **365**, 1417.

66 Jun, H., Yuwono, V., Paramonov, S.E. and Hartgerink, J.D. (2005) *Advanced Materials*, **17**, 2612.

67 Paramonov, S.E., Jun, H.W. and Hartgerink, J.D. (2006) *Biomacromolecules*, **7**, 24.

68 Capito, R.M., Azevedo, H.S., Velichko, Y.S., Mata, A. and Stupp, S.I. (2008) *Science*, **319**, 1812.

69 Stupp, S.I., Donners, J.J.J.M., Li, L.S. and Mata, A. (2005) *MRS Bulletin*, **30**, 864.

70 Geim, A.K., Dubonos, S.V., Grigorieva, I.V., Novoselov, K.S., Zhukov, A.A. and Shapoval, S.Y. (2003) *Nature Materials*, **2**, 461.

71 Ruiz, S.A. and Chen, C.S. (2007) *Soft Matter*, **3**, 168.

72 Sniadecki, N.J., Tan, J., Anguelouch, A., Ruiz, S.A., Reich, D.H. and Chen, C.S. (2004) *Molecular Biology of the Cell*, **15**, 54.

73 Roberts, C., Chen, C.S., Mrksich, M., Martichonok, V., Ingber, D.E. and Whitesides, G.M. (1998) *Journal of the American Chemical Society*, **120**, 6548.

74 Houseman, B.T. and Mrksich, M. (2001) *Biomaterials*, **22**, 943.

75 Mrksich, M. and Whitesides, G.M. (1995) *Trends in Biotechnology*, **13**, 228.

76 Chen, C.S., Mrksich, M., Huang, S., Whitesides, G.M. and Ingber, D.E. (1997) *Science*, **276**, 1425.

77 Chen, C.S., Mrksich, M., Huang, S., Whitesides, G.M. and Ingber, D.E. (1998) *Biotechnology Progress*, **14**, 356.

78 Mrksich, M. (2005) *MRS Bulletin*, **30**, 180.

79 Lahann, J. and Langer, R. (2005) *MRS Bulletin*, **30**, 185.

80 Jiang, X.Y., Ferrigno, R., Mrksich, M. and Whitesides, G.M. (2003) *Journal of the American Chemical Society*, **125**, 2366.

81 Yeo, W.S., Yousaf, M.N. and Mrksich, M. (2003) *Journal of the American Chemical Society*, **125**, 14994.

82 Huck, W.T.S. (2007) *Angewandte Chemie - International Edition*, **46**, 2754.

83 Piner, R.D., Zhu, J., Xu, F., Hong, S.H. and Mirkin, C.A. (1999) *Science*, **283**, 661.

84 Chun, A.L., Moralez, J.G., Webster, T.J. and Fenniri, H. (2005) *Biomaterials*, **26**, 7304.

85 Chun, A.L., Moralez, J.G., Fenniri, H. and Webster, T.J. (2004) *Nanotechnology*, **15**, S234.

86 Fields, G.B., Lauer, J.L., Dori, Y., Forns, P., Yu, Y.C. and Tirrell, M. (1998) *Biopolymers*, **47**, 143.

87 Stroumpoulis, D., Zhang, H.N., Rubalcava, L., Gliem, J. and Tirrell, M. (2007) *Langmuir*, **23**, 3849.

88 Sargeant, T.D., Rao, M.S., Koh, C.Y. and Stupp, S.I. (2008) *Biomaterials*, **29**, 1085.

89 Biesalski, M.A., Knaebel, A., Tu, R. and Tirrell, M. (2006) *Biomaterials*, **27**, 1259.

90 Tirrell, M., Kokkoli, E. and Biesalski, M. (2002) *Surface Science*, **500**, 61.

91 Charest, J.L., Eliason, M.T., Garcia, A.J. and King, W.P. (2006) *Biomaterials*, **27**, 2487.

92 Kunzler, T.P., Huwiler, C., Drobek, T., Voros, J. and Spencer, N.D. (2007) *Biomaterials*, **28**, 5000.

93 Mrksich, M., Chen, C.S., Xia, Y.N., Dike, L.E., Ingber, D.E. and Whitesides, G.M. (1996) *Proceedings of the National Academy of Sciences of the United States of America*, **93**, 10775.

94 Lussi, J.W., Michel, R., Reviakine, I., Falconnet, D., Goessl, A., Csucs, G., Hubbell, J.A. and Textor, M. (2004) *Progress in Surface Science*, **76**, 55.

95 Harrington, D.A., Cheng, E.Y., Guler, M.O., Lee, L.K., Donovan, J.L., Claussen, R.C. and Stupp, S.I. (2006) *Journal of Biomedical Materials Research Part A*, **78**, 157.

96 Zhu, H., Ji, J. and Shen, J. (2004) *Biomacromolecules*, **5**, 1933.

97 Zhu, H., Ji, J., Barbosa, M.A. and Shen, J. (2004) *Journal of Biomedical Materials Research Applied Biomaterials*, **71**, 159.

1
Spin-Polarized Scanning Tunneling Microscopy

Mathias Getzlaff

1.1
Introduction and Historical Background

Until the 1980s an idealized and rather unrealistic view was found in surface physics for a lack of techniques which allowed real-space imaging. During this time, surfaces were often assumed to be perfect – that is, imperfections such as step edges, dislocations or adsorbed atoms were neglected. Most of the important information was gained rather indirectly by spatially averaging methods or experimental techniques with insufficient resolution.

However, in 1982, with the invention of the scanning tunneling microscope by G. Binnig and H. Rohrer [1, 2], the situation changed dramatically. This instrument allowed, for the first time, the topography of surfaces to be imaged in real space with both lateral and vertical atomic resolution.

Subsequently, a number of different spectroscopic modes were introduced which provided additional access to electronic behavior, thus allowing the correlation of topographic and electronic properties down to the atomic scale.

In 1988, Pierce considered the possibility of making the scanning tunneling microscope sensitive towards the spin of tunneling electrons by using spin-sensitive tip materials as a further development [3], and this was also predicted – theoretically – by Minakov *et al.* [4]. As a step towards this aim, Allenspach *et al.* [5] replaced the electron gun of a scanning electron microscope with a scanning tunneling microscope tip. Thus, in field emission mode the electrons impinged on a magnetic surface, and the spin polarization of the emitted electrons was subsequently monitored; this, at least in principle, would allow magnetic imaging with nanometer resolution.

However, it was the first 'direct' realization by Wiesendanger *et al.* [6] that opened the possibility of imaging the magnetic properties at atomic resolution. Moreover, the importance of this proposal was not restricted only to basic studies but was also applicable to research investigations. Meanwhile, a rapidly increasing interest emerged from an industrial point of view, a concept which became even more

Nanotechnology. Volume 6: Nanoprobes. Edited by Harald Fuchs

ISBN: 978-3-527-31733-2

important when considering the need for dramatic increases in the storage density of devices such as computer hard drives. Clearly, further developments in this area will require tools that allow high spatial resolution magnetic imaging for an improved understanding of nanoscaled objects such as magnetic domains and domain wall structures.

In this chapter, we will describe the successful development and implementation of spin-polarized scanning tunneling microscopy (SP-STM), and will also show – by means of selected examples – how our understanding of surface magnetic behavior has vastly increased in recent years.

1.2
Spin-Polarized Electron Tunneling: Considerations Regarding Planar Junctions

First, let us assume that two ferromagnetically or antiferromagnetically coupling layers are separated by an insulator (Figure 1.1, left part). The following discussion can also be extended to ferromagnetic nanoparticles located within an insulating matrix (Figure 1.1, right part).

In order for an electric current to occur, there is the prerequisite that the thickness of the barrier should be small enough so as to allow quantum mechanical tunneling. It is also essential for this discussion that this process is assumed to conserve the spin orientation.

The dependence of the tunneling current on the relative magnetization is shown in Figure 1.2, assuming two ferromagnetic thin layers. The total resistivity for the parallel alignment is less than for the antiparallel orientation. This effect, which is known as tunneling magnetoresistance (TMR), represents a band structure effect that relies on the spin resolved density of states (DOS) at the Fermi level. In comparison, giant magnetoresistance (GMR) is caused by a spin-dependent scattering at the interfaces (further information is available in Ref. [7]).

In order to discuss the behavior of two ferromagnetic electrodes separated by an insulating barrier, the model of Jullière [8] is used; this employs the assumptions that the tunneling process is spin-conserving, and that the tunneling current is proportional to the density of states of the corresponding spin orientation in each electrode.

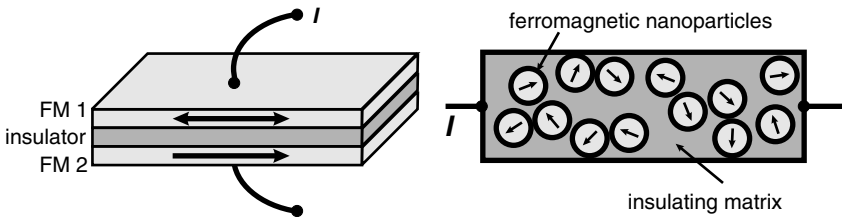


Figure 1.1 Tunneling magnetoresistance can occur when ferromagnetic thin films are separated by an insulating layer (left), and when ferromagnetic nanoparticles are embedded in an insulating matrix (right).

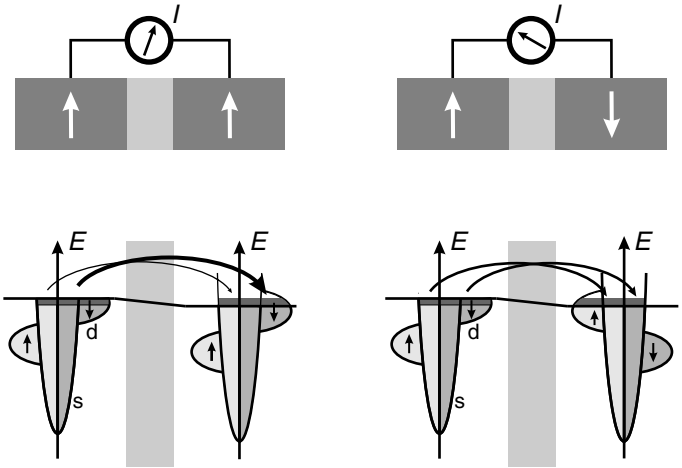


Figure 1.2 Dependence of the tunneling current on the relative magnetization of two ferromagnetic layers. For a parallel orientation a large quantity of spin down electrons at the Fermi energy can tunnel into empty down states; this results in a high tunneling current. In contrast, for an antiparallel orientation the quantity of empty down states is significantly lower, leading to a reduced tunneling current.

In this situation, the tunneling current for a parallel magnetization is given by:

$$I^{\uparrow\uparrow} \propto n_1^{\uparrow} n_2^{\uparrow} + n_1^{\downarrow} n_2^{\downarrow} \tag{1.1}$$

with n_i being the electron density of electrode i at the Fermi level E_F. For the antiparallel orientation, the tunneling current amounts to:

$$I^{\uparrow\downarrow} \propto n_1^{\uparrow} n_2^{\downarrow} + n_1^{\downarrow} n_2^{\uparrow} \tag{1.2}$$

With $a_i = n_i^{\uparrow}/(n_i^{\uparrow} + n_i^{\downarrow})$ being the part of majority electrons of electrode i, and $1-a_i = n_i^{\downarrow}/(n_i^{\uparrow} + n_i^{\downarrow})$ that of the minority electrons, the spin polarization P_i of electrode i is given by:

$$P_i = \frac{n_i^{\uparrow} - n_i^{\downarrow}}{n_i^{\uparrow} + n_i^{\downarrow}} = 2a_i - 1 \tag{1.3}$$

This allows the differential conductance for a parallel orientation to be expressed as:

$$G^{\uparrow\uparrow} = G_p \propto a_1 a_2 + (1-a_1)(1-a_2) = \frac{1}{2}(1 + P_1 P_2) \tag{1.4}$$

and for an antiparallel orientation as:

$$G^{\uparrow\downarrow} = G_{ap} \propto a_1(1-a_2) + (1-a_1)a_2 = \frac{1}{2}(1 - P_1 P_2) \tag{1.5}$$

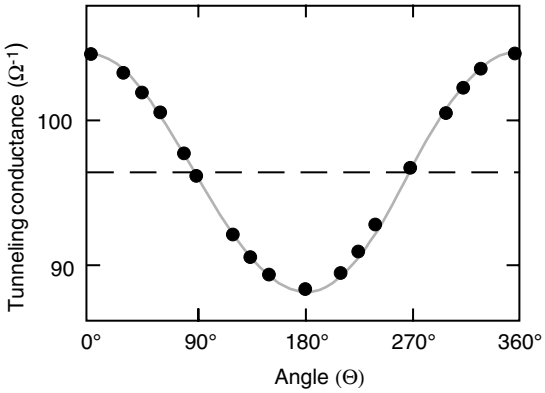


Figure 1.3 Dependence of the tunneling conductance (inverse resistance) of a planar Fe-Al_2O_3-Fe junction on the angle Θ between the magnetization vectors of both electrodes. (Data taken from Ref. [9]).

The magnitude of the TMR effect is given by:

$$\mathrm{TMR} = \frac{G^{\uparrow\uparrow} - G^{\uparrow\downarrow}}{G^{\uparrow\downarrow}} = \frac{G_\mathrm{p} - G_{ap}}{G_{ap}} = \frac{R_{ap} - R_\mathrm{p}}{R_\mathrm{p}} \tag{1.6}$$

where R_p (R_{ap}) is the resistance for the (anti)parallel orientation, respectively. By using the spin polarization, the TMR effect can be written as:

$$\mathrm{TMR} = \frac{\Delta R}{R_\mathrm{p}} = \frac{2P_1P_2}{1 - P_1P_2} \tag{1.7}$$

In the above considerations, it has been assumed that the magnetizations in both ferromagnetic electrodes are oriented parallel or antiparallel. However, the differential conductance also depends on the angle Θ between both directions of magnetization. This behavior is shown in Figure 1.3 for an Fe–Al_2O_3–Fe junction. Thus, until now the situation has been discussed for $\Theta = 0^\circ$ and $\Theta = 180^\circ$, respectively. For an arbitrary angle Θ, the differential conductance can be expressed as:

$$G = G_0 \cdot (1 + P_1P_2\cos\Theta) \tag{1.8}$$

with $G_0 = (G_\mathrm{p} + G_{ap})/2$ being the spin-averaged conductance.

1.3 Spin-Polarized Electron Tunneling in Scanning Tunneling Microscopy (STM): Experimental Aspects

The substitution of a ferromagnetic electrode (as discussed above) with a ferromagnetic probe tip represents the situation in SP-STM. The insulating barrier is realized by the vacuum between the sample and the tip, which are separated by a distance of several Ångstroms, thus allowing the laterally resolved determination of magnetic

properties. As a consequence, the zero bias anomaly – that is, the decrease in the TMR with increasing bias voltage in planar junctions – is not present for SP-STM investigations because the anomaly can be attributed to scattering of electrons at defects in amorphous barriers [10].

The following sections describe the two fundamental experimental aspects concerning spin-polarized electron tunneling in an STM experiment. Here, different probe tips and modes of operation are employed in order to obtain magnetic information from a sample.

1.3.1 Probe Tips for Spin-Polarized Electron Tunneling

In order to realize SP-STM, the probe tip should fulfill most of the following conditions:

- The higher the spin polarization of the apex atom, the more pronounced the magnetic information (cf. Equations 1.4–1.7) in comparison to electronic and topographic information. Due to the typical reduction of this spin polarization by adsorbates from the residual gas, even under ultra-high vacuum (UHV) conditions, a clean environment or an inert tip material is certainly advantageous [11].
- The sensitivity can also be improved by periodically reversing the magnetization direction, thus *directly* probing the local tunneling magnetoresistance.
- The interaction between tip and sample should be as low as possible because the stray field of a ferromagnetic tip may modify or destroy the sample's domain structure.
- Controlling the orientation of the magnetization axis of the tip parallel or perpendicular to the sample surface allows the domain structure of any sample to be imaged, independent of whether its easy axis is in-plane or out-of-plane.

1.3.1.1 Ferromagnetic Probe Tips

With regards to stray field minimization, bulk tips made from ferromagnetic 3*d* transition metals (see Refs. [12, 13]), with their high content of magnetic material and high saturation magnetization are an unfavorable choice. This mostly restricts their application to ferri- and antiferromagnetic samples [14], which are practically insensitive to external fields. Nevertheless, the stray field can be reduced either by using a material which exhibits a low saturation magnetization, or by using thin-film tips with a film thickness comparable to (or less than) the tip–sample separation [15]. The spin dependence of image potential states can also be used as a sensitive probe of surface magnetism [16], allowing high-resolution magnetic imaging at tip–sample distances larger than in normal tunneling experiments, and thereby reducing the stray field of the ferromagnetic tip.

The first path was realized by Wulfhekel and Kirschner [17–23], who periodically switched the magnetization direction of an amorphous CoFeNiSiB tip with a small coil wound around the tip (see Figure 1.4). Such a tip must exhibit a low coercivity, as this allows minimization of the coil size and thus the coil's stray field at the sample

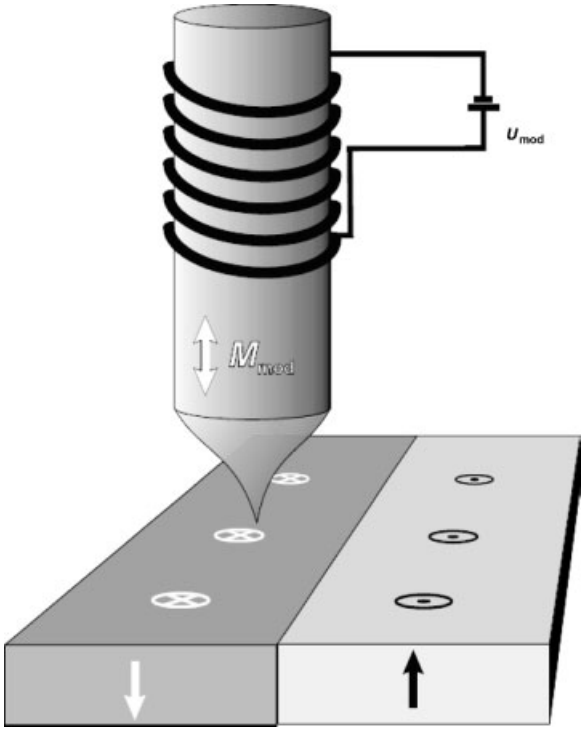


Figure 1.4 Schematic representation of the SP-STM as operated by Wulfhekel and Kirschner [17]. A soft magnetic tip is periodically magnetized in opposite directions along the tip axis by a small coil wound around the tip thus being sensitive to an out-of-plane magnetization. The resultant variation in tunneling current is measured using a lock-in technique. (From Ref. [24], with permission of IOP Publishing Ltd.)

position, as well as a low saturation magnetization. The most important precondition is an extremely low magnetostriction in order to suppress any modulation of the tip length due to the periodic remagnetization process. Consequently, the tips will be sensitive to the perpendicular component of the sample magnetization [21]. In order to realize an in-plane sensitivity, the same type of technique is applicable; however, the 'tip' would now consist of a ferromagnetic ring, the magnetization of which is also periodically switched by using a coil [25, 26].

In order to realize the second situation (see Ref. [27]), an *in situ* preparation of magnetic thin films is necessary. Typically, a polycrystalline tungsten (W) wire is electrochemically etched (this is important for tip stability). The W tip is heated to at least 2200 K upon introduction into the UHV chamber; otherwise, the magnetic coating material is frequently lost during the approach. This high-temperature flash removes oxides and other contaminants, thereby enhancing the binding between the

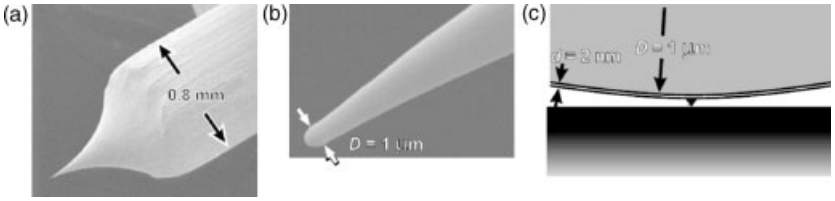


Figure 1.5 Scanning electron microscopy (SEM) images of an electrochemically etched, polycrystalline W tip after a high-temperature flash at $T > 2200$ K. (a) The overview shows the shaft of the tip, which exhibits a diameter of 0.8 mm; (b) High-resolution image of the very end of the tip. The tip apex has an angle of about 15° and the tip diameter amounts to approximately 1 µm; (c) Schematic representation of the tip apex (in scale). The magnetic film is very thin compared to the curvature of the tip. Most likely, a small magnetic cluster protrudes from the tip, and this is responsible for the lateral resolution of the SP-STM. (From Ref. [24], with permission of IOP Publishing Ltd.)

tip surface and the magnetic overlayer. While the overall shape of the tip [as shown in the scanning electron microscopy (SEM) overview image of Figure 1.5a] remains almost unaffected by the high-temperature treatment, the high-resolution SEM image shown in Figure 1.5b reveals that the tip diameter is increased to 1 µm, most likely due to melting of the tip apex. Following this high-temperature treatment, the tips are magnetically coated with a magnetic film exhibiting a thickness of several monolayers (MLs). In contrast to bulk tips, the magnetization direction is governed by the shape anisotropy. The anisotropy of thin film tips can thus be adjusted by selecting an appropriate film material. For example, while 3–10 ML Fe [27] and <50 ML Cr [28] are almost always sensitive to the in-plane component of the magnetization, 7–9 ML Gd [29], 10–15 ML $Gd_{90}Fe_{10}$ [30] and 25–45 ML Cr [30] are usually perpendicularly magnetized at low temperature. The well-known spin reorientation transition of Co films on Au (see Refs [31, 32]) which occur with increasing thickness of the magnetic material allows tuning of the magnetically sensitive direction of the tip with the *same* set of coating materials. For thin Co coverages (<8 ML) on a Au-coated W tip, an out-of-plane magnetic sensitivity is achieved, whereas for thicker Co films the in-plane component of the sample magnetization can be probed [33].

At least qualitatively, this observation can easily be understood. Two anisotropy terms are relevant: (i) the shape anisotropy which arises due to the pointed shape of the tip; and (ii) the surface and interface anisotropy of the film. The first term will always try to force the magnetization along the tip axis – that is, perpendicular to the sample surface. In contrast, the second term is material-dependent. Due to the rather large curvature of the tip as compared to the thickness of the coating film (see Figure 1.5c), the effective surface and interface anisotropy of a thin film can be deduced from an equivalent film on a flat W(110) substrate. For instance, in the case of 10 ML Fe the two anisotropy terms favor different directions. While the shape anisotropy still tries to force the magnetization along the tip axis, the ferromagnetic

film on W(110) exhibits a strong in-plane anisotropy [35] which obviously overcomes the shape anisotropy. An external field of 2 T is required to force the tip magnetization out of the easy (in-plane) into the hard (out-of-plane) direction [34]; this is consistent with results of Elmers and Gradmann [35] concerning thin film systems.

Even at room temperature magnetic thin film tips can be used for several days without losing their spin sensitivity. Initially, this is surprising as any surface is continuously exposed to the residual gas in the vacuum chamber which, depending on the reactivity of the sample under investigation, leads to a more or less rapid – but continuous – degradation of the surface spin polarization [36]. However, the geometry of the tunnel junction must be taken into account as it differs from that of an open surface. While residual gas particles may impinge onto a flat, uncovered surface from the entire half-space, the tip apex is almost completely shadowed by the sample, as shown schematically in Figure 1.5c. Thereby, gas transport onto the tip apex is dramatically reduced. Of course, the same argument can be applied to the sample surface which is, however, only valid for the particular location of the sample surface that is right under the tip apex. As this location varies when scanning the tip across the sample, the shadowing is only temporarily effective for any particular site of the sample surface, whereas the tip is shadowed at all times.

1.3.1.2 **Antiferromagnetic Probe Tips**

Despite the fact that the magnetic dipole interaction between the sample and the tip is considerably reduced for ferromagnetic ultrathin film coatings on a nonmagnetic tip, in comparison to thicker coatings or even bulk ferromagnetic tips, an additional influence cannot be ruled out. One straightforward and experimentally feasible solution, however, is to use an antiferromagnetically coated (see Ref. [30]) or a bulk antiferromagnetic tip consisting of, for example, a MnNi alloy [37–40]. The tip should exhibit no significant stray field, since opposite contributions compensate on an atomic scale, thus allowing the nondestructive imaging and investigation of spin structures even for magnetically soft samples. The spin sensitivity is determined solely by the orientation of the magnetic moment of the atom that forms the very end of the tip apex; the orientation of all other magnetic moments plays no role. Furthermore, the tip is insensitive to external fields, which allows direct access to intrinsic sample properties in field-dependent studies.

In order to demonstrate this insensitivity, we can refer to an investigation conducted by Kubetzka *et al.* [30]. Here, the response of an identically prepared system to an applied perpendicular magnetic field is shown using a ferromagnetic tip on the one hand, and an antiferromagnetic tip on the other hand. Figure 1.6a shows a series of dI/dU maps – that is, maps of the differential conductance, of 1.95 ML Fe on W(110) recorded with a ferromagnetic GdFe tip. Figure 1.6a(i) shows an overview of the initial state, while Figure 1.6a(ii) is taken at higher resolution, as indicated by the frame in Figure 1.6a(i). Because the coverage is slightly below 2 ML, narrow ML areas can be seen with a bright appearance at the chosen bias voltage. These ML areas efficiently decouple double layer (DL) regions on adjacent terraces. At 350 mT [see Figure 1.6a(iii)] the domain distribution is asymmetric; the bright domains have

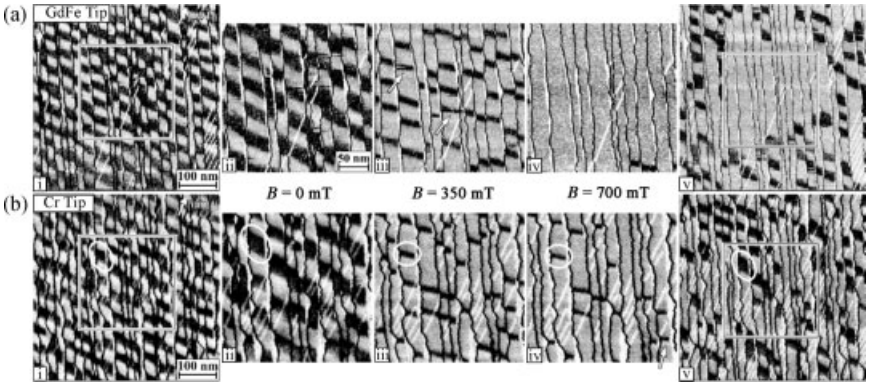


Figure 1.6 (a) d*I*/d*U* maps of 1.95 ML Fe/W (110) recorded with a GdFe tip (out-of-plane contrast): (i) $500 \times 500\,nm^2$ overview of magnetic initial state; (ii) $250 \times 250\,nm^2$ zoom-in; (iii) Asymmetry at B = 350 mT: dark domains are compressed and form 360° walls; (iv) saturation is observed within the field of view; (v) the influence of the tip's stray field becomes obvious in the overview recorded at $B = 0$ mT; (b) Analogous series of an identically prepared sample, recorded with a Cr-coated tip: (i, ii) magnetic initial state; (iii) asymmetry at B = 350 mT; (iv) a large fraction of the walls has survived at 700 mT, in contrast to (a); (v) the scanned area exhibits no significant difference in comparison to its surrounding. (Reprinted with permission from Ref. [30]; copyright (2002) American Physical Society.)

grown and the dark domains have shrunk. In some places the magnetic contrast changes abruptly from one horizontal scan line to the next (see arrows), this being the result of a rearrangement of the sample's magnetic state during the imaging process. At 700 mT [see Figure 1.6a(iv)] the sample has almost reached saturation within the field of view. However, it becomes obvious in the overview image, subsequently recorded at the same location in zero applied field [see Figure 1.6a(v)], that this field value does not reflect intrinsic sample properties. A large fraction of the dark domains has survived outside the region which was scanned previously at 700 mT. Thus, superpositioning of the applied field and the additional field emerging from the magnetic coating of the tip is much more efficient than the applied field alone.

Figure 1.6b shows an analogous series of images of a sample which was identically prepared but imaged with an antiferromagnetic Cr-covered tip. This exhibits an out-of-plane sensitivity, like the GdFe tips [30]. A dark domain is marked as an example to be recognized in all five images. The domain structure in Figure 1.6b(i)–(iii) displays no significant difference to the corresponding structures in Figure 1.6a. Since a rearrangement of the domain structure during imaging is not observed throughout this series, the occurrence of such events in Figure 1.6a can be attributed to the GdFe tip's stray field. As in Figure 1.6a(iii), the dark domains are compressed at 350 mT, which proceeds at 700 mT. At this field value, and in contrast to Figure 1.6a, a large fraction of the dark domains has survived. In the overview image of Figure 1.6b(v), which was taken again subsequently in a zero-applied field, the previously scanned area exhibits no significant difference in domain distribution in

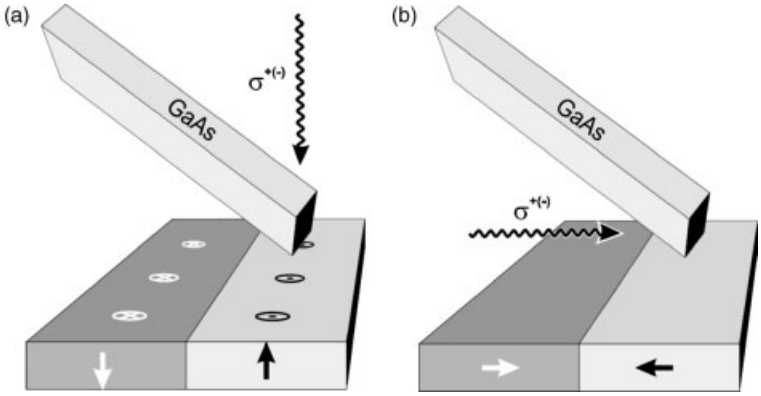


Figure 1.7 Schematic experimental set-up to realize spin-polarized electron tunneling using optically pumped GaAs tips, as proposed in Ref. [3]. (a) Spins aligned perpendicular to the sample surface can be detected by a GaAs tip excited by helical light incident along the surface normal; (b) When the incident light is along the sample surface the experiment is sensitive to electron spins parallel to the surface plane. (Reprinted from Ref. [24], with permission of IOP Publishing Ltd.)

comparison to its surrounding. This result directly demonstrates the advantage of a stray field free tip.

1.3.1.3 Optically Pumped GaAs Probe Tips

The concept of spin-polarized tunneling between a magnetic surface and an optically pumped GaAs tip (as discussed below) was first proposed by Pierce [3]. This experimental approach allows both the sign and the polarization direction of photoexcited electrons to be modified, simply by choosing an appropriate laser light helicity and experimental geometry. Varying the spin polarization of the tunneling electrons with a simultaneously constant intensity of the incident light enables the magnetic effects to be separated from the topographic and electronic effects. The corresponding schematic arrangement (see Figure 1.7) proves that the experiment can be made sensitive to either the out-of-plane or the in-plane magnetization direction by changing the direction of the incident light to be parallel and perpendicular to the sample surface, respectively.

Optically pumped *p*-type GaAs is widely used as a source to produce spin-polarized electrons close to the Fermi level. The physical principle is based on two properties of this material: (i) It is a direct band-gap semiconductor; and (ii) the degeneracy of the *p*-like valence band is lifted by spin–orbit interaction into a fourfold degenerate $p_{3/2}$ level (Γ_8 band edge) and a twofold degenerate $p_{1/2}$ level (Γ_7 band edge). The spin–orbit splitting amounts to $\Delta_{so} = 0.34$ eV. If circularly polarized light (σ^+ or σ^-) with an energy slightly above the energy gap of $E_{gap} = 1.52$ eV of p-GaAs is irradiated onto the sample, the electronic transition in GaAs must fulfill the optical selection rule $\Delta m_j = m_f - m_i = \pm 1$, where $m_{f,i}$ is the angular momentum of the final and initial states, respectively (see Figure 1.8). When using σ^+ light, the relative transmission probability into $m_j = -1/2$ states is threefold higher than that into $m_j = -3/2$ states

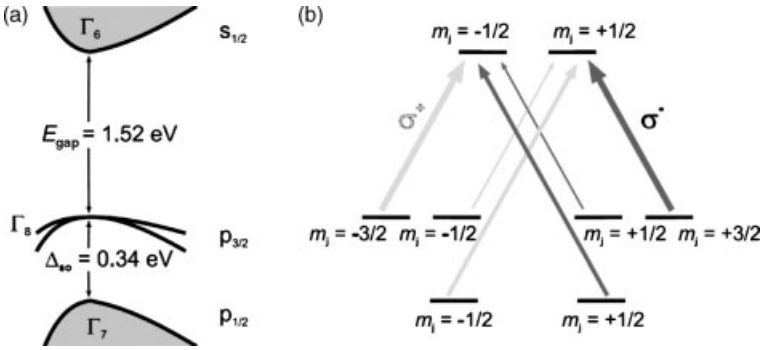


Figure 1.8 (a) Schematic band structure of GaAs in the vicinity of the Γ-point of the Brillouin zone. The width of the band gap between the conduction (Γ_6) and the fourfold degenerate $p_{3/2}$ valence band edge (Γ_8) amounts to 1.52 eV. Another 0.34 eV lower there is the twofold degenerate $p_{1/2}$ level (Γ_7); (b) Allowed transitions between different m_j sublevels for circularly polarized light of opposite helicity (σ^+ and σ^-). The transition probability is represented by the thickness of the arrows. (Reprinted from Ref. [24], with permission of IOP Publishing Ltd.)

(and vice versa for σ^- light). As a result, while the theoretical limit of the electron spin polarization in photoemission is $\pm 50\%$, values as high as 43% have been achieved experimentally [41].

The preliminary experiments on GaAs–insulator–ferromagnet tunnel junctions were reported in Ref. [42]. The specimen was prepared by cleaving a GaAs crystal in air along the (1 1 0) plane. A 20–40 Å layer of Al was then evaporated onto the GaAs (1 1 0) surface and subsequently oxidized. Onto this insulating barrier was then deposited a 150 Å thick Co film, which was itself protected by a 50 Å Au cap layer. The chosen experimental set-up required that the light would traverse the ferromagnetic layer and the insulator before reaching the semiconductor. After magnetization of the Co film perpendicular to the plane, a dependence of the tunneling current on the helicity of the light was measured, which suggested the existence of spin-polarized transport. However, an even stronger signal was detected when there was no tunneling barrier between the semiconductor and the ferromagnet' this was explained by "...an intensity modulation of the circularly polarized light upon transmission through the magnetically ordered layer, determined by the polar magneto-optic coefficients" [42]. The reduction of the helical asymmetry when using a tunnel barrier, compared to a situation without any barrier, was explained in a later analysis [43, 44] by a negative tunneling conductance. As this method to create spin-polarized electrons involves neither a magnetic material nor magnetic fields, it offers excellent conditions for application in SP-STM.

In spite of this favorable situation, GaAs tips have not yet been used successfully for the imaging of magnetic surfaces. Similar to the experiments performed with planar junctions, this may be caused by difficulties in separating spin-polarized tunneling from magneto-optical effects [45–49].

Compared to the proposal of Pierce [3] (as shown in Figure 1.7), the first successful observation of spin-polarized electron tunneling with the scanning tunneling

microscope, using a GaAs electrode [50], was obtained by exchanging the role of tip and sample – that is, by using a Ni tip and a GaAs(110) surface. Moreover, instead of optically pumping the GaAs sample and thereby producing spin-polarized charge carriers in the GaAs, the reverse process was used, with spin-polarized electrons being injected from the Ni tip into the conduction band of GaAs. Upon transition of the injected electrons from this metastable state in the conduction band into the final state in the valence band recombination, luminescence was seen to occur and the circular polarization of the emitted photons was analyzed.

Evidence for a second explanation for the failure of magnetic imaging with GaAs electrodes – namely an insufficient lifetime of the spin carriers at the tip apex – comes from analogous STM-excited luminescence experiments performed with single crystalline Ni(110) tips and a stepped GaAs(110) sample [51]. With the tip positioned above flat terraces, a high-spin injection efficiency was measured. However, the intensity of the recombination luminescence on the upper terrace was found to have decreased by a factor of 1000. Simultaneously, the polarization decreased by a factor of 6. This observation was explained by a reduction of either the spin injection efficiency or of the spin relaxation lifetime, and attributed to the metallic nature of the step edge caused by midgap states of the (111) surface. As a sharp tip must possess numerous step edges around the apex atom, it is a straightforward conclusion that the spin relaxation lifetime may be drastically reduced at the very end of the tip.

1.3.1.4 **Nonmagnetic Probe Tips**

Surprisingly, even nonmagnetic tips can be used to image certain magnetic sample properties, as demonstrated by Bode *et al.* [52, 53] and by Pietzsch *et al.* [54]. The images shown in Figure 1.9a and b [54] were taken for slightly less than 2 ML Fe on

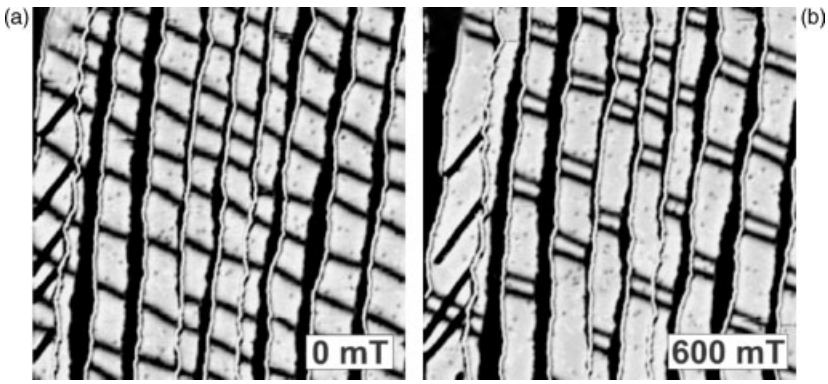


Figure 1.9 Domain walls as observed with a nonmagnetic W tip. (a) No external field applied; (b) Taken at 600 mT. The external field enforces pair formation. Sensitivity to the spin orientation inside the walls is lost, and all walls are imaged as dark lines. Thus, the image provides information on the magnetization lying along an easy or a hard direction. Note the five lines in the left stripe running in a direction bottom-left to top-right; these are not domain walls but are dislocation lines. (Reprinted from Ref. [54], with permission of Springer. Copyright (2004).)

W(110), which allows a comparison with the results shown in Figure 1.6 (which were obtained using a magnetic tip). However, a tungsten tip without any magnetic coating was now used, whereupon the dark lines revealed the presence of domain walls. The main difference between the two measurements was that the nonmagnetic tip did not provide sensitivity to the spin orientation inside the walls. Instead, both walls of a pair were imaged equally, in contrast to the observation made with the ferromagnetic tip [55] (cf. Figure 1.26). This is a consequence of the fact that the measurement made with the W tip does not involve spin-polarized tunneling; rather, it is the spin-averaged electronic structure of the sample that gives rise to the signal. The electronic structure of the DL stripes is locally modified due to the presence of a domain wall. In other words, the electronic structure is sensitive to whether the magnetization is in an easy or a hard direction. First-principle calculations have shown [52, 53] that the spin–orbit-induced mixing of different *d*-states depends on the magnetization direction, and changes the local density of those states that are detectable by non-spin-polarized STS.

As an important implication of this effect, the magnetic nanostructure of surfaces can be investigated with conventional nonmagnetic tips. This has the clear advantage that there is definitely no dipolar magnetic stray field from the tip that could interfere with the sample. In addition, the preparation of a magnetic tip is omitted.

1.3.2 Modes of Operation

In the following sections, different modes of operation enabling to achieve a magnetic contrast using magnetic probe tips will briefly be discussed from a theoretical point of view according to Ref. [56].

1.3.2.1 Constant Current Mode

In the situation when the tip and sample are magnetic, the tunneling current can be described as a sum of a spin-averaged I_0 and a spin-dependent term I_{sp} [56] (cf. Equation 1.8):

$$I(\vec{r}_t, U, \theta) = I_0(\vec{r}_t, U) + I_{sp}(\vec{r}_t, U, \theta) \tag{1.9}$$

$$= const. \cdot (n_t \tilde{n}_s(\vec{r}_t, U) + \vec{m}_t \tilde{\vec{m}}_s(\vec{r}_t, U)) \tag{1.10}$$

with n_t being the non-spin-polarized local density of states (LDOS) at the tip apex, $\tilde{n}_s$ the energy-integrated LDOS of the sample, and $\vec{m}_t$ and $\tilde{\vec{m}}_s$ the vectors of the (energy-integrated) spin-polarized LDOS with:

$$\tilde{\vec{m}}_s(\vec{r}_t, U) = \int \vec{m}_s(\vec{r}_t, E) \mathrm{d}E \tag{1.11}$$

and θ the angle between the magnetization direction of tip and sample. For a non-spin-polarized STM experiment – that is, using either a nonmagnetic tip or a nonmagnetic sample – the second term, I_{sp}, vanishes.

The constant current mode is restricted to some limited cases, which is partly due to the integral in Equation 1.11 and taken over all energies between the Fermi energy E_F and eU, with U being the applied bias voltage because I_{sp} becomes reduced if the spin polarization changes sign between E_F and eU. Furthermore, the magnetically induced corrugation is small compared to the topographic and electronic corrugation; this is due to the exponential dependence of the tunneling current on the distance between tip and sample. Nevertheless, it is still possible to obtain information of complex atomic-scale spin structures at ultimate magnetic resolution (as shown in Ref. [56]), although it is necessary to understand the influence of the tip [57, 58].

1.3.2.2 Spectroscopy of the Differential Conductance

The difficulties of separating the topographic, electronic and magnetic information can be overcome by measuring the differential conductance, $\mathrm{d}I/\mathrm{d}U$, with a magnetic tip [56]:

$$\mathrm{d}I/\mathrm{d}U(\vec{r}_t, U) \propto n_t\tilde{n}_s(\vec{r}_t, E_F + eU) + \vec{m}_t\tilde{\vec{m}}_s(\vec{r}_t, E_F + eU) \tag{1.12}$$

In this situation, the measured quantity no longer depends on the energy-integrated spin polarization, but rather on the spin polarization in a narrow energy window ΔE around $E_F + eU$.

The differential conductance is determined experimentally by adding a small ac-voltage to the dc-bias voltage at a frequency which is significantly above the cut-off frequency of the feedback loop that ensures a constant current. The amplitude of the ac-voltage is responsible for the width ΔE. The current modulation is amplified by means of a lock-in technique.

The electronically homogeneous surfaces maps of differential conductance reflect the magnetic behavior, since any variation of the signal must be due to the second spin-dependent term, I_{sp}. The situation becomes more complicated for electronically heterogeneous surfaces; nevertheless, a careful comparison between spin-averaged and spin-resolved measurements often allows a distinction to be made between topographic and electronic contrast compared to the magnetically induced information. However, this set of experiment requires measurements to be made with both nonmagnetic and magnetic probe tips.

The determination of differential conductance also provides access to the spin polarization of a surface which is *locally resolved* [59–61].

The recording of inelastic tunneling spectra (i.e. the second derivative of the conductance $\mathrm{d}^2I/\mathrm{d}U^2$) with a magnetic tip in an external magnetic field, it becomes possible to study directly – that is, without a separating insulating layer – magnon excitations in magnetic nanostructures [62].

1.3.2.3 Local Tunneling Magnetoresistance

As an alternative method, the local tunneling magnetoresistance $\mathrm{d}I/\mathrm{d}\vec{m}_t$ between the magnetic tip and magnetic sample can be determined by modulation of the tip magnetization direction and determining the variation of the tunneling current using a lock-in technique. This type of measurement was first proposed by Johnson

and Clarke [63] and later accomplished by Wulfhekel and Kirschner [17]. By taking the derivative of Equation 1.10 one obtains:

$$\mathrm{d}I/\mathrm{d}\vec{m}_t \propto \tilde{\vec{m}}_s \tag{1.13}$$

Thus, the signal is proportional to the energy-integrated spin-polarized LDOS. One significant advantage of this technique relates to the detailed knowledge of the magnetization of the probe tip. A nonvanishing signal is also obtained only if a local magnetization is present. Furthermore, this method allows the investigation of samples in a single domain state; this situation differs from the spectroscopy of differential conductance, which demands that different magnetic domains are simultaneously visible in a single image. Consequently, due to the direct detection of magnetic information, knowledge of the electronic properties is no longer required.

It must be borne in mind, however, that the interpretation of chemically heterogeneous surfaces (e.g. of alloys) remains difficult. Both, the sign and magnitude, of the element-specific spin polarization may vary, thereby avoiding any direct identification of the domain structure. Nonetheless, the chemical contrast plays an additional role.

1.4
Magnetic Arrangement of Ferromagnets

In this section, we will demonstrate the magnetic arrangement of ferromagnetic systems for systems which exhibit localized magnetic moments, and also which represent an itinerant or band magnet. A typical representative of the former group is the rare-earth metal gadolinium, while the transition metal cobalt is typical of the latter group.

1.4.1
Rare-Earth Metals: Gd/W(110)

In analogy to the low-temperature experiments performed with ferromagnet–insulator–superconductor planar tunneling junctions [64, 65], where the quasiparticle density of states of superconducting aluminum is split by a magnetic field into spin up and spin down parts, two spin-polarized electronic states with opposite polarization can be used to probe the magnetic orientation of the sample relative to the tip, thus enabling spin-polarized scanning tunneling spectroscopy (SP-STS) [60].

The principle of SP-STS is shown schematically in Figure 1.10, using a sample which exhibits an exchange split-surface state with a relatively small exchange splitting, Δ_{ex}. This situation is, for example, realized for the Gd(0001) [66–70], Tb (0001) [71] and Dy(0001) surfaces [72]. If Δ_{ex} is too large, then one spin component would be too far from the Fermi level and not accessible by STS, as for example in the case of Fe(001), where Δ_{ex} amounts to 2.1 eV and only the minority band appears as a

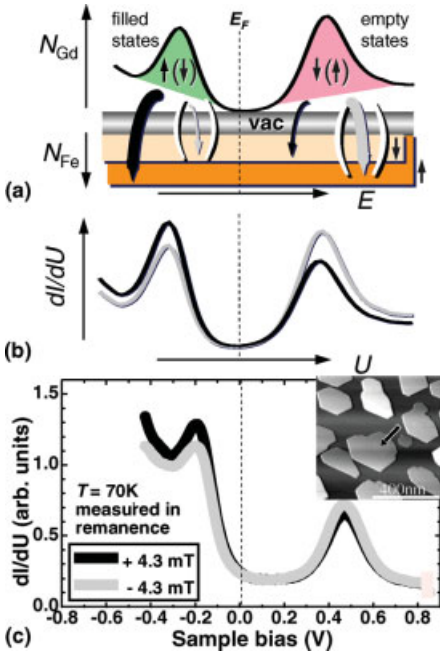


Figure 1.10 (a) The principle of SP-STS using a sample with an exchange split surface state, for example, Gd(0001), and a magnetic Fe tip with a constant spin polarization close to E_F. Due to the spin valve effect the tunneling current of the surface state spin component being parallel to the tip is enhanced at the expense of its spin counterpart; (b) This should lead to a reversal in the dI/dU signal at the surface state peak position upon switching the sample magnetically; (c) Exactly this behavior could be observed in the tunneling spectra measured with the tip positioned above an isolated Gd island (see arrow in the inset). (Reprinted with permission from Ref. [27]; copyright (1998), American Physical Society.)

peak in the dI/dU spectra just above the Fermi level [73]. In contrast, the majority (minority) part of the Gd(0001) surface state has a binding energy of −220 meV (500 meV) at 20 K [71]; that is, the exchange splitting only amounts to 700 meV, far below the Curie temperature of 293 K.

In the following section we consider vacuum tunneling between a Gd(0001) surface and a tip material; for simplicity, we assume a constant spin polarization (see Figure 1.10a, lower part). If the magnetization direction of the tip remains constant, then two possible magnetic orientational relationships occur between the tip and sample – parallel or antiparallel – under the assumption that the magnetization of the tip and sample is aligned. Since, however, both the majority and the minority component of the Gd(0001) surface state appear in the tunneling spectra, the spins of one component of the surface state will in any case be *parallel* with the tip, while the spins of the other component will be *antiparallel*. Therefore, the spin valve effect will act differently on the two spin components; due to the strong spin dependence of the density of states, the spin component of the surface state parallel to the tip magnetization is enhanced at the expense of its counterpart being antiparallel.

Consequently, by comparing tunneling dI/dU spectra measured above domains with opposite magnetization, one expects a reversal in contrast at the majority and minority peak positions (see Figure 1.10b).

Tunneling spectra measured in an external magnetic field with an in-plane sensitive ferromagnetic probe tip positioned above an isolated Gd(0001) island show exactly the expected behavior (see Figure 1.10c). The sample was magnetized in a magnetic field of $+4.3$ mT applied parallel to the sample surface, and the spectra were subsequently measured in remanence with the tip positioned above the Gd island (this is marked by an arrow in the inset of Figure 1.10c). The direction of the magnetic field was then reversed (to -4.3 mT) and further dI/dU spectra were monitored at the same location. Figure 1.10c shows the tunneling spectra measured in remanence after the application of a positive or negative field. A comparison of the spectra reveals that for a positive field, the differential conductance dI/dU measured at a sample bias which corresponds to the binding energy of the occupied (majority) part of the surface state, is higher than for negative field. The opposite is true for the empty (minority) part. Freestanding Gd islands on W(110) were chosen for this experiment, since it is known from Kerr effect measurements [74] that the coercivity is only 1.5 mT – that is, much lower than the applied field. Thus, it can be safely concluded that the magnetization of the sample was switched by the external field while the tip magnetization remained unchanged.

Further information relating to magnetic imaging of the Gd(0001) surface can be found in Refs [36, 68, 70], while data concerning the surface of another rare earth metal, Dy(0001), are available in Ref. [72].

1.4.2
Transition Metals: Co(0001)

The domain structure of the surface of a Co(0001) single crystal has been studied by Wulfhekel *et al.* [17–23]. The uniaxial magnetocrystalline anisotropy of hcp-Co points along the *c*-axis – that is, perpendicular to the (0001) surface. However, the total energy of the sample is minimized by the formation of surface closure domains where the magnetization locally tilts towards the surface plane, thereby reducing the dipolar energy. As the magnetocrystalline anisotropy energy and dipolar energy are similar in size, and the in-plane components of the magnetocrystalline anisotropy energy are almost degenerate, a complicated dendritic pattern is formed at the surface. Figure 1.11a shows the typical dendritic-like perpendicular domain pattern of Co(0001) as measured by Wulfhekel *et al.* [23]. Due to the fact that the tip magnetization is intentionally modulated by a small coil, the bright and dark locations in Figure 1.11a can be assigned to specific magnetic orientations, namely the magnetization vector points out of or into the surface, respectively. A sharp contrast can be recognized in Figure 1.11b, which is completely absent in the topographic image [21]. The absence of any correlation confirms that this contrast is not caused by different local structural or electronic properties; rather, it represents a domain wall separating two regions of different magnetization directions. This

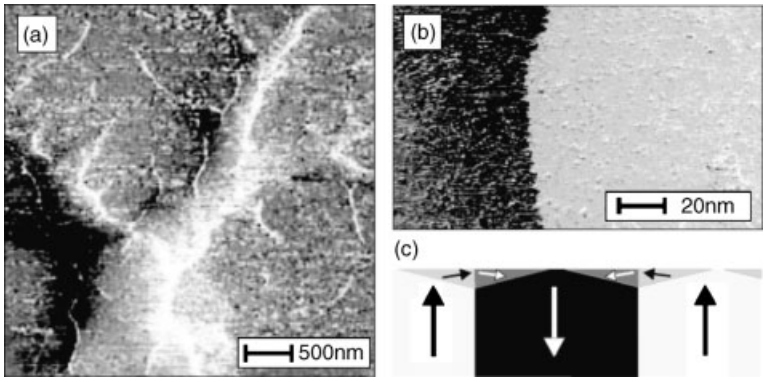


Figure 1.11 (a) Magnetic domain image on Co(0001) obtained by SP-STM; (b) A sharp domain wall at the end of a branch, at high magnification; (c) Schematic cross-section of the closure domain pattern of Co. (Reprinted with permission from Ref. [23]; copyright (2003), EDP Sciences.)

domain wall is found to correspond to the domain wall across two canted surface domains (see Figure 1.11c).

Further information concerning the ferromagnetic transition metal Fe on W(110) and Mo(110) is provided below. The spin-resolved electronic properties of dislocation lines that occur during thin film growth of Fe films on W(110) are described in Ref. [75], while details of the complex magnetic structure of Fe on W(001) are reported in Refs [76, 77]. The easy magnetization axis was shown to be layer-dependent, whereas the second and third Fe layers were magnetized along $\langle 110 \rangle$ or equivalent directions; the easy axis of the fourth layer was rotated by 45°.

1.5 Spin Structures of Antiferromagnets

The lateral averaging of magnetically sensitive techniques often fails in the imaging of antiferromagnetic surfaces because the overall detected magnetization is equal to zero. Here, we will show how SP-STM can be used to overcome this difficulty. The first example is the topological anti-ferromagnetism of Cr(001); a second example, namely the atomic resolution of magnetic behavior, will be demonstrated using the antiferromagnetic Mn monolayer on W(110) and Mn_3N_2 films on MgO(001).

The antiferromagnetic nature was additionally reported for the first monolayer of Fe on W(001) [77]. The antiferromagnetic coupling between a whole atomic layer and a ferromagnetic substrate was investigated for Mn on Fe(001) [25, 26, 78]; a Co layer on a Cu-capped Co substrate exhibits a ferromagnetic or antiferromagnetic coupling, depending on the interlayer thickness [79]. Magnetite Fe_3O_4 represents a ferrimagnet with a high spin polarization near the Fermi level. SP-STM was used to investigate the corresponding (001) [39, 40, 80–83] and (111) surfaces [40, 84].

1.5.1
Topological Anti-Ferromagnetism of Cr(001)

The Cr(001) surface for which the topological step structure is directly linked to the magnetic structure represents a topological anti-ferromagnet; that is, each terrace exhibits a ferromagnetic alignment of the magnetic moments, although between two adjacent terraces the magnetization possesses an antiparallel orientation that was predicted, on a theoretical basis, by Blügel *et al.* [85].

By using the scanning possibilities, the antiferromagnetic coupling between neighbored terraces of a Cr(001) surface can be imaged directly [70, 86–94]. The topography (see Figure 1.12a) presents a regular step structure with terrace widths of about 100 nm [88]. The line section in the bottom of Figure 1.12a shows that all step edges in the field of view are of single atomic height – that is, 1.4 Å. This topography should lead to a surface magnetization that periodically alternates between adjacent terraces and, indeed, this was observed experimentally (see Figure 1.12b). The line section of the differential conductance drawn along the same path as in Figure 1.12a indicates two discrete levels with sharp transitions at the positions of the step edges.

The typical domain wall width, as measured on a stepped Cr(001) surface, amounts to 120–170 nm [86]. In analogy to ferromagnetic domain walls (these are discussed in detail in Chapter 9), this value is determined by intrinsic material parameters – that is, the strength of the exchange coupling and the magnetocrystalline anisotropy. Clearly, the domain wall width cannot remain unchanged very close to a screw

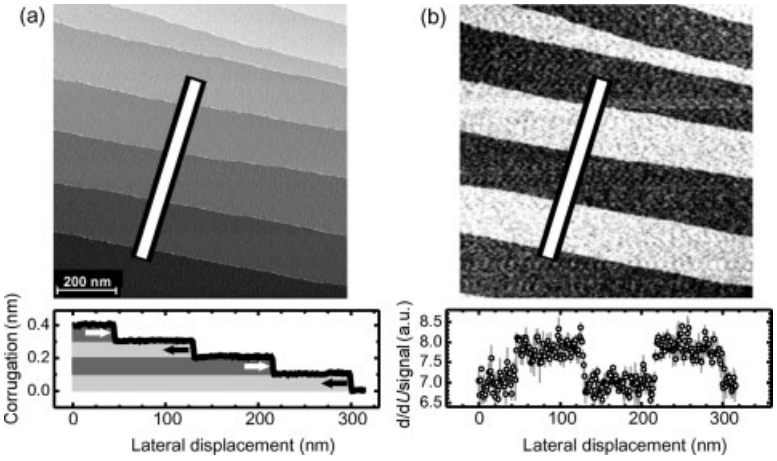


Figure 1.12 (a) Topography and (b) spin resolved map of the d*I*/d*U* signal of a clean and defect free Cr(0 0 1) surface as measured with a ferromagnetic Fe-coated tip. The bottom panels show averaged sections drawn along the line. Adjacent terraces are separated by steps of monatomic height. (Reprinted with permission from [88]; copyright (2003) by the American Physical Society.).

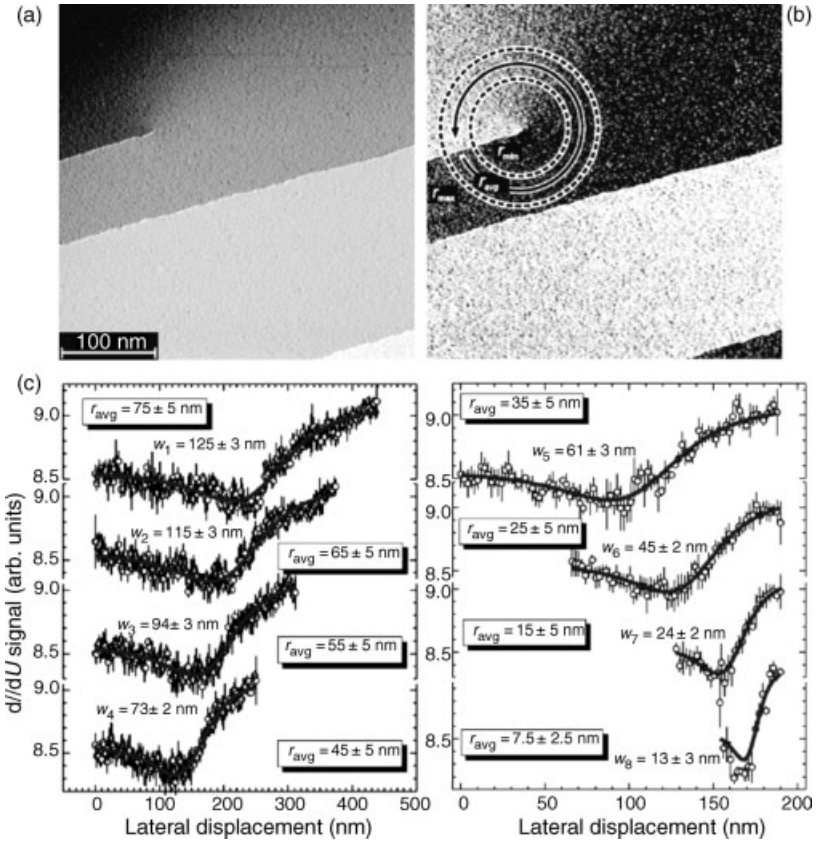


Figure 1.13 (a) Topography and (b) magnetic d*I*/d*U* signal of a Cr(001) surface with a single screw dislocation. The magnetic frustration leads to the formation of a domain wall between the dislocation; (c) Circular sections drawn at different radii around the center of the screw dislocation. (Reprinted with permission from Ref. [88]; copyright (2003), American Physical Society.)

dislocation where the circumference becomes comparable with or smaller than the intrinsic domain wall width.

The dependence of domain wall width on the distance from the screw dislocation of the Cr(001) surface is shown in Figure 1.13a [88]. Here, approximately 100 nm from the next step edge, a single screw dislocation can be recognized in the upper left corner of the image. The magnetic d*I*/d*U* map of Figure 1.13b reveals that this screw dislocation is the starting point of a domain wall which propagates towards the upper side of the image. Starting at the tail of the arrow (zero lateral displacement), eight circular line sections are drawn counterclockwise around the screw dislocation at different radii r_{avg}, from 75 nm down to 7.5 nm; these data are plotted in Figure 1.13c. The domain walls were fitted using the model provided in Chapter 9. The results are shown as gray lines in Figure 1.13c; except for the smallest average

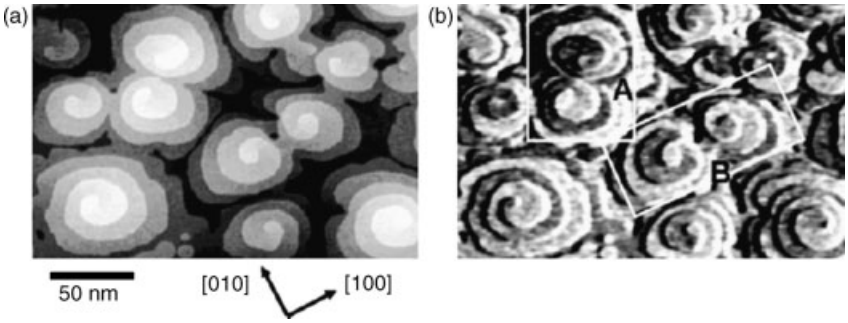


Figure 1.14 STM images of (a) the topography and (b) dI/dU magnetic signal obtained simultaneously from the same area of a 9 nm-thick Cr(001) film. (Reprinted with permission from Ref. [93]; copyright (2007), Elsevier.)

radius ($r_{avg} = 7.5$ nm), an excellent agreement with the experimental data was found. At an average radius $r_{avg} = 75$ nm, the domain wall width amounted to 125 nm, being in close agreement with the intrinsic domain wall width of Cr(001) as determined far away from screw dislocations. This was not surprising, as the circumference amounted to about 500 nm – much larger than the intrinsic domain wall width. However, as soon as r_{avg} was reduced below 60 nm a significant reduction in domain wall width was observed, although the circumference still exceeded the intrinsic domain wall width. The results showed clearly that the domain wall width was always considerably narrower than the circumference of the cross-section.

We can now discuss a more complex structure, namely the influence of the distance and chirality between two adjacent spiral terraces on magnetic structures on Cr(001) films [90, 93]. Figure 1.14a shows a topographic STM image of a 9 nm-thick Cr(001) film [93] where the feature of the surface morphology is that the Cr layers form high-density, self-organized spiral terraces. Each terrace is displaced by a monatomic step height, and a screw dislocation is clearly visible in the center of each spiral pattern. The typical diameter of these spiral terraces is 50 nm. A complex spin frustration and characteristic magnetic ordering is present, being restricted by the topological asymmetry of the spiral terraces. Figure 1.14b shows the dI/dU magnetic image obtained simultaneously in the same area of Figure 1.14a, exhibiting a magnetic contrast. A comparison of the two images of Figure 1.14a and b reveals that most parts of the observed magnetic contrasts are consistent with a topological antiferromagnetic structure. The maximum magnetic contrast corresponds to the topological antiferromagnetic order, and a deviation from the maximum magnetic contrast can be recognized as the spin frustration which appear in the region near the screw dislocations and between two spirals.

The magnetic structure can be deduced from the observed dI/dU magnetic signal intensity by assuming the orientation of the tip magnetization parallel to the bcc [100] direction. For example, the derived magnetic structures of two adjacent spirals (the regions A and B indicated in Figure 1.14b) are shown in Figure 1.15 by arrows. Although the two adjacent spirals have the same chirality, the sign is opposite

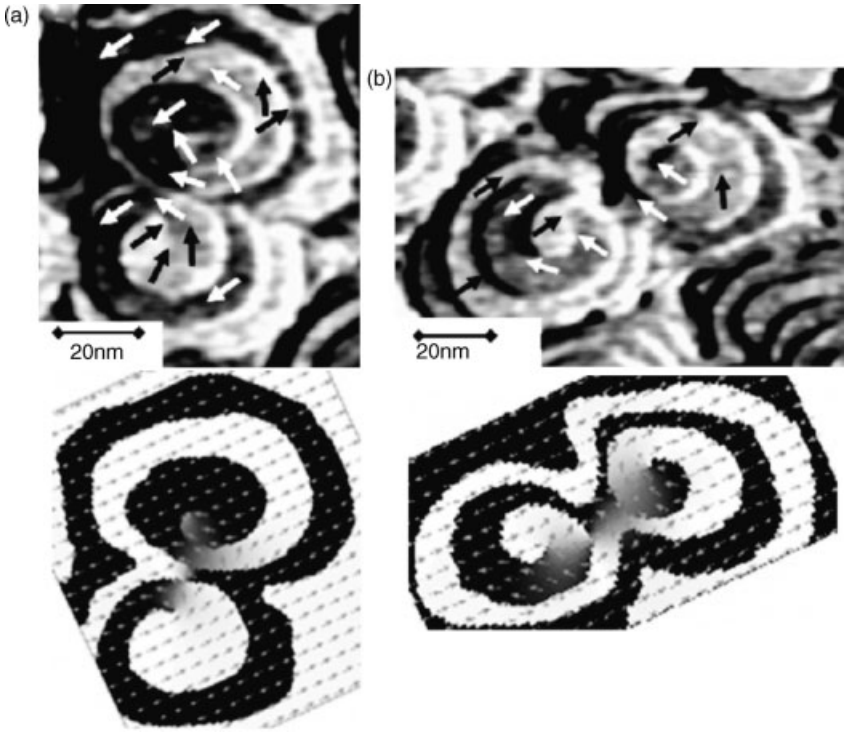


Figure 1.15 Observed (upper) and simulated (lower) magnetic structures of two adjacent spirals: (a) region A and (b) region B indicated in Figure 1.14b. The directions of the magnetization are represented by the arrows. The distance between the two screw dislocations is 20 nm (region A) and 32 nm (region B). (Reprinted with permission from Ref. [93]; copyright (2007), Elsevier.)

between regions A and B; the distance between the two screw dislocations d_s is 20 and 32 nm for the regions A and B, respectively. Although the magnetization rotates gradually around the center of the spiral in the case of $d_s = 75$ nm [90], this does not occur for $d_s = 20$ nm and 32 nm, which suggests that the spin-frustrated region decreases with decreasing d_s. There seems to be little difference in the sign of chirality of spirals.

In order to understand its origin, we can calculate the magnetic structure of these spiral terraces [93]; the result is shown in the two lower panels in Figure 1.15. The white (black) contrast represents the magnetization to be parallel (antiparallel) to the [100] direction. The topological antiferromagnetic order appears in a series of adjacent terraces, as well as in the most part of spiral terraces; the frustrated regions (the gray regions in the simulated figures) are evident between the center of adjacent spirals. It should be noted that the observed and calculated magnetic structures are clearly asymmetric with respect to the straight line connecting two screw dislocations, in spite of the different d_s-values. The simulated spin alignments are in good qualitative agreement with the observed results.

1.5.2
Magnetic Spin Structure of Mn with Atomic Resolution

The deposition of Mn on W(110) in the submonolayer regime results in a pseudomorphic growth; that is, Mn mimics the bcc symmetry as well as the lattice constant of the underlying substrate [95]. By using a clean W tip, atomic resolution could be achieved on the Mn islands, as demonstrated by Heinze *et al.* [96] (see Figure 1.16a). Additional information is provided in Refs [70, 97]. The diamond-shaped unit cell of the (1×1)-grown Mn ML is clearly visible. The line section drawn along the dense-packed $[1\,\bar{1}\,1]$ direction exhibits a periodicity of 0.27 nm, which almost perfectly fits the expected nearest-neighbor distance of 0.274 nm. The measured corrugation amplitude amounts to 15 pm. A calculated STM image for a conventional tip without spin polarization is given for comparison (see inset of Figure 1.16a). Clearly, the qualitative agreement between theory and experiment is excellent.

In a second set of experiments [96], different ferromagnetic tips were used. Since it is known from first-principles computations that the easy magnetization axis of the Mn ML on W(110) is in-plane [96], the experiment required a magnetic tip with a magnetization axis in the plane of the surface in order to maximize the effects. This condition is fulfilled by Fe-coated probe tips [27]. Figure 1.16b shows an STM image taken with such a tip, where the periodic parallel stripes along the [001] direction of the surface can be recognized. The periodicity along the $[1\,\bar{1}0]$ direction amounts to 4.5 Å, which corresponds to the size of the magnetic c(2×2) unit cell. The inset in Figure 1.16b shows the calculated STM image for the magnetic ground state, that is, the c(2×2)-antiferromagnetic configuration. Thus, theory and experiment give a consistent picture. Even the predicted faint constrictions of the stripes along the [001] direction related to the pair of second-smallest reciprocal lattice vectors are visible in the measurement. Again, experimental and theoretical data can be compared more quantitatively by drawing line sections along the dense-packed $[1\,\bar{1}1]$ direction (see Figure 1.16b). The result, which is plotted in Figure 1.16c, reveals that the periodicity, when measured with a Fe-coated probe, is twice the nearest-neighbor distance (i.e. 0.548 nm).

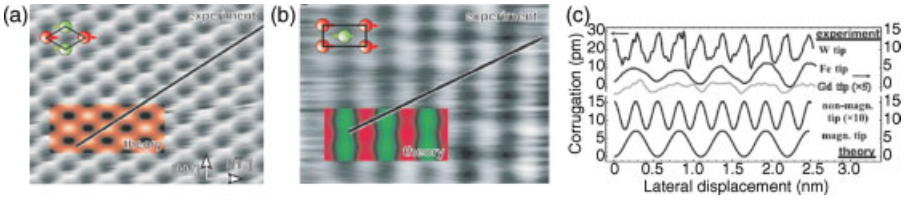


Figure 1.16 Comparison of experimental and theoretical STM images of a Mn ML on W(110) with (a) a nonmagnetic W tip and (b) a magnetic Fe tip; (c) Experimental and theoretical line sections for the images in (a) and (b). The unit cell of the calculated magnetic ground-state configuration is shown in (a) and (b) for comparison. The image size is 2.7×2.2 nm. (Reprinted with permission from Ref. [96]; AAAS.)

The pronounced dependence of the effect on the magnetization direction of the tip can be exploited to gain further information on the magnetization direction of the sample. This is done by using a tip that exhibits an easy magnetization axis that is almost perpendicular to the one of the sample surface. This condition is fulfilled by a W tip coated with about 7 ML Gd [29]. The gray line in Figure 1.16c represents a typical line section as measured with such a Gd-coated probe tip. Indeed, the corrugation amplitude was always much smaller than that for Fe-coated tips and never exceeded 1 pm, thus supporting the theoretical results that the easy axis of the Mn atoms is in-plane.

In the following section it will be shown, by reference to the studies of Yang *et al.* [98] and Smith *et al.* [99], that both magnetic and nonmagnetic atomic-scale information can be obtained simultaneously in the constant current mode for another Mn-based system which consists of Mn_3N_2 films grown on MgO(001) with the c axis parallel to the growth surface, which is (010). The surface geometrical unit cell, containing six Mn atoms and four N atoms (3 : 2 ratio), can be denoted as $c(1 \times 1)$, whereas the surface magnetic unit cell is just (1×1).

The bulk structure of Mn_3N_2 exhibits a face-centered tetragonal (fct) rock salt-type structure. The bulk magnetic moments of the Mn atoms are ferromagnetic within (001) planes, lie along the [100] direction, and are layerwise antiferromagnetic along [001]. Besides the magnetic superstructure, every third (001) layer along the c direction has all N sites vacant, which results in a bulk unit cell exhibiting $c = 12.13$ Å (six atomic layers).

Figure 1.17a presents a SP-STM image [98] of the surface acquired using a Mn-coated W tip thus being sensitive to the in-plane direction. Although the row structure with period $c/2$ is observed, a modulation with period c of the height of the rows is additionally obvious. The modulation shown in Figure 1.17b is evident for both domains D1 and D2 by the area-averaged line profiles taken from inside the boxed regions on either side of the domain boundary. For the domain D1 (red line), the modulation amplitude is about a factor of 2 larger than for the domain D2 (blue line). As the height modulation is proportional to $m_t m_s \cos\theta$, with m_t and m_s being the

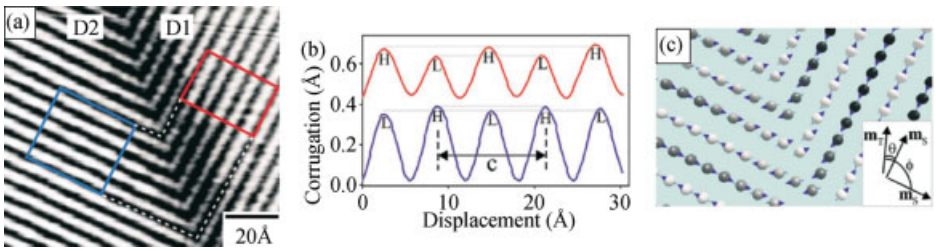


Figure 1.17 (a) SP-STM image acquired using a Mn-coated W tip; (b) Two area-averaged line profiles (red and blue) corresponding to the regions inside the red and blue rectangles in (a); (c) Simulated SP-STM map: contrast: white ↔ black ⇒ θ: 0 ↔ π. The inset shows the moments of tip ($\vec{m}_t$) and the sample ($\vec{m}_s$) for the two different domains and the angles between them. Each ball represents a magnetic atom. (Reprinted with permission from Ref. [98]; copyright (2002), American Physical Society.)

moment of the tip and sample, respectively, and θ the angle between them (cf. Equation 1.8), it is simple to show that $\theta = \arctan(\Delta z_2/\Delta z_1)$, where z_1 and z_2 are the height modulation in domains D1 and D2, respectively. In the case shown here, with $\Delta z_1 = 0.04$ Å and $\Delta z_2 = 0.02$ Å, θ amounts to ≈27°.

A high peak (H) on one side of the domain boundary converts to a low peak (L) on the other side. This inversion is simulated in Figure 1.17c by a simple antiferromagnetic model configuration of spin moments and a tip spin at the angle $\theta = 27°$. The gray scale for each magnetic atom is proportional to $m_t m_s \cos\theta$ (white: $\theta = 0$; black: $\theta = \pi$). Clearly, the inversion occurs when the difference $\phi - \theta = \pi/2$, where θ and φ are the two different angles between tip and sample moments in domains D1 and D2, respectively.

The data can now be used to separate the magnetic and nonmagnetic components. Beginning with the SP-STM image shown in Figure 1.18a [98], the average height profile $z(x)$ where x is along [001] (Figure 1.18b, dark blue line) and also $z(x + c/2)$ (Figure 1.18b, light blue line) are plotted. Clearly, by taking the difference and sum of these two functions, the magnetic component with periodicity c and the nonmagnetic component with period $c/2$ can be extracted: $m_t m_s \cos[\theta(x)] \sim [z(x) - z(x + c/2)]/2$. This is further justified if it is assumed that the bulk magnetic symmetry is maintained at the surface. When using this procedure, the resulting magnetic profile for the data of Figure 1.18 has a period of c and a trapezoidal wave shape, as shown in Figure 1.18c (violet line). The nonmagnetic profile is also shown in

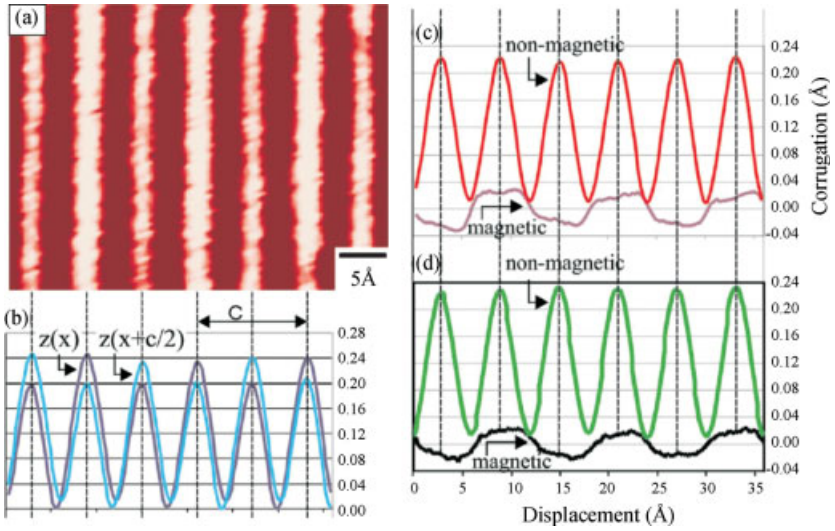


Figure 1.18 (a) SP-STM image acquired using a Mn-coated W tip; (b) Area-averaged line profile $z(x)$ of the whole image of (a) (dark blue), and $z(x + c/2)$ (light blue); (c) The resulting nonmagnetic component (red) and magnetic component (violet) for the Mn-coated tip; (d) Nonmagnetic (green) and magnetic (black) components for the Fe-coated tip on a similar sample region. (Reprinted with permission from Ref. [98]; copyright (2002), American Physical Society.)

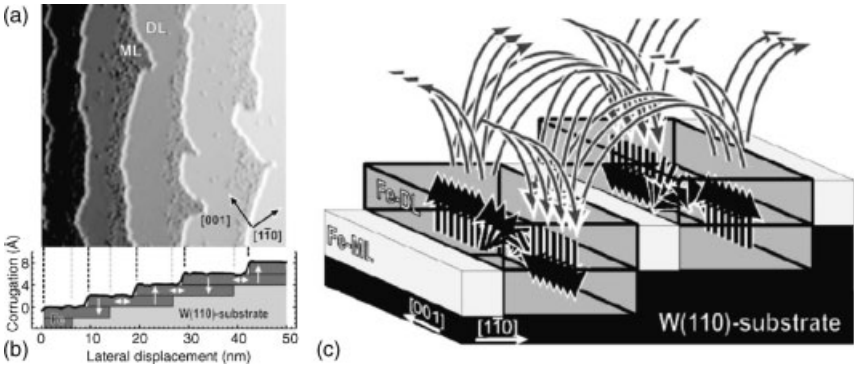


Figure 1.19 (a) Topographic STM image (scan range: 50 nm × 50 nm) of 1.6 ML Fe/W(110) after annealing to 450 K; (b) Line section measured at the bottom edge of the STM image. The local coverage alternates between one and two atomic layers. White arrows symbolize the easy magnetization directions of the mono- and the double layer; that is, in-plane and perpendicular to the surface, respectively; (c) Adjacent perpendicularly magnetized DL stripes exhibit an antiparallel dipolar coupling. Within domain walls the Fe DL on W(110) locally exhibits an in-plane magnetization. (Reprinted with permission from Ref. [104]; Elsevier.)

Figure 1.18c (red line) exhibiting a period of $c/2$ and a sinusoidal shape, the same as for the average line profile acquired with a nonmagnetic tip. The magnetic component amplitude is about 20% of the nonmagnetic component amplitude.

1.6 Magnetic Properties of Nanoscaled Wires

The behavior of perpendicularly magnetized Fe double layer nanowires epitaxially grown on a stepped W(110) single crystal [29, 55, 69, 100–103] with an average terrace width of about 9 nm is presented as an example of magnetic wires exhibiting a width in the nanometer range (see Figure 1.19a) [104]. This study was carried out at low temperatures, below about 10 K. At higher temperatures a reorientation to an in-plane easy axis occurs with the spin reorientation temperature being coverage-dependent for samples with a coverage between 1.5 and 2.2 atomic layers [105]. The sample was prepared by the deposition of 1.6 ML Fe on the W(110) substrate held at elevated temperature of 450 K. This preparation procedure leads to step-flow growth of the second Fe ML on top of the closed first Fe layer, thereby creating a system of nanowires with alternating Fe ML and DL coverage elongated along the step edges of the substrate (this situation is shown schematically in Figure 1.19b, which also contains the line section corresponding to the line shown in Figure 1.19a). The coverage range between 1.4 and 1.8 ML Fe/W(110) is characterized by magnetic saturation at relatively low external perpendicular fields combined with the absence of a hysteresis – that is, zero remanence. As shown schematically in Figure 1.19c, this antiparallel order is a consequence of the dipolar coupling which reduces the

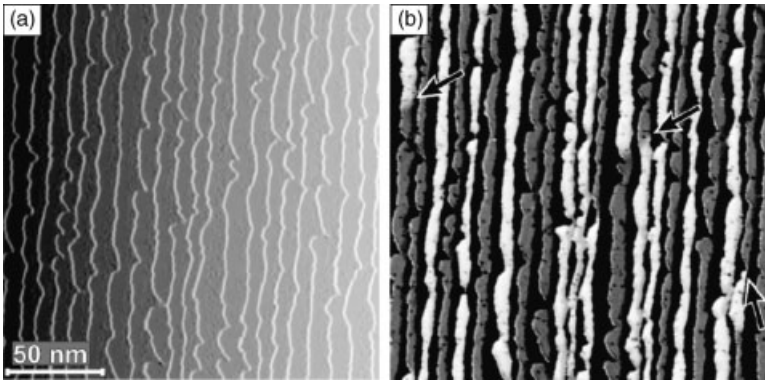


Figure 1.20 (a) STM topograph and (b) magnetic dI/dU image of Fe nanowires on W(110). Both images were measured simultaneously. The sample exhibits a demagnetized antiferromagnetic ground state which is energetically favorable due to flux closure between adjacent perpendicularly magnetized Fe nanowires [106]. (Reprinted with permission from Ref. [104]; Elsevier.)

magnetic stray field of the perpendicularly magnetized Fe DL. At domain walls the magnetization vector may locally be oriented along the hard magnetic axis – that is, in-plane.

Tunneling spectroscopy was used to image the corresponding magnetic domain structure. Since it is known from full dI/dU spectroscopy curves how the contrast must be interpreted (see Section 1.3.2), it is no longer necessary to measure the entire spectra at every pixel of the image as this is very time-consuming (about 10–20 h per image for the investigation discussed here [104]). Instead, the dI/dU signal at a fixed sample bias already gives a good contrast. Figure 1.20 shows the simultaneously recorded topography (panel a) and the dI/dU signal (panel b) of 1.5 ML Fe/W(110). The measurement time for this image was about 30 min. Due to its different electronic properties, the Fe ML appears dark, but this is not related to the magnetic properties. Instead, the ML is known to exhibit an in-plane magnetization [107] which cannot be detected using Gd-tips which are sensitive only to out-of-plane magnetization [29]. Clearly, the magnetic domain structure is dominated by DL nanowires which are magnetized alternately up and down, although exceptions from this model can easily be recognized. Several domain walls within single Fe nanowires are visible; some of these are marked with arrows in Figure 1.20b. There are also numerous adjacent nanowires which couple ferromagnetically rather than antiferromagnetically. It is likely that these DL nanowires approach very close to each other – or may even touch – so that the exchange coupling dominates.

Imaging by SP-STM can also be used to deduce macroscopic magnetic properties, a situation demonstrated by Pietzsch *et al.* [104, 108] for a system of Fe nanowires as just discussed above. These authors showed that spin-resolved dI/dU maps in a varying external field could be used to obtain the magnetic hysteresis curve of the

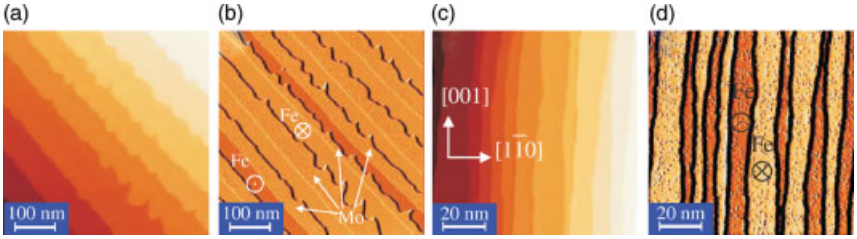


Figure 1.21 (a, c) Topographic STM images and (b, d) simultaneously measured differential conductance d*I*/d*U* maps of 0.5 ps-Fe (a, b) and of 1.0 ps-ML Fe (c, d) grown on Mo(110) at 700 K, respectively. Images (c, d) have been measured on the vicinal surface of the Mo substrate. Black and white lines in (b, d) located at step edges are artifacts due to scanning too quickly over the step edges. The black dots observed in (d) are due to adsorbates [24, 107]. Images and conductance maps were measured using a W/10 ML Au/4 ML Co magnetic tip. The Fe nanostructures reveal an out-of-plane magnetic contrast. (Reprinted with permission from Ref. [109]; copyright (2005), American Physical Society.)

surface area under investigation; that is, SP-STM enables the observation of magnetic hysteresis down to the nanometer scale.

Replacing W(110) with Mo(110) provides the unique possibility of observing the modification of magnetic properties of the Fe nanostructures, but leaving the structure and morphology almost unaffected [33, 109]. The magnetic easy axis is directed along the [001] direction for Fe/Mo(110) [110], while the easy axis is [1 $\bar{1}$0] for Fe/W(110) films [111]. The pseudomorphic ML (ps-ML) Fe/Mo(110) nanostructures are perpendicularly magnetized at low temperatures [112], whereas the ps-ML Fe/W (110) is magnetized in-plane along the [1 $\bar{1}$0] direction [113].

Figure 1.21a shows the topography and Figure 1.21b the simultaneously recorded d*I*/d*U* map of 0.5 ps-ML Fe deposited onto the Mo(110) single crystal at 700 K [109]. Monatomic Mo terraces decorated with the regular narrow Fe nanostripes grown by step-flow growth at the step edges are visible. The location of the Fe atoms on the Mo (110) surface is better visible on the d*I*/d*U* map (see Figure 1.21b) due to the element specific contrast resulting from the differences of the spin-averaged d*I*/d*U* signal which are connected with the local electronic surface properties that are different for Fe and Mo [112]. Uncovered Mo surfaces are indicated in Figure 1.21b by white arrows. The Fe nanowires show two different colors, representing two different values of the local d*I*/d*U* signal for equivalent surface regions (ML Fe/Mo(110)) for which the spin-averaged conductance signals should be the same [24]. All STM images and conductance maps shown in Figure 1.21 were measured using W tips covered by 10 ML Au and subsequently by 4 ML Co. It is known [114] that 4 ML Co prepared on W(110)/Au reveal an out-of-plane magnetic easy axis. Therefore, it may be expected that the magnetization of the tip would show perpendicular to the front plane of the tip – that is, along the tip axis – leading to an out-of-plane magnetic sensitivity of the tip. The contrast observed for the Fe nanostructures results from the perpendicularly magnetized Fe nanostripes, in agreement with Ref. [112].

The perpendicularly magnetized ML Fe nanostripes shown in Figure 1.21b are not antiferromagnetically ordered; that is, only two of the stripes are magnetized 'up',

whereas the orientation of the magnetization for the remaining stripes shows in the opposite direction ('down'). This means that the dipolar coupling between adjacent ML Fe nanowires is weak. The strength of the dipolar coupling between adjacent stripes increases with the stripe width and decreases with the distance between adjacent stripes [113]. The distance between adjacent ML Fe nanowires can be diminished down to a minimum by an increase of Fe coverage up to 1 ML. The topography of the 1 ML Fe deposited onto the vicinal surface of the Mo(110) crystal is presented in Figure 1.21c. Narrow ML Fe nanowires obtained on the vicinal surface are antiferromagnetically ordered, as demonstrated on the conductivity map (see Figure 1.21d).

1.7
Nanoscale Elements with Magnetic Vortex Structures

The domain structure of nanoscale magnetic elements has attracted considerable interest. The dependence of the domain structure on shape [115, 116], size [117] and edge structure [118] has been explored in many experiments. Further information on nanoislands gained by SP-STM are provided concerning Fe islands on W (110) [102, 119, 120], Fe islands on Mo(110) [112, 121], Co islands on Cu(111) [122, 123], Co islands on Pt(111) [124–126], Co islands on Au(11) [14], and FeAu alloy islands on Mo(110) [127, 128].

Due to their small dimensions, these elements are often superparamagnetic. This issue is addressed in Refs [112, 121], making use of time-dependent SP-STM studies. In a further development, Krause *et al.* showed [129] that superparamagnetic Fe nanoislands with typical sizes of 100 atoms could be addressed and locally switched using a magnetic probe tip. SP-STM thus provides an improved understanding of the underlying mechanism due to the feasibility to separate and quantify three fundamental contributions involved in magnetization switching, namely the current-induced spin torque, heating the island by the tunneling current, and Oerstedt field effects.

However, the influence of the thickness in conjunction with the magnetocrystalline anisotropy concerning nanoscale elements has rarely been studied [130]. Micromagnetic calculations have shown [131] that the lowest-energy domain configuration of permalloy rectangles depends critically on the thickness. With increasing thickness, a transition from the so-called C-state via the Landau-type or vortex configuration into a diamond state (double-vortex) was found. This behavior is caused by the thickness-dependent contribution of the magnetostatic energy, which must be paid wherever the magnetization is perpendicular to the element's rim. At a certain critical thickness it is energetically favorable to avoid the stray field by magnetizing the element along the edges throughout the entire particle, leading to a so-called *flux closure arrangement*.

This thickness-dependent behavior was corroborated experimentally by Bode *et al.* [120] using Fe islands on W(110). The left column of Figure 1.22 [120] shows the topography for different heights of the Fe islands, with the mean island height h

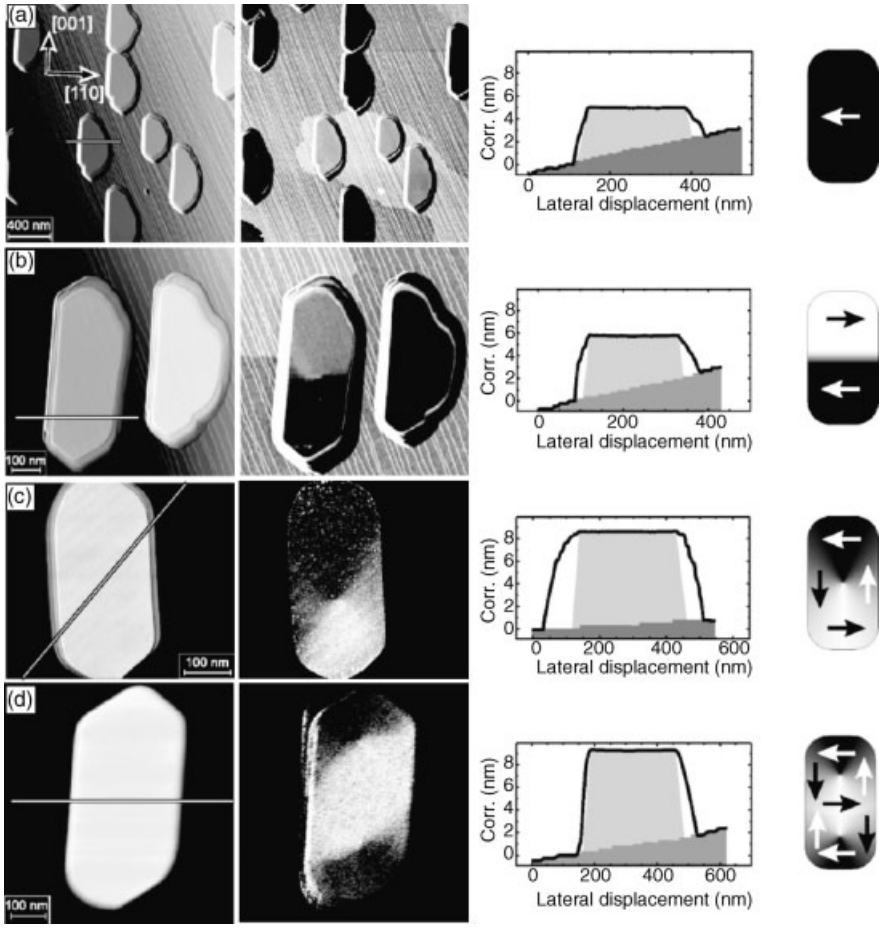


Figure 1.22 Topographic images (first column), spin-resolved dI/dU maps (second column) and topographic line sections (third column) of Fe islands on W(110) with different mean island heights h: (a) 53.5 nm, (b) 54.5 nm, (c) 57.5 nm, and (d) 58.5 nm. The data were obtained at $T = 14$ K. The resulting island domain configurations are schematically represented in the fourth column. (Reprinted with permission from Ref. [120]; copyright (2004), American Institute of Physics.)

varying between 3.5 and 8.5 nm. The lateral dimensions of the islands, irrespective of their height, are almost equal. The islands shown in Figure 1.22a exhibit an average height of approximately 3.5 nm (see line section). In the right-hand panel of Figure 1.22a the different magnetization states of islands can be distinguished by means of different dI/dU intensities. This variation results from spin-polarized tunneling between the magnetic STM tip (due to a coating with more than 100 ML Cr the tip is sensitive to the in-plane direction [28]) and the magnetic sample, and therefore represents different relative in-plane orientations of the magnetization in tip and sample. As no significant variation was found on top of the atomically flat

island surface, it would appear that these Fe islands are single domain, as shown schematically in the right-hand panel of Figure 1.22a. Evidently, there exists a close correlation between the magnetization direction of individual Fe islands and the surrounding Fe ML; dark (bright) Fe islands are always surrounded by a dark (bright) ML. With increasing thickness the magnetic pattern of the Fe islands becomes more and more complex such that, at $h = 54.5$ nm (see Figure 1.22b) a two-domain state is present. The island in Figure 1.22c exhibits a height h of 57.5 nm, while the corresponding spin-resolved dI/dU map shows the typical pattern of a single vortex state. A diamond state is found on the even higher island shown in Figure 1.22d ($h = 58.5$ nm).

Thus, the magnetic ground state is expected to be a vortex, as can be understood by the following consideration. If the dimensions of the particles are too small they do not form a single domain state, as this would require a relatively high stray field (or dipolar) energy. On the other hand, if the dimensions were too large, such domains would be formed such as those found in macroscopic pieces of magnetic material, because the additional cost of domain wall energy cannot be compensated by the reduction in stray field energy. By exhibiting a vortex configuration, the magnetization curls continuously around the particle center, drastically reducing the stray field energy and avoiding domain wall energy. Yet, the question arises as to the diameter of this core. An earlier investigation conducted by Shinjo *et al.* [132], using magnetic force microscopy, suggested an upper limit of about 50 nm caused by the intrinsic lateral resolution that was due to detection of the stray field. In the following section, it will be seen that an enhanced lateral resolution can be obtained using the technique of SP-STM, as shown by Wachowiak *et al.* [28, 102].

In order to gain a detailed insight into the magnetic behavior of the vortex core, a zoom into the central region was carried out where the rotation of the magnetization into the surface normal is expected. Maps of the dI/dU signal measured with different Cr-coated tips that are sensitive to the in-plane and out-of-plane component of the local sample magnetization are shown in Figure 1.23a and b, respectively [28]. The dI/dU signal as measured along a circular path around the vortex core (the circle in Figure 1.23a) exhibits a cosine-like modulation, indicating that the in-plane component of the local sample magnetization curls continuously around the vortex core. Figure 1.23b, which was measured with an out-of-plane sensitive tip on an identically prepared sample, exhibits a small bright area approximately in the center of the island. Therefore, the dI/dU map of Figure 1.23b confirms that the local magnetization in the vortex core is tilted normal to the surface (Figure 1.23c). The line section across the vortex core enables the width to be determined at about 9 nm, which is not restricted due to lateral resolution.

1.8 Individual Atoms on Magnetic Surfaces

The measurement of spin polarization states of individual atoms and understanding how atomic spins behave in a condensed matter environment, are essential steps

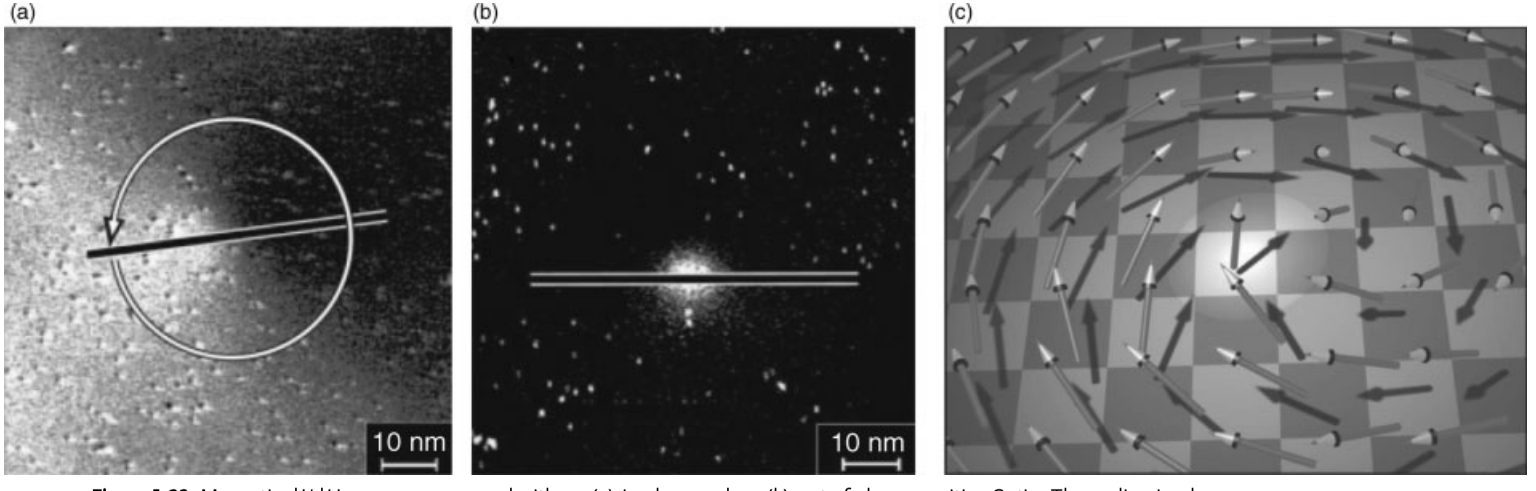


Figure 1.23 Magnetic dI/dU maps as measured with an (a) in-plane and an (b) out-of-plane sensitive Cr tip. The curling in-plane magnetization around the vortex core is recognizable in (a) and the perpendicular magnetization of the vortex core is visible as a bright area in (b); (c) Schematic arrangement of a magnetic vortex core. Far from the vortex core the magnetization curls continuously around the center with the orientation in the surface plane. In the center of the core the magnetization is perpendicular to the plane (highlighted). (Reprinted with permission from Ref. [28]; AAAS.)

towards the creation of devices, the functionality of which can be engineered at the level of individual atomic spins.

The direct observation of a spin polarization state of isolated adatoms remains challenging because isolated atoms have a low magnetic anisotropy energy that causes their spin to fluctuate over time due to environmental interactions. In the following section, measurements made by Yayon *et al.* [133] concerning the spin polarization state of *individual* Fe and Cr adatoms on a metal surface, are described. In order to fix the adatom spin in time, the adatoms were deposited onto ferromagnetic Co nanoislands, thereby coupling the adatom spin to the island magnetization through the direct exchange interaction.

Cobalt islands were chosen as a calibrated substrate where different magnetization states ('up' ↑ and 'down' ↓ with respect to the surface plane) are easily accessed [122]. The left-hand part of Figure 1.24 shows a representative topograph of Fe adatoms adsorbed onto triangular Co islands on the Cu(111) surface. The spatial oscillations seen on the Cu(111) surface are due to interference of surface state electrons scattered from the adatoms and Co islands [134].

In the right-hand part of Figure 1.24, panel (a) shows a color-scaled spin-polarized dI/dU map, together with topographic contour lines (measured simultaneously) for Fe and Cr atoms codeposited on two Co islands with opposite magnetization. The Fe and Cr atoms can easily be distinguished by their topographic signatures (Cr atoms

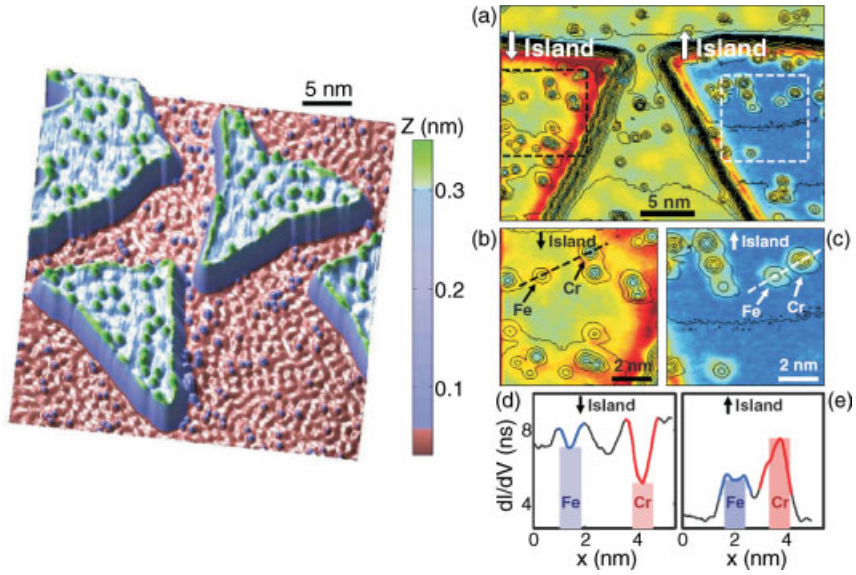


Figure 1.24 Left: Topograph of Fe adatoms adsorbed onto triangular Co islands on Cu(111) at $T = 4.8$ K. Fe adatoms are seen as green protrusions on the Co islands and blue protrusions on the bare Cu(111) surface. Right: (a) Spin-polarized dI/dU map of Fe and Cr adatoms on ↓ and ↑ Co islands on Cu(111); (b, c) Zoom-ins of areas marked by dashed lines on ↓ and ↑ islands in (a); (d, e) Line scans through the centers of Fe and Cr adatoms on ↓ and ↑ islands, respectively (marked by dashed lines in b and c). (Reprinted with permission from Ref. [133]; copyright (2007), American Physical Society.)

protrude 0.07 nm from the island surface, while Fe atoms protrude 0.04 nm). Spin-contrast between adatoms sitting on the two islands is seen as line cuts through Fe and Cr atoms (see Figure 1.24b–e). Fe atoms sitting on the ↓ island exhibit a larger dI/dU signal than those on the ↑ island, while Cr atoms on the ↓ island show a smaller dI/dU signal than those on the ↑ island. This confirms the parallel nature of the Fe/Co island spin coupling and the antiparallel nature of the Cr/Co island spin coupling over this energy range. SP-STS thus clearly reveals single adatom spin contrast: Each type of adatom yields a distinct spectrum, and over the energy range of the Co island surface-state Fe and Cr adatoms show opposite spin polarization. However, this measurement does not unambiguously determine the direction of the total spin of the adatom because the total spin is an integral over all filled states, whereas the spectra shown here were recorded over a finite energy range.

For a better understanding of the magnetic coupling between adatoms and islands, a density functional theory (DFT) calculation was also carried out [133]. The adsorption energies of Fe and Cr atoms on a ferromagnetic 2 ML film of Co on Cu(1 1 1) were calculated with the adatom moment fixed parallel and antiparallel to the magnetization of the Co film. The resulting values (see Figure 1.25) showed that Fe adatoms preferred a ferromagnetic alignment to the Co film, while Cr adatoms preferred an antiferromagnetic alignment. Comparison with the spin-polarized measurements implied that the Fe and Cr adatoms exhibit a negative spin polarization over the energy range of the Co island surface state.

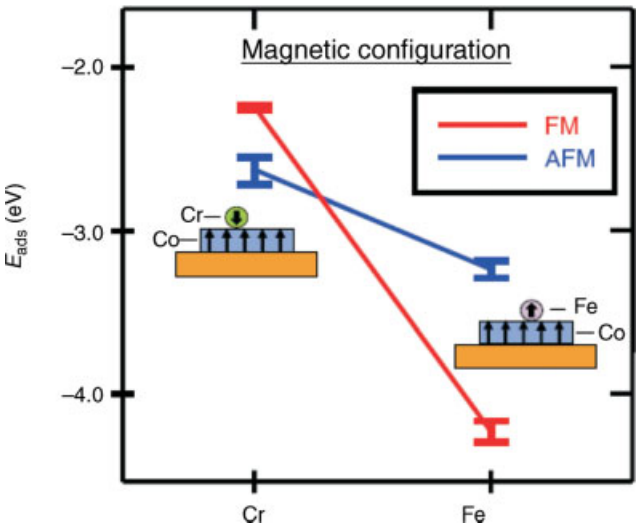


Figure 1.25 Calculated binding energies of ferromagnetic and antiferromagnetic configurations for Fe and Cr adatoms on a 2 ML high Co film on Cu(111). Error bars indicate the energy difference between hcp and fcc adatom adsorption sites. Cartoons depict the lowest-energy magnetic coupling configuration for Fe and Cr adatoms on the Co film. (Reprinted with permission from Ref. [133]; copyright (2007), American Physical Society.)

As discussed in Section 1.5.1, the Cr(001) surface exhibits a topological antiferromagnetic order. By increasing the number of adatoms, however, a small proportion of the Fe atoms on this surface will also exhibit an antiferromagnetic coupling to the underlying Cr(001) terraces [135, 136].

It is known from spin-polarized photoemission experiments that even nonmagnetic atoms such as oxygen (see Ref. [137] for O/Fe(110) and Ref. [138] for O/Co (0001)), sulfur [139] and iodine [140] become polarized upon chemisorption onto ferromagnetic surfaces. For each of these systems, SP-STM allows a deeper insight on the basis of its atomic resolution. For example, it was found that individual oxygen atoms on an Fe DL would induce highly anisotropic scattering states which were of minority spin character only [141]. This spin-dependent electron scattering at the single impurity level opens the possibility of understanding the origin of magnetoresistance phenomena on the atomic scale.

In the case of Fe islands, it has been reported that magnetic domains can be observed even after the deposition of a sulfur layer [142], and can act as a passivation species. These findings can be understood on the basis of the above discussion, also taking into account the fact that spin-polarized electrons from the interface with binding energies near the Fermi level are not fully damped but rather exhibit an attenuation length of at least several monolayers. Additionally, this mean free path is spin-dependent [143], such that an appropriate adsorbate layer may allow to extend the SP-STM to operate even under ambient conditions.

1.9 Domain Walls

The motion of domain walls is often hindered by lattice defects, leading to Barkhausen jumps in magnetic hysteresis curves. By using a high lateral resolution, the effective pinning of domain walls by screw-and-edge dislocations was first presented by Krause *et al.* [144] for Dy films on W(110).

Here, we will describe the details of two further aspects concerning domain walls which require an effective lateral resolution of SP-STM. First, we report on the behavior of domain walls in external magnetic fields, as investigated by Kubetzka *et al.* [103]; second, we will outline details of the widths of domain walls in nanoscale systems, referring to the studies conducted by Pratzer *et al.* [107].

The formation and stability of 360° domain walls plays a crucial role in the remagnetization processes of thin ferromagnetic films, with possible implications for the performance and development of magnetoresistive and magnetic random access memory (MRAM) devices. These are formed in external fields applied along the easy direction of the magnetic material when pairs of 180° walls with the same sense of rotation are forced together. Their stability against remagnetization into the uniform state is a manifestation of a hard axis anisotropy perpendicular to the rotational plane of the wall. This anisotropy may be of crystalline origin or – in films with an in-plane magnetic easy direction – due to the shape anisotropy.

Within Fe DL wires on W(110) that are separated by narrow regions with ML coverage (see Figure 1.26a [103]), two types of 180° walls can be distinguished using a ferromagnetic probe tip prepared by coating a W tip with several ML of Fe and therefore being sensitive to the in-plane magnetization component [27]. The wall width amounts to $w = 7$ nm. The intermediate $\mathrm{d}I/\mathrm{d}U$ signal (gray) corresponds to a perpendicular magnetization oriented either up or down. These two cases cannot be distinguished with a tip exhibiting a pure in-plane sensitivity unless the symmetry is broken by an external field. Figure 1.26 shows $\mathrm{d}I/\mathrm{d}U$ maps in an increasing perpendicular magnetic field of up to 800 mT. Areas magnetized parallel to the field direction grow at the expense of antiparallel ones, and pairs of 180° walls are forced together, which is equivalent to the formation of 360° walls. As expected, their lateral extension decreases with increasing field value. A closer inspection of these field-dependent measurements reveals that: (i) the magnetization rotates along every single nanowire with a defined chirality, and that (ii) the rotational sense is the same in each of the 12 wires within the imaged area. However, as the azimuthal angle of the tip magnetization is unknown, the absolute sense of rotation cannot be determined. For the same reason, it cannot be decided on the basis of these data alone whether the walls are of Bloch or Néel type, although the facts that the closed DL film is magnetized in-plane along $[1\,\bar{1}\,0]$ at elevated temperatures and the domain walls are oriented along the same direction, are indicative of their Bloch-type character. The first of the above observations is to be expected for stability reasons, as neighboring walls of opposite chirality (unwinding or untwisted walls [108]) will attract each other and can easily annihilate, in contrast to winding walls. As a consequence, the cooling process of the sample from above the Curie temperature to the measurement temperature of 14 K will result in a defined chirality within every individual wire, since such a structure is more stable against thermal fluctuations. The observed average distance between neighboring walls does not therefore necessarily reflect the magnetic ground state at low temperatures, as it might be a relic from the cooling process which is effectively frozen in a metastable state. The second of the observations will be discussed in Section 1.10.

With increasing external magnetic field, the tip's magnetization is successively rotated from in-plane towards the perpendicular direction. Its in-plane direction is also reversed during data acquisition at 400 mT (see Figure 1.26, black arrow), and this causes an inverted contrast for the remaining upper part of the image. At this field value a group of five 360° walls has formed a row (see ellipse), a correlation that might arise from their in-plane stray field. At 800 mT most of the 360° walls within the scanned area have been remagnetized by a rotation via the hard [001] in-plane direction.

Attention is now drawn to the Fe ML located in-between the Fe DL. Figure 1.27 [107] shows the topography (panel a) and the magnetic $\mathrm{d}I/\mathrm{d}U$ signal (panel b) of 1.25 ML Fe/W(110) grown at $T = 500$ K. Several domain walls separating dark and bright domains of the Fe ML can clearly be recognized in the overview of Figure 1.27b. Because of their different electronic properties, the DL stripes appear dark at this particular sample bias. At approximately the center of the white box in Figure 1.27b can be seen a bright spot; this is caused by a domain wall in this particular DL. The

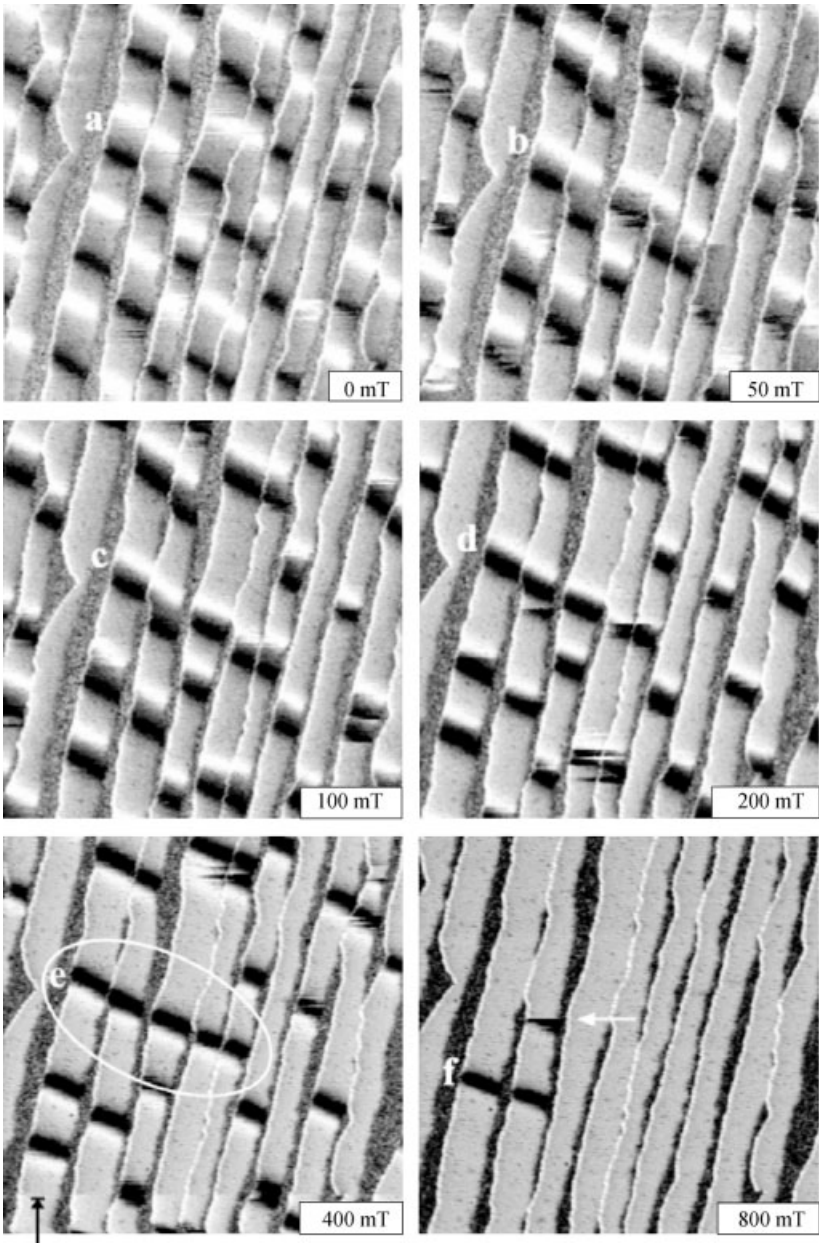


Figure 1.26 d*I*/d*U* maps of the surface area exhibiting Fe DL wires on W(110) imaged in an increasing perpendicular magnetic field. Pairs of 180° domain walls are gradually forced together, which is equivalent to the formation and compression of 360° walls. At 800 mT, most of these have vanished; that is, the Fe film is in magnetic saturation. With increasing external magnetic field the tip's magnetization is gradually forced from the in-plane towards the perpendicular direction. (Reprinted with permission from Ref. [103]; copyright (2003), American Physical Society.)

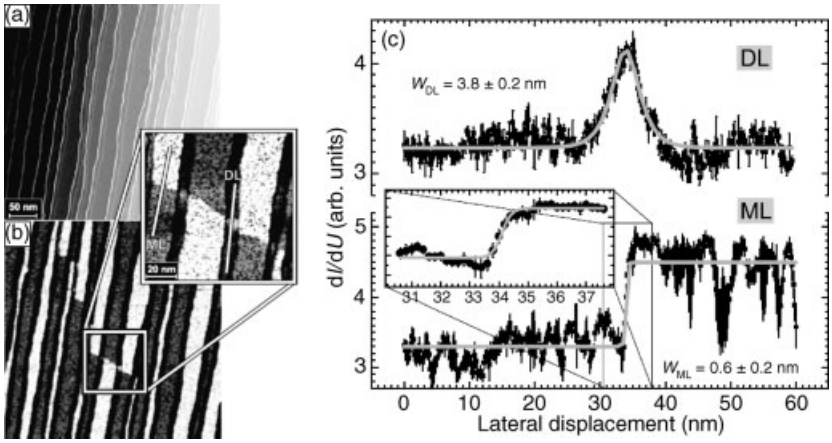


Figure 1.27 (a) Topographic and (b) spin-resolved dI/dU image showing the in-plane magnetic domain structure of 1.25 ML Fe/W (110). Several ML and DL domain walls can be recognized in the higher magnified inset; (c) Line sections showing domain wall profiles of the ML (bottom) and the DL (top). The inset shows that the ML domain wall width is on the atomic scale, with $w_{ML} = 6 \pm 2$ Å. (Reprinted with permission from Ref. [107]; copyright (2001), American Physical Society.)

inset of Figure 1.27b presents this location at higher magnification. Averaged line sections drawn along the white lines across domain walls in the ML and the DL are plotted in Figure 1.27c lower and upper, respectively, where clearly the ML domain wall is much narrower than the DL wall. The inset of Figure 1.27c shows the data in the vicinity of the ML domain wall in greater detail, and reveals a domain wall width $w < 1$ nm. In order to allow a more quantitative discussion the measured data have been fitted with a theoretical tanh function of a 180° wall profile [145]. This can be extended to an arbitrary angle between the magnetization axis of tip and sample ϕ [107, 146] by

$$\gamma(x) = \gamma_0 + \gamma_{sp} \cos\left\{\arccos\left[\tanh\left(\frac{x - x_0}{w/2}\right)\right] + \phi\right\} \tag{1.14}$$

where $\gamma(x)$ is the dI/dU signal measured at position x, x_0 is the position of the domain wall, w is the wall width, and γ_0 and γ_{sp} are the spin-averaged and spin-polarized dI/dU signals, respectively. Due to the Fe-coated tip which exhibits in-plane sensitivity, it is known that $\phi_{DL} = \pi/2$ and $\phi_{ML} = 0$. The best fit to the wall profile of the DL is achieved with $w_{DL} = 3.8 \pm 0.2$ nm. The profile of the ML domain wall is much narrower. If the fit procedure is performed over the full length of the line section $w_{ML} = 0.50 \pm 0.26$ nm, whereas $w_{ML} = 0.66 \pm 0.18$ nm is found if the fit is applied to the data in the inset of Figure 1.27c; this confirms the result of the analysis of the magnetization curves – that is, an almost atomically sharp domain wall.

A domain wall width of only six to eight atomic rows was also observed for an antiferromagnetic Fe monolayer on W(001) [147]. Such a narrow domain wall width can, in theory, be understood to arise from band structure effects, also taking into account intra-atomic noncollinear magnetism [148].

With regards to noncollinear effects it is important to distinguish between interatomic and intra-atomic magnetism. The first type is well known, and has been observed experimentally for small magnetic clusters [149] and in magnetic layers [150]; it has also been described on a theoretical basis [56, 151–157]. Interatomic magnetism can be understood within the Heisenberg model, taking into account *atomic* magnetic moments which are nonparallel for different atoms. Intra-atomic noncollinear effects arise from the tunneling current which flows through orbitals of the *same* atom, whilst that directly at the tip apex possesses a spin density orientation that is *noncollinear* [55].

1.10
Chiral Magnetic Order

Due to the inversion symmetry of bulk crystals, homochiral spin structures are unable to exist. However, as low-dimensional systems lack structural inversion symmetry, these single-handed spin arrangements may occur due to the Dzyaloshinskii–Moriya interaction [158, 159] arising from the spin-orbit scattering of electrons in an inversion asymmetric crystal field. This observation is now discussed with reference to the studies of Bode *et al.* [160], which were carried out in the same system Mn/W(110) for which atomic resolution of the magnetic properties had been demonstrated [96] (see Section 1.5.2).

In metallic itinerant magnets, spin-polarized electrons of the valence band hop across the lattice and exert the Heisenberg exchange interaction between magnetic spin moments $\vec{S}$ located on atomic sites i and j:

$$E_{exch} = \sum_{i,j} J_{ij}\, \vec{S}_i \cdot \vec{S}_j \tag{1.15}$$

As a consequence of a coulombic interaction, the exchange interaction is isotropic. Owing to the presence of a spin–orbit interaction, which connects the lattice with the spin symmetry, the broken parity of the lattice at an interface or surface gives rise to an additional interaction that breaks the inversion invariance of the Heisenberg Hamiltonian in Equation 1.15. This Dzyaloshinskii–Moriya interaction [158, 159]

$$E_{\mathrm{DM}} = \sum_{i,j} \vec{D}_{ij} \cdot (\vec{S}_i \times \vec{S}_j) \tag{1.16}$$

arises from the spin–orbit scattering of hopping electrons in an inversion asymmetric crystal field (where $\vec{D}_{ij}$ is the Dzyaloshinskii vector). In such an environment the scattering sequence of spin-polarized electrons, for example $i \rightarrow j \rightarrow i$ versus $j \rightarrow i \rightarrow j$, is noncommutative. The presence of this interaction has far-reaching consequences. Depending on the sign, the symmetry properties and the magnitude of $\vec{D}_{ij}$, uniaxial ferromagnetic or antiferromagnetic structures fail to exist and are instead replaced by a directional noncollinear magnetic structure of one specific chirality $\vec{C} = \vec{S}_i \times \vec{S}_{i+1}$, being either right-handed ($C>0$) or left-handed

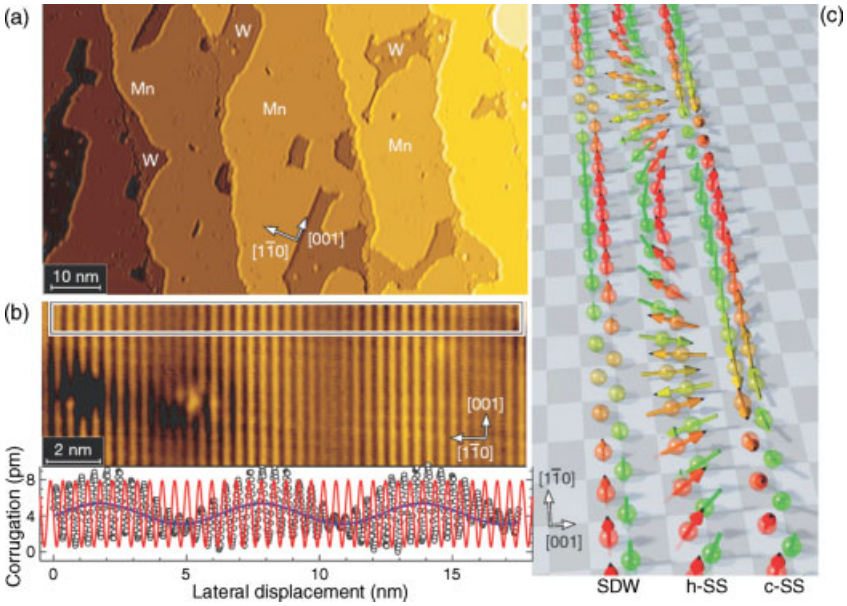


Figure 1.28 SP-STM of the Mn monolayer on W(110) and potential spin structures. (a) Topography of 0.77 atomic layers of Mn on W(110); (b) High-resolution constant-current image (upper panel) of the Mn monolayer taken with a Cr-coated tip. The stripes along the [001] direction are caused by spin-polarized tunneling between the magnetic tip and the sample. The averaged line section (lower panel) reveals a magnetic corrugation with a nominal periodicity of 0.448 nm and a long wavelength modulation. Comparison with a sine wave (red), expected for perfect antiferromagnetic order, reveals a phase shift of π between adjacent antinodes. In addition, there is an offset modulation (blue line) which is attributed to a varying electronic structure owing to spin–orbit coupling; (c) Artist's impression of the considered spin structures: a spin density wave (SDW), a helical spin spiral (h-SS) and a left-handed cycloidal spin spiral (c-SS). (Reprinted with permission from Ref. [160]; copyright (2007), MacMillan Publishers Ltd.)

($C<0$). In fact, J, D and also the anisotropy constants K, create a parameter space containing magnetic structures of unprecedented complexity [161], including 2-D and 3-D cycloidal, helicoidal or toroidal spin structures, or even vortices.

Figure 1.28a shows the topography of 0.77 atomic layers of Mn grown on a W(110) substrate [160]. The magnetic structure can be directly imaged with SP-STM using magnetically coated W tips. Figure 1.28b shows a high spatial resolution constant-current image measured on the atomically flat Mn layer using a Cr-coated probe tip which is sensitive to the in-plane magnetization [28]. The SP-STM data reveal periodic stripes running along the [001] direction, with an inter-stripe distance matching the surface lattice constant along the $[1\,\bar{1}\,0]$ direction (this was discussed earlier in Section 1.5.2). The additional important observation is, however, that the line section in the lower panel of Figure 1.28b, representing the magnetic amplitude, is not *constant* but is rather *modulated*, with a period of about 6 nm. Further, the magnetic corrugation is not simply a symmetric modulation superimposed on a

constant offset I_0. Instead, an additional long-wave modulation of I_0 (blue line) is present which is ascribed to spin–orbit coupling-induced variations of the spin-averaged electronic structure. When using in-plane-sensitive tips, the minima of the magnetic corrugation are found to coincide with the minima of the long-wave modulation of the spin-averaged local density of states. Within the field of view (see Figure 1.28b), three antinodes of the magnetic corrugation are visible. Comparing the experimental data with a sine function (red), representing perfect antiferromagnetic order, reveals a phase shift of π between adjacent antinodes.

The long wavelength modulation of the magnetic amplitude observed in Figure 1.28b may be explained by two fundamentally different spin structures: (i) a spin density wave (SDW) as it occurs, for example, in bulk Cr; or (ii) a spin spiral. Whereas, a SDW is characterized by a sinusoidal modulation of the size of the magnetic moments and the absence of spin rotation, the spin spiral consists of magnetic moments of approximately constant magnitude but whose directions rotate continuously. Spin spirals that are confined to a plane perpendicular or parallel to the propagation direction are denoted as either helical spirals (h-SS) or cycloidal spirals (c-SS). Figure 1.28c shows an artist's impression of a SDW, a h-SS and a c-SS. The magnetic contrast vanishes in either case twice over one magnetic period because: (i) the sample magnetic moments themselves vanish periodically; or (ii) the magnetic moments beneath the tip apex are orthogonal with respect to the tip magnetization m_t. The two cases can, however, be distinguished by addressing different components of the sample magnetization. Whereas, in case (i) a maximum spin contrast is always achieved at lateral positions where the magnetic moments are largest, and independent of m_t, in case (ii) a rotation of m_t can shift the position of maximum spin contrast.

Such a rotation of m_t can be achieved by subjecting an in-plane-sensitive Fe-coated tip to an appropriate external magnetic field (see sketches in Figure 1.29 [160]). For samples without a net magnetic moment, it is expected that the sample magnetization would remain unaffected. The SP-STM images and line sections of Figure 1.29 show data taken at a perpendicular field of 0 T (Figure 1.29a), 1 T (Figure 1.29b) and 2 T (Figure 1.29c). By using the encircled adsorbate as a marker, a maximum

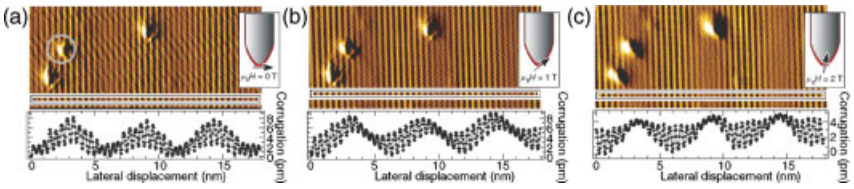


Figure 1.29 Field-dependent SP-STM measurements. Magnetically sensitive constant-current images of the Mn monolayer on W(110) (top panels) and corresponding line sections (bottom panels) taken with a ferromagnetic Fe-coated tip at external fields of (a) 0 T, (b) 1 T and (c) 2 T. As sketched in the insets, the external field rotates the tip magnetization from in-plane (a) to out-of-plane (c), shifting the position of maximum spin contrast. This proves that the Mn layer does not exhibit a spin density wave but rather a spin spiral rotating in a plane orthogonal to the surface. (Reprinted with permission from Ref. [160]; copyright (2007), MacMillan Publishers Ltd.)

magnetic contrast at this lateral position in zero field is observed, indicating large in-plane components of the sample magnetization here. This is also corroborated by the line section, which – in agreement with the in-plane-sensitive measurements of Figure 1.28b – shows a high magnetic corrugation at the maximum of the spin-averaged long-wave modulation. With an increasing external field the position of high magnetic corrugation shifts to the left (see Figure 1.29b) until a node reaches the adsorbate at 2 T (Figure 1.29c). The line sections reveal that the magnetic field shifts the position of high magnetic corrugation, but leaves the long-wave spin-averaged modulation unaffected. At 2 T – that is, with an almost perfectly out-of-plane magnetized tip – a maximum magnetic contrast is achieved and the spin-averaged signal exhibits a minimum (see line section of Figure 1.29c). Although this observation rules out a SDW, it provides clear proof of a spin spiral with magnetic moments rotating from in-plane (imaged in Figure 1.29a) to out-of-plane (imaged in Figure 1.29c).

The islands exhibit a spin spiral of only one chirality, as would be expected for a Dzyaloshinskii–Moriya interaction-driven magnetic configuration. The azimuthal orientation of the tip magnetization, however, cannot be reliably controlled, and consequently it is not possible to test experimentally whether the observed spin spiral is helical or cycloidal.

References

1 Binnig, G. and Rohrer, H. (1982) *Helvetica Physica Acta*, **55**, 726.

2 Binnig, G. and Rohrer, H. (1987) *Reviews of Modern Physics*, **59**, 615.

3 Pierce, D.T. (1988) *Physica Scripta*, **38**, 291.

4 Minakov, A.A. (1990) *Surface Science*, **236**, L377.

5 Allenspach, R. and Bischof, A. (1988) *Applied Physics Letters*, **54**, 587.

6 Wiesendanger, R., Güntherodt, H.-J., Güntherodt, G., Gambino, R.J. and Ruf, R. (1990) *Physical Review Letters*, **65**, 247.

7 Getzlaff, M. (2007) *Fundamentals of Magnetism*, Springer, Berlin.

8 Jullière, M. (1975) *Physics Letters A*, **54**, 225.

9 Miyazaki, T. and Tezuka, N. (1995) *Journal of Magnetism and Magnetic Materials*, **139**, L231.

10 Ding, H.F., Wulfhekel, W., Henk, J., Bruno, P. and Kirschner, J. (2003) *Physical Review Letters*, **90**, 116603.

11 Wiesendanger, R., Bürgler, D., Tarrach, G., Schaub, T., Hartmann, U., Güntherodt, H.-J., Shvets, I.V. and Coey, J.M.D. (1991) *Applied Physics A: Materials Science and Processing*, **53**, 349.

12 Wiesendanger, R., Bürgler, D., Tarrach, G., Wadas, A., Brodbeck, D., Güntherodt, H.-J., Güntherodt, G., Gambino, R.J. and Ruf, R. (1991) *Journal of Vacuum Science and Technology B*, **9**, 519.

13 Wiesendanger, R., Bode, M., Kleiber, M., Löhndorf, M., Pascal, R., Wadas, A. and Weiss, D. (1997) *Journal of Vacuum Science and Technology B*, **15**, 1330.

14 Rastei, M.V. and Bucher, J.P. (2006) *Journal of Physics: Condensed Matter*, **18**, L619.

15 Wadas, A. and Hug, H.J. (1992) *Journal of Applied Physics*, **72**, 203.

16 Kubetzka, A., Bode, M. and Wiesendanger, R. (2007) *Applied Physics Letters*, **91**, 012508.

17 Wulfhekel, W. and Kirschner, J. (1999) *Applied Physics Letters*, **75**, 1944.

18 Wulfhekel, W., Ding, H.F. and Kirschner, J. (2000) *Journal of Applied Physics*, **87**, 6475.

19 Wulfhekel, W., Ding, H.F., Lutzke, W., Steierl, G., Vázquez, M., Marin, P., Hernando, A. and Kirschner, J. (2001) *Applied Physics A: Materials Science and Processing*, **72**, 463.

20 Ding, H.F., Wulfhekel, W., Chen, C., Barthel, J. and Kirschner, J. (2001) *Materials Science and Engineering B*, **84**, 96.

21 Ding, H.F., Wulfhekel, W. and Kirschner, J. (2002) *Europhysics Letters*, **57**, 100.

22 Wulfhekel, W., Hertel, R., Ding, H.F., Steierl, G. and Kirschner, J. (2002) *Journal of Magnetism and Magnetic Materials*, **249**, 368.

23 Ding, H.F., Wulfhekel, W., Schlickum, U. and Kirschner, J. (2003) *Europhysics Letters*, **63**, 419.

24 Bode, M. (2003) *Reports on Progress in Physics*, **66**, 523.

25 Schlickum, U., Wulfhekel, W. and Kirschner, J. (2003) *Applied Physics Letters*, **83**, 2016.

26 Schlickum, U., Janke–Gilman, N., Wulfhekel, W. and Kirschner, J. (2004) *Physical Review Letters*, **92**, 107203.

27 Bode, M., Getzlaff, M. and Wiesendanger, R. (1998) *Physical Review Letters*, **81**, 4256.

28 Wachowiak, A., Wiebe, J., Bode, M., Pietzsch, O., Morgenstern, M. and Wiesendanger, R. (2002) *Science*, **298**, 577.

29 Pietzsch, O., Kubetzka, A., Bode, M. and Wiesendanger, R. (2000) *Physical Review Letters*, **84**, 5212.

30 Kubetzka, A., Bode, M., Pietzsch, O. and Wiesendanger, R. (2002) *Physical Review Letters*, **88**, 057201.

31 Allenspach, R., Stampanoni, M. and Bischof, A. (1990) *Physical Review Letters*, **65**, 3344.

32 Pütter, S., Ding, H.F., Millev, Y.T., Oepen, H.P. and Kirschner, J. (2001) *Physical Review B - Condensed Matter*, **64**, 092409.

33 Prokop, J., Kukunin, A. and Elmers, H.J. (2006) *Physical Review B - Condensed Matter*, **73**, 014428.

34 Kubetzka, A., Ferriani, P., Bode, M., Heinze, S., Bihlmayer, G., von Bergmann, K., Pietzsch, O., Blügel, S. and Wiesendanger, R. (2005) *Physical Review Letters*, **94**, 087204.

35 Elmers, H.J. and Gradmann, U. (1990) *Applied Physics A: Materials Science and Processing*, **51**, 255.

36 Bode, M., Getzlaff, M. and Wiesendanger, R. (1999) *Journal of Vacuum Science and Technology A - Vacuum Surfaces and Films*, **17**, 2228.

37 Murphy, S., Osing, J. and Shvets, I.V. (2003) *Surface Science*, **547**, 139.

38 Murphy, S., Osing, J. and Shvets, I.V. (1999) *Applied Surface Science*, **144–145**, 497.

39 Ceballos, S.F., Mariotto, G., Murphy, S. and Shvets, I.V. (2003) *Surface Science*, **523**, 131.

40 Murphy, S., Ceballos, S.F., Mariotto, G., Berdunov, N., Jordan, K., Shvets, I.V. and Mukovskii, Y.M. (2005) *Microscopy Research and Technique*, **66**, 85.

41 Pierce, D.T., Celotta, R.J., Wang, G.C., Unertl, W.N., Galejs, A., Kuyatt, C.E. and Mielczarek, S. (1980) *Review of Scientific Instruments*, **51**, 478.

42 Prins, M.W.J. and Abraham, D.L. (1993) H. van Kempen. *Journal of Magnetism and Magnetic Materials*, **121**, 152.

43 Prins, M.W.J., van Kempen, H., van Leuken, H., de Groot, R.A., van Roy, W. and de Boeck, J. (1995) *Journal of Physics - Condensed Matter*, **7**, 9447.

44 Jansen, R., Prins, M.W.J. and van Kempen, H. (1998) *Physical Review B - Condensed Matter*, **57**, 4033.

45 Mukasa, K., Sueoka, K., Hasegawa, H., Tazuke, Y. and Hayakawa, K. (1995) *Materials Science and Engineering B*, **31**, 69.

46 Prins, M.W.J., Jansen, R. and van Kempen, H. (1996) *Physical Review B - Condensed Matter*, **53**, 8105.

47 Suzuki, Y., Nabhan, W. and Tanaka, K. (1997) *Applied Physics Letters*, **71**, 3153.

48 Nabhan, W., Suzuki, Y., Shinohara, R., Yamaguchi, K. and Tamura, E. (1999) *Applied Surface Science*, **144–145**, 570.

49 Jansen, R., Schad, R. and van Kempen, H. (1999) *Journal of Magnetism and Magnetic Materials*, **198–199**, 668.

50 Alvarado, S.F. and Renaud, P. (1992) *Physical Review Letters*, **68**, 1387.

51 LaBella, V.P., Bullock, D.W., Ding, Z., Emery, C., Venkatesan, A., Oliver, W.F., Salamo, G.J., Thibado, P.M. and Mortazavi, M. (2001) *Science*, **292**, 1518.

52 Bode, M., Heinze, S., Kubetzka, A., Pietzsch, O., Nie, X., Bihlmayer, G., Blügel, S. and Wiesendanger, R. (2002) *Physical Review Letters*, **89**, 237205.

53 Bode, M., Kubetzka, A., Heinze, S., Pietzsch, O., Wiesendanger, R., Heide, M., Nie, X., Bihlmayer, G. and Blügel, S. (2003) *Journal of Physics: Condensed Matter*, **15**, S679.

54 Pietzsch, O., Kubetzka, A., Bode, M. and Wiesendanger, R. (2004) *Applied Physics A: Materials Science and Processing*, **78**, 781.

55 Bode, M., Pietzsch, O., Kubetzka, A., Heinze, S. and Wiesendanger, R. (2001) *Physical Review Letters*, **86**, 2142.

56 Wortmann, D., Heinze, S., Kurz, P., Bihlmayer, G. and Blügel, S. (2001) *Physical Review Letters*, **86**, 4132.

57 Hofer, W.A. and Fisher, A.J. (2002) *Surface Science*, **498**, L65.

58 Hofer, W.A. and Fisher, A.J. (2003) *Journal of Magnetism and Magnetic Materials*, **267**, 139.

59 Kubetzka, A., Pietzsch, O., Bode, M. and Wiesendanger, R. (2003) *Applied Physics A: Materials Science and Processing*, **76**, 873.

60 Wiesendanger, R., Bode, M. and Getzlaff, M. (1999) *Applied Physics Letters*, **75**, 124.

61 Yamada, T.K., Vázquez de Parga, A.L., Bischoff, M.M.J., Mizoguchi, T. and van Kempen, H. (2005) *Microscopy Research and Technique*, **66**, 93.

62 Balashov, T., Takács, A.F., Wulfhekel, W. and Kirschner, J. (2006) *Physical Review Letters*, **97**, 187201.

63 Johnson, M. and Clarke, J. (1990) *Journal of Applied Physics*, **67**, 6141.

64 Tedrow, P.M. and Meservey, R. (1973) *Physical Review B - Condensed Matter*, **7**, 318.

65 Meservey, R. and Tedrow, P.M. (1994) *Physics Reports - Review Section of Physics Letters*, **238**, 173.

66 Bode, M., Getzlaff, M., Heinze, S., Pascal, R. and Wiesendanger, R. (1998) *Applied Physics A: Materials Science and Processing*, **66**, S121.

67 Getzlaff, M., Bode, M., Heinze, S., Pascal, R. and Wiesendanger, R. (1998) *Journal of Magnetism and Magnetic Materials*, **184**, 155.

68 Getzlaff, M., Bode, M. and Wiesendanger, R. (1999) *Surface Review and Letters*, **6**, 591.

69 Bode, M., Pietzsch, O., Kubetzka, A. and Wiesendanger, R. (1055) *Journal of Electron Spectroscopy and Related Phenomena*, **2001**, 114–16.

70 Wiesendanger, R. and Bode, M. (2001) *Solid State Communications*, **119**, 341.

71 Bode, M., Getzlaff, M., Kubetzka, A., Pascal, R., Pietzsch, O. and Wiesendanger, R. (1999) *Physical Review Letters*, **83**, 3017.

72 Berbil-Bautista, L., Krause, S., Bode, M. and Wiesendanger, R. (2007) *Physical Review B - Condensed Matter*, **76**, 064411.

73 Stroscio, J.A., Pierce, D.T., Davies, A., Celotta, R.J. and Weinert, M. (1995) *Physical Review Letters*, **75**, 2960.

74 Farle, M. and Lewis, W.A. (1994) *Journal of Applied Physics*, **75**, 5604.

75 Bode, M., von Bergmann, K., Pietzsch, O., Kubetzka, A. and Wiesendanger, R. (2006) *Journal of Magnetism and Magnetic Materials*, **304**, 1.

76 von Bergmann, K., Bode, M. and Wiesendanger, R. (2004) *Physical Review B - Condensed Matter*, **70**, 174455.

77 von Bergmann, K., Bode, M., Kubetzka, A., Pietzsch, O. and Wiesendanger, R. (2005) *Microscopy Research and Technique*, **66**, 61.

78 Yamada, T.K., Bischoff, M.M.J., Heijnen, G.M.M., Mizoguchi, T. and

van Kempen, H. (2003) *Physical Review Letters*, **90**, 056803.

79 Okuno, S.N., Kishi, T. and Tanaka, K. (2002) *Physical Review Letters*, **88**, 066803.

80 Wiesendanger, R., Shvets, I.V., Bürgler, D., Tarrach, G., Güntherodt, H.-J., Coey, J.M.D. and Gräser, S. (1992) *Science*, **255**, 583.

81 Wiesendanger, R., Shvets, I.V., Bürgler, D., Tarrach, G., Güntherodt, H.-J. and Coey, J.M.D. (1992) *Zeitschrift für Physik D*, **86**, 1.

82 Wiesendanger, R., Shvets, I.V., Bürgler, D., Tarrach, G., Güntherodt, H.-J. and Coey, J.M.D. (1992) *Europhysics Letters*, **19**, 141.

83 Shvets, I.V., Wiesendanger, R., Bürgler, D., Tarrach, G., Güntherodt, H.-J. and Coey, J.M.D. (1992) *Journal of Applied Physics*, **71**, 5489.

84 Berdunov, N., Murphy, S., Mariotto, G. and Shvets, I.V. (2004) *Physical Review Letters*, **93**, 057201.

85 Blügel, S., Pescia, D. and Dederichs, P.H. (1989) *Physical Review B - Condensed Matter*, **39**, 1392.

86 Kleiber, M., Bode, M., Ravlić, R. and Wiesendanger, R. (2000) *Physical Review Letters*, **85**, 4606.

87 Kleiber, M., Bode, M., Ravlić, R., Tezuka, N. and Wiesendanger, R. (2002) *Journal of Magnetism and Magnetic Materials*, **240**, 64.

88 Ravlić, R., Bode, M., Kubetzka, A. and Wiesendanger, R. (2003) *Physical Review B - Condensed Matter*, **67**, 174411.

89 Kawagoe, T., Suzuki, Y., Bode, M. and Koike, K. (2003) *Journal of Applied Physics*, **93**, 6575.

90 Kawagoe, T., Iguchi, Y., Miyamachi, T., Yamasaki, A. and Suga, S. (2005) *Physical Review Letters*, **95**, 207205.

91 Hänke, T., Krause, S., Berbil-Bautista, L., Bode, M., Wiesendanger, R., Wagner, V., Lott, D. and Schreyer, A. (2005) *Physical Review B - Condensed Matter*, **71**, 184407.

92 Kawagoe, T., Iguchi, Y., Yamasaki, A., Suzuki, Y., Koike, K. and Suga, S. (2005) *Physical Review B - Condensed Matter*, **71**, 014427.

93 Kawagoe, T., Iguchi, Y. and Suga, S. (2007) *Journal of Magnetism and Magnetic Materials*, **310**, 2201.

94 Dreyer, M., Lee, J., Krafft, C. and Gomez, R. (2005) *Journal of Applied Physics*, **97**, 10E703.

95 Bode, M., Hennefarth, M., Haude, D., Getzlaff, M. and Wiesendanger, R. (1999) *Surface Science*, **432**, 8.

96 Heinze, S., Bode, M., Kubetzka, A., Pietzsch, O., Nie, X., Blügel, S. and Wiesendanger, R. (2000) *Science*, **288**, 1805.

97 Bode, M., Heinze, S., Kubetzka, A., Pietzsch, O., Hennefarth, M., Getzlaff, M., Wiesendanger, R., Nie, X., Bihlmayer, G. and Blügel, S. (2002) *Physical Review B - Condensed Matter*, **66**, 014425.

98 Yang, H., Smith, A.R., Prikhodko, M. and Lambrecht, W.R.L. (2002) *Physical Review Letters*, **89**, 226101.

99 Smith, A.R., Yang, R., Yang, H., Dick, A., Neugebauer, J. and Lambrecht, W.R.L. (2005) *Microscopy Research and Technique*, **66**, 72.

100 Bode, M., Kubetzka, A., Pietzsch, O. and Wiesendanger, R. (2001) *Applied Physics A: Materials Science and Processing*, **72**, S149.

101 Vedmedenko, E.Y., Kubetzka, A., von Bergmann, K., Pietzsch, O., Bode, M., Kirschner, J., Oepen, H.P. and Wiesendanger, R. (2004) *Physical Review Letters*, **92**, 077207.

102 Wiesendanger, R., Bode, M., Kubetzka, A., Pietzsch, O., Morgenstern, M., Wachowiak, A. and Wiebe, J. (2004) *Journal of Magnetism and Magnetic Materials*, **272–276**, 2115.

103 Kubetzka, A., Pietzsch, O., Bode, M. and Wiesendanger, R. (2003) *Physical Review B - Condensed Matter*, **67**, 020401.

104 Bode, M., Kubetzka, A., Pietzsch, O. and Wiesendanger, R. (2002) *Surface Science*, **514**, 135.

105 von Bergmann, K., Bode, M. and Wiesendanger, R. (2006) *Journal of Magnetism and Magnetic Materials*, **305**, 279.

106 Hauschild, J., Gradmann, U. and Elmers, H.J. (1998) *Applied Physics Letters*, **72**, 3211.

107 Pratzer, M., Elmers, H.J., Bode, M., Pietzsch, O., Kubetzka, A. and Wiesendanger, R. (2001) *Physical Review Letters*, **87**, 127201.

108 Pietzsch, O., Kubetzka, A., Bode, M. and Wiesendanger, R. (2001) *Science*, **292**, 2053.

109 Prokop, J., Kukunin, A. and Elmers, H.J. (2005) *Physical Review Letters*, **95**, 187202.

110 Usov, V., Murphy, S. and Shvets, I.V. (2004) *Journal of Magnetism and Magnetic Materials*, **283**, 357.

111 Gradmann, U., Korecki, J. and Waller, G. (1986) *Applied Physics A: Materials Science and Processing*, **39**, 101.

112 Bode, M., Pietzsch, O., Kubetzka, A. and Wiesendanger, R. (2004) *Physical Review Letters*, **92**, 067201.

113 Elmers, H.J., Hauschild, J. and Gradmann, U. (1999) *Physical Review B - Condensed Matter*, **59**, 3688.

114 Allenspach, R., Stampanoni, M. and Bischof, A. (1990) *Physical Review Letters*, **65**, 3344.

115 Yi, G., Aitchison, P.R., Doyle, W.D., Chapman, J.N. and Wilkinson, C.D.W. (2002) *Journal of Applied Physics*, **92**, 6087.

116 Kirk, K.J., McVitie, S., Chapman, J.N. and Wilkinson, C.D.W. (2001) *Journal of Applied Physics*, **89**, 7174.

117 Gomez, R.D., Luu, T.V., Pak, A.O., Kirk, K.J. and Chapman, J.N. (1999) *Journal of Applied Physics*, **85**, 6163.

118 Herrmann, M., McVitie, S. and Chapman, J.N. (2000) *Journal of Applied Physics*, **87**, 2994.

119 Kubetzka, A., Pietzsch, O., Bode, M. and Wiesendanger, R. (2001) *Physical Review B - Condensed Matter*, **63**, 140407.

120 Bode, M., Wachowiak, A., Wiebe, J., Kubetzka, A., Morgenstern, M. and Wiesendanger, R. (2004) *Applied Physics Letters*, **84**, 948.

121 Bode, M., Kubetzka, A., von Bergmann, K., Pietzsch, O. and Wiesendanger, R. (2005) *Microscopy Research and Technique*, **66**, 117.

122 Pietzsch, O., Kubetzka, A., Bode, M. and Wiesendanger, R. (2004) *Physical Review Letters*, **92**, 057202.

123 Pietzsch, O., Okatov, S., Kubetzka, A., Bode, M., Heinze, S., Lichtenstein, A. and Wiesendanger, R. (2006) *Physical Review Letters*, **96**, 237203.

124 Rusponi, S., Weiss, N., Chen, T., Epple, M. and Brune, H. (2005) *Applied Physics Letters*, **87**, 162514.

125 Meier, F., von Bergmann, K., Ferriani, P., Wiebe, J., Bode, M., Hashimoto, K., Heinze, S. and Wiesendanger, R. (2006) *Physical Review B - Condensed Matter*, **74**, 195411.

126 Meier, F., von Bergmann, K., Wiebe, J., Bode, M. and Wiesendanger, R. (2007) *Journal of Physics D*, **40**, 1306.

127 Kukunin, A., Prokop, J. and Elmers, H.J. (2006) *Acta Physica Polonica A*, **109**, 371.

128 Prokop, J., Kukunin, A. and Elmers, H.J. (2007) *Physical Review B - Condensed Matter*, **75**, 144423.

129 Krause, S., Berbil-Bautista, L., Herzog, G., Bode, M. and Wiesendanger, R. (2007) *Science*, **317**, 1537.

130 Evoy, S., Carr, D.W., Sekaric, L., Suzuki, Y., Parpia, J.M. and Craighead, H.G. (2000) *Journal of Applied Physics*, **87**, 404.

131 Hertel, R. (2002) *Zeitschrift Fur Metallkunde*, **93**, 957.

132 Shinjo, T., Okuno, T., Hassdorf, R., Shigeto, K. and Ono, T. (2000) *Science*, **289**, 930.

133 Yayon, Y., Brar, V.W., Senapati, L., Erwin, S.C. and Crommie, M.F. (2007) *Physical Review Letters*, **99**, 067202.

134 Crommie, M.F., Lutz, C.P. and Eigler, D.M. (1993) *Nature*, **363**, 524.

135 Ravlić, R., Bode, M. and Wiesendanger, R. (2003) *Journal of Physics: Condensed Matter*, **15**, S2513.

136 Bode, M., Ravlić, R., Kleiber, M. and Wiesendanger, R. (2005) *Applied Physics A: Materials Science and Processing*, **80**, 907.

137 Getzlaff, M., Bansmann, J. and Schönhense, G. (1999) *Journal of Magnetism and Magnetic Materials*, **192**, 458.

138 Getzlaff, M., Bansmann, J. and Schönhense, G. (1996) *Journal of Electron Spectroscopy and Related Phenomena*, **77**, 197.

139 Getzlaff, M., Westphal, C., Bansmann, J. and Schönhense, G. (2000) *Journal of Electron Spectroscopy and Related Phenomena*, **107**, 293.

140 Getzlaff, M., Bansmann, J. and Schönhense, G. (1993) *Physics Letters A*, **182**, 153.

141 von Bergmann, K., Bode, M., Kubetzka, A., Heide, M., Blügel, S. and Wiesendanger, R. (2004) *Physical Review Letters*, **92**, 046801.

142 Berbil-Bautista, L., Krause, S., Hänke, T., Bode, M. and Wiesendanger, R. (2006) *Surface Science*, **600**, L20.

143 Getzlaff, M., Bansmann, J. and Schönhense, G. (1993) *Solid State Communications*, **87**, 467.

144 Krause, S., Berbil-Bautista, L., Hänke, T., Vonau, F., Bode, M. and Wiesendanger, R. (2006) *Europhysics Letters*, **76**, 637.

145 Hubert, A. and Schäfer, R. (1998) *Magnetic Domains*, Springer, Berlin.

146 Kubetzka, A., Pietzsch, O., Bode, M., Ravlić, R. and Wiesendanger, R. (2003) *Acta Physica Polonica B*, **104**, 259.

147 Bode, M., Vedmedenko, E.Y., von Bergmann, K., Kubetzka, A., Ferriani, P., Heinze, S. and Wiesendanger, R. (2006) *Nature Materials*, **5**, 477.

148 Nakamura, K., Takeda, Y., Akiyama, T., Ito, T. and Freeman, A.J. (2004) *Physical Review Letters*, **93**, 057202.

149 Douglass, D.C., Cox, A.J., Bucher, J.P. and Bloomfield, L.A. (1993) *Physical Review B - Condensed Matter*, **47**, 12874.

150 von Bergmann, K., Heinze, S., Bode, M., Vedmedenko, E.Y., Bihlmayer, G., Blügel, S. and Wiesendanger, R. (2006) *Physical Review Letters*, **96**, 167203.

151 Pappas, D.P., Popov, A.P., Anisimov, A.N., Reddy, B.V. and Khamna, S.N. (1996) *Physical Review Letters*, **76**, 4332.

152 Hobbs, D., Kresse, G. and Hafner, J. (2000) *Physical Review B - Condensed Matter*, **62**, 11556.

153 Kurz, Ph., Bihlmayer, G., Hirai, K. and Blügel, S. (2001) *Physical Review Letters*, **86**, 1106.

154 Heinze, S., Kurz, P., Wortmann, D., Bihlmayer, G. and Blügel, S. (2002) *Applied Physics A: Materials Science and Processing*, **75**, 25.

155 Kurz, Ph., Förster, F., Nordström, L., Bihlmayer, G. and Blügel, S. (2004) *Physical Review B - Condensed Matter*, **69**, 024415.

156 Lizárraga, R., Nordström, L., Bergqvist, L., Bergman, A., Sjöstedt, E., Mohn, P. and Eriksson, O. (2004) *Physical Review Letters*, **93**, 107205.

157 Heinze, S. (2006) *Applied Physics A: Materials Science and Processing*, **85**, 407.

158 Dzialoshinskii, I.E. (1957) *Soviet Physics JETP*, **5**, 1259.

159 Moriya, T. (1960) *Physical Review*, **120**, 91.

160 Bode, M., Heide, M., von Bergmann, K., Ferriani, P., Heinze, S., Bihlmayer, G., Kubetzka, A., Pietzsch, O., Blügel, S. and Wiesendanger, R. (2007) *Nature*, **447**, 190.

161 Rößler, U.K., Bogdanov, A.N. and Pfleiderer, C. (2006) *Nature*, **442**, 797.

2
Nanoscale Imaging and Force Analysis with Atomic Force Microscopy

Hendrik Hölscher, André Schirmeisen, and Harald Fuchs

2.1
Principles of Atomic Force Microscopy

2.1.1
Basic Concept

The direct measurement of the force interaction between distinct molecules has been a challenge for scientists for many years. In fact, only very recently was a demonstration given that these forces can be determined for a single atomic bond, by using the powerful technique of atomic force microscopy (AFM). But how is it possible to measure interatomic forces, which may be as small as one billionth of one Newton?

The answer to this question is surprisingly simple: It is the same mechanical principle used by a pair of kitchen scales, where a spring with a defined elasticity is elongated or compressed due to the weight of the object to be measured. The compression Δz of the spring (with spring constant c_z) is a direct measure of the force F exerted, which in the regime of elastic deformation obeys Hooke's law:

$$F = c_z \times \Delta z. \tag{2.1}$$

The only difference from the kitchen scale is the sensitivity of the measurement. In AFM, the 'spring' is a bendable cantilever with a stiffness of 0.01 N m^{-1} to 10 N m^{-1}. As interatomic forces are in the range of some nN, the cantilever will be deflected by 0.01 nm to 100 nm. Consequently, the precise detection of the cantilever bending is the key feature of an atomic force microscope. If a sufficiently sharp tip is directly attached to the cantilever, it would be possible to measure the interacting forces between the last atoms of the tip and the sample through the bending of the cantilever.

In 1986, Binnig, Quate and Gerber presented exactly this concept for the first atomic force microscope [1]. These authors measured the deflection of a cantilever with sub-Ångström precision by a scanning tunneling microscope [2] and used a gold foil as the spring. The tip was a piece of diamond glued to this home-made cantilever

Nanotechnology. Volume 6: Nanoprobes. Edited by Harald Fuchs

ISBN: 978-3-527-31733-2

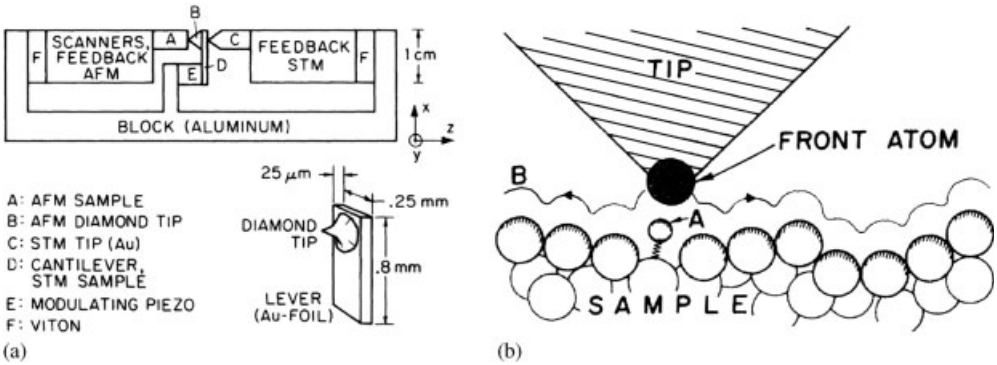


Figure 2.1 (a) The basic concept of the first atomic force microscope built in 1986 by Binnig, Quate and Gerber. A sharp diamond tip glued to a gold foil scanned the surface, while the bending of the cantilever was detected with scanning tunneling microscopy; (b) The ultimate goal was to measure the force between the front atom of the tip and a specific sample atom. (Reproduced from Ref. [1].)

(see Figure 2.1), and by using this set-up the group was able to image sample topographies down to the nanometer scale.

2.1.2 Current Experimental Set-Ups

During the past few years the experimental set-up has been modified while AFM has become a widespread research tool. Some 20 years after its invention, the commercial atomic force microscope is available from a variety of manufacturers. Although most of these instruments are designed for specific applications and environments, they are typically based on the following types of sensors detection method and scanning principles.

2.1.2.1 Sensors

Cantilevers are produced by standard microfabrication techniques, mostly from silicon and silicon nitride as rectangular or V-shaped cantilevers. Spring constants and resonance frequencies of cantilevers depend on the actual mode of operation. For contact AFM measurements these are about 0.01 to $1\,\mathrm{N\,m^{-1}}$ and 5–100 kHz, respectively. In a typical atomic force microscope, cantilever deflections ranging from 0.1 Å to a few micrometers are measured, which corresponds to a force sensitivity ranging from 10^{-13} N to 10^{-5} N.

Figure 2.2 shows two scanning electron microscopy (SEM) images of a typical rectangular silicon cantilever. When using this imaging technique, the length (l), width (w) and thickness (t) can be precisely measured. The spring constant c_z of the cantilever can then be determined from these values [3].

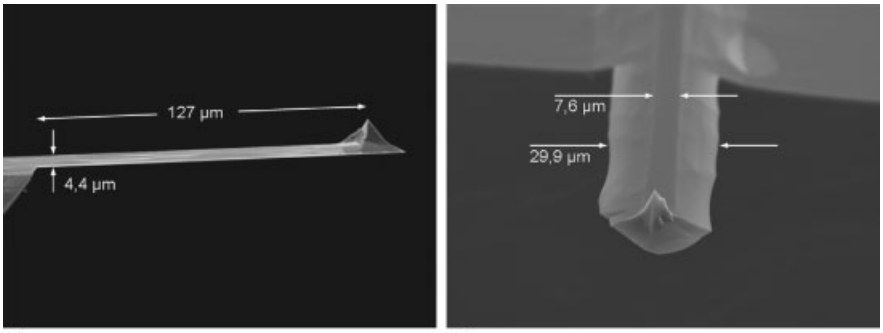


Figure 2.2 (a) Scanning electron microscopy image of a rectangular silicon cantilever with a width of 127 µm and a thickness of 4.4 µm; (b) A different view of the same cantilever reveals that the cross-section of the cantilever is trapezoidal and the cantilever has two geometric widths – a smaller one on the tip side and a broader one on the reverse side. Hence, most manufacturers provide a 'mean width', which for the cantilever shown is (7.6 + 29.9)/2 = 18.75 µm. The trapezoidal shape of the cantilever is caused by the anisotropic etching of the silicon during microfabrication of the cantilever. (Images courtesy of Boris Anczykowski, nanoAnalytics GmbH; used with kind permission.)

$$c_z = E_{\mathrm{Si}} \frac{w}{4} \left(\frac{t}{l}\right)^3 \tag{2.2}$$

where $E_{\mathrm{Si}} = 1.69 \times 10^{11}\,\mathrm{N\,m^{-2}}$ is the Youngs's modulus. The typical dimensions of silicon cantilevers are as follows: lengths of 100–300 µm; widths of 10–30 µm; and thicknesses of 0.3–5 µm.

The torsion of the cantilever due to lateral forces between tip and surface depends also on the height of the tip, h. The torsional spring constant can be calculated from [3]

$$c_{\mathrm{tor}} = \frac{G}{3} \frac{wt^3}{lh^2} \tag{2.3}$$

where $G_{\mathrm{Si}} = 0.68 \times 10^{11}\,\mathrm{N\,m^{-1}}$ is the shear modulus of silicon.

As the dimensions of cantilevers given by the manufacturer are only average values, the high-accuracy calibration of the spring constant requires the measurement of length, width and thickness for each individual cantilever. The length and width can be measured with sufficient accuracy using an optical microscope, but the thickness requires high-resolution techniques such as SEM. In order to avoid this time- and cost-consuming measurement one can determine the cantilever thickness from its eigenfrequency in normal direction [3–5]

$$t = \underbrace{\frac{4\sqrt{3}}{0.596861^2\pi} \sqrt{\frac{\rho_{\mathrm{Si}}}{E_{\mathrm{Si}}}}}_{\approx 7.23 \times 10^{-4}\ \mathrm{s/m}} l^2 f_0 \tag{2.4}$$

where the density of silicon $\rho_{\mathrm{Si}} = 2330\,\mathrm{kg\,m^{-3}}$.

The formulas presented above are only valid for rectangular cantilevers. Equations for V-shaped cantilevers can be found in Refs. [6, 7].

2.1.2.2 Detection Methods

In addition to changes in the cantilevers, the detection methods used to measure the minute bendings have also been improved. Today, commercial AFM instruments use the so-called *laser beam deflection* scheme shown in Figure 2.3. The bending and torsion of cantilevers can be detected by a laser beam reflected from their reverse side [8, 9], while the reflected laser spot is detected with a sectioned photo-diode. The different parts are read out separately. For this, a four-quadrant diode is normally used, in order to detect the normal as well as the torsional movements of the cantilever. With the cantilever at equilibrium the spot is adjusted such that the upper and lower sections show the same intensity. Then, if the cantilever bends up or down, the spot will move and the difference signal between the upper and lower sections will provide a measure of the bending.

The sensitivity can be improved by interferometer systems adapted by several research groups (see Refs [10–13]). It is also possible to use cantilevers with integrated deflection sensors based on piezoresistive films [14–16]. As no optical parts are

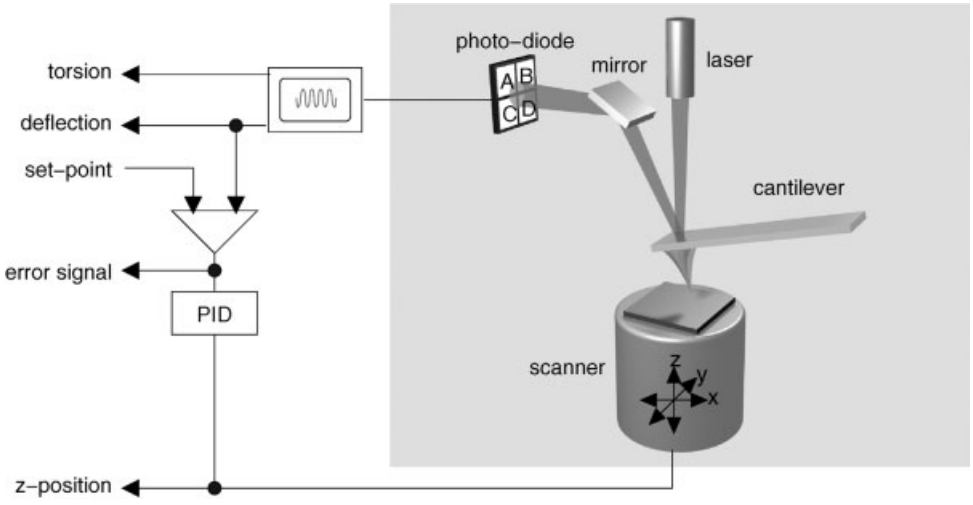


Figure 2.3 Principle of an atomic force microscope using the laser beam deflection method. Deflection (normal force) and torsion (friction) of the cantilever are measured simultaneously by measuring the lateral and vertical deflection of a laser beam while the sample is scanned in the x–y-plane. The laser beam deflection is determined using a four-quadrant photo diode. If A, B, C and D are proportional to the intensity of the incident light of the corresponding quadrant, the signal $(A + B) - (C + D)$ is a measure for the deflection, and $(A + C) - (B + D)$ is a measure of the torsion of the cantilever. A schematic of the feedback system is shown by the solid lines. The actual deflection signal of the photo-diode is compared with the set-point chosen by the experimentalist. The resultant error signal is fed into the PID controller, which moves the z-position of the scanner in order to minimize the deflection signal.

required in the experimental set-up of an atomic force microscope, when using these cantilevers the design can be very compact [17]. An extensive comparison of the different detection methods available can be found in Ref. [4].

2.1.2.3 Scanning and Feedback System

As the surface is scanned, the deflection of the cantilever is kept constant by a feedback system that controls the vertical movement of the scanner (shown schematically in Figure 2.3). The system functions as follows: (i) The current signal of the photo-diode is compared with a preset value; (ii) the feedback system, which includes a proportional, integral and differential (PID) controller, then varies the z-movement of the scanner to minimize the difference. As a consequence, the tip-sample force is kept practically constant for an optimal set-up of the PID parameters.

While the cantilever is moving relative to the sample in the x–y-plane of the surface by a piezoelectric scanner (see Figure 2.4), the current z-position of the scanner is recorded as a function of the lateral x–y-position with (ideally) sub-Ångström precision. The obtained data represents a map of equal forces, which is then analyzed and visualized by computer processing.

A principle of a simple laboratory class experiment – the imaging of a test grid – is shown in Figure 2.5. The comparison between the topography and the error signal shows that the PID controller needs some time at the step edges to correct the actual deflection error.

2.1.3 Tip–Sample Forces in Atomic Force Microscopy

A large variety of sample properties related to tip-sample forces can be detected using the atomic force microscope. The obtained contrast depends on the operational mode and

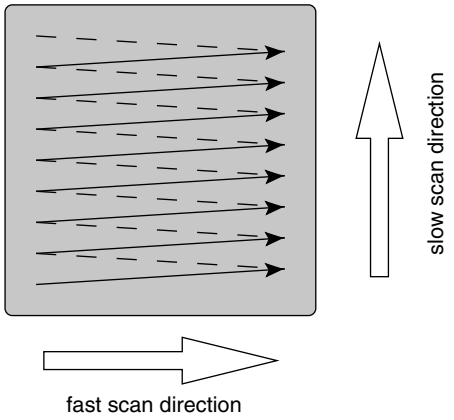


Figure 2.4 Schematic of the scan process. The cantilever scans the sample surface systematically in the x- and y-directions. Typical scan sizes ranging from less than 1 nm × 1 nm to 150 μm × 150 μm can be used.

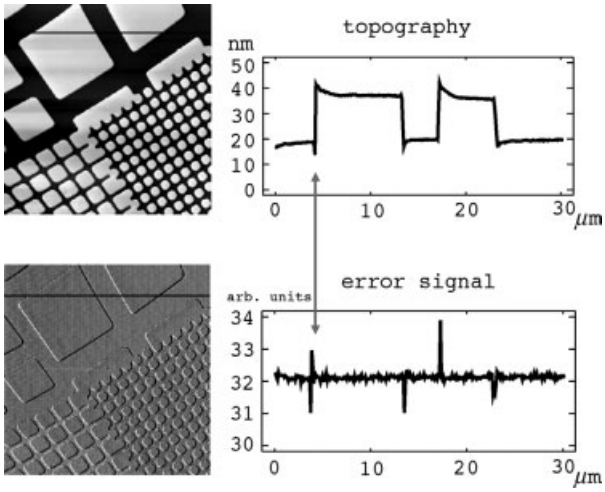


Figure 2.5 A simple laboratory class experiment demonstrating the scanning process of an atomic force microscope in contact mode. The cantilever is scanned over a simple test grid made of silicon. The feedback keeps the deflection (and therefore the force) constant, and the *z*-position of the scanner is interpreted as the topography of the sample. The resultant map is plotted as a gray-scale image (lighter areas correspond to higher topography, upper left graph). The simultaneously plotted error signal (lower graph) shows that the feedback fails to keep the deflection constant at the step edges.

the actual tip-sample interactions. Before discussing details of the operational modes of AFM, however, we must first specify the most important tip–sample interactions.

Figure 2.6 shows the typical shape of the interaction force curve that the tip senses during an approach towards the sample surface. Upon approach of the tip towards the

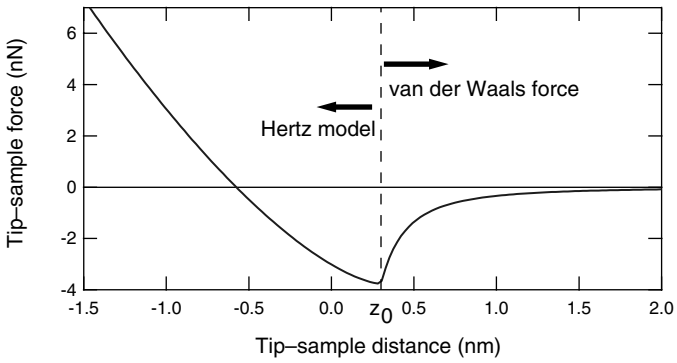


Figure 2.6 Tip–sample model force after the DMT-M model for air Equation 2.9, using the parameters described in the text. The dashed line marks the position z_0, where the tip touches the surface.

sample, the negative attractive forces (which represent, for example, van der Waals or electrostatic interaction forces) increase until a minimum is reached. This turn-around point is due to the onset of repulsive forces, caused by Pauli repulsion, which will start to dominate upon further approach. Eventually, the tip is pushed into the sample surface and elastic deformation will occur.

In general, the effective tip–sample interaction force is a sum of different force contributions, and these can be roughly divided into *attractive* and *repulsive* components. The most important forces are summarized as follows.

2.1.3.1 Van der Waals Forces

These forces are caused by fluctuating induced electric dipoles in atoms and molecules. The distance-dependence of this force for two distinct molecules follows $1/z^7$. For simplicity, solid bodies are often assumed to consist of many independent noninteracting molecules, and the van der Waals forces of these bodies are obtained by simple summation. For example, for a sphere over a flat surface the van der Waals force is given by

$$F_{vdW}(z) = -\frac{A_H R}{6z^2}, \tag{2.5}$$

where R is the radius of the sphere and A_H is the Hamaker constant, which is typically in the range of $\approx$0.1 aJ [18]. This geometry is often used to approximate the van der Waals forces between the tip and sample. Due to the $1/z^2$ dependency, van der Waals forces are considered long-range forces compared to the other forces that occur in AFM.

2.1.3.2 Capillary Forces

Capillary forces are important under ambient conditions. Water molecules condense at the sample surface (and also on the tip) and cause the occurrence of an adsorption layer. Consequently, the atomic force microscope tip penetrates through this layer when approaching the sample surface. At the tip–sample contact, a water meniscus is formed which causes a very strong attractive force [19]. For soft samples these forces often lead to unwanted deformations of the surface; however, this effect can be circumvented by measuring directly in liquids. Alternatively, capillary forces can be avoided by performing the experiments in a glovebox with dry gases, or in vacuum.

2.1.3.3 Pauli or Ionic Repulsion

These forces are the most important in conventional contact mode AFM. The Pauli exclusion principle forbids that the charge clouds of two electrons showing the same quantum numbers can have some significant overlap; first, the energy of one of the electrons must be increased, and this yields a repulsive force. In addition, an overlap of the charge clouds of electrons can cause an insufficient screening of the nuclear charge, leading to ionic repulsion of coulombic nature. The Pauli and the ionic repulsion are nearly hard-wall potentials. Thus, when the tip and sample are in intimate contact most of the (repulsive) interaction is carried by the atoms directly at the interface. The Pauli repulsion is of purely quantum mechanical origin, and semi-

empirical potentials are mostly used to allow an easy and fast calculation. One well-known model is the *Lennard–Jones* potential, which combines short-range repulsive interactions with long-range attractive van der Waals interactions:

$$V_{LJ}(z) = E_0\left(\left(\frac{r_0}{z}\right)^{12} - 2\left(\frac{r_0}{z}\right)^{6}\right) \tag{2.6}$$

where E_0 is the bonding energy and r_0 the equilibrium distance. In this case, the repulsion is described by an inverse power law with $n = 12$. The term with $n = 6$ describes the attractive van der Waals potential between two atoms/molecules.

2.1.3.4 Elastic Forces

Elastic forces and deformations can occur if the tip is in contact with the sample. As this deformation affects the effective contact area, knowledge about the elastic forces and the corresponding deformation mechanics of the contact is an important issue in AFM. The repulsive forces that occur during the elastic indentation of a sphere into a flat surface were analyzed as early as 1881 by H. Hertz (see Refs [20, 21]),

$$F_{\text{Hertz}}(z) = \frac{4}{3}E^*\sqrt{R}(z_0 - z)^{3/2} \quad \text{for} \quad z \le z_0, \tag{2.7}$$

where the effective elastic modulus E^*

$$\frac{1}{E^*} = \frac{(1-\mu_t^2)}{E_t} + \frac{(1-\mu_s^2)}{E_s} \tag{2.8}$$

depends on the Young's moduli $E_{t,s}$ and the Poisson ratios $\mu_{t,s}$ of the tip and surface, respectively. Here, R is the tip radius and z_0 is the point of contact.

This model does not include adhesion forces, however, which must be considered at the nanometer scale. Two extreme cases were analyzed by Johnson *et al.* [22] and Derjaguin *et al.* [23]. The model of Johnson, Kendall and Roberts (the JKR model) considers only the adhesion forces *inside* the contact area, whereas the model of Derjaguin, Muller and Toporov (the DMT model) includes only the adhesion *outside* the contact area. Various models analyzing the contact mechanics in the intermediate regime were suggested by other authors (see Ref. [24] for a recent overview).

However, in many practical cases it is sufficient to assume that the geometric shape of the tip and sample does not change until contact has been established at $z = z_0$, and that afterwards, the tip–sample forces are given by the DMT-M theory, denoting Maugis' approximation to the earlier DMT model [24]. In this approach, an offset $F_{\text{vdW}}(z_0)$ is added to the well-known Hertz model, which accounts for the adhesion force between tip and sample surface. Therefore, the DMT-M model is often also referred to as *Hertz-plus-offset model* [24]. The resulting overall force law is given by

$$F_{\text{DMT-M}}(z) = \begin{cases} -\dfrac{A_H R}{6z^2} & \text{for} \quad z \ge z_0\,, \\ \dfrac{4}{3}E^*\sqrt{R}(z_0 - z)^{3/2} - \dfrac{A_H R}{6z_0^2} & \text{for} \quad z < z_0\,. \end{cases} \tag{2.9}$$

Figure 2.6 shows the resulting tip–sample force curve for the DMT-M model. The following parameters were used, representing typical values for AFM measurements under ambient conditions: $A_H = 0.2$ aJ; $R = 10$ nm; $z_0 = 0.3$ nm; $\mu_t = \mu_s = 0.3$; $E_t = 130$ GPa; and $E_s = 1$ GPa.

2.1.3.5 Frictional Forces

Frictional forces counteract the movement of the tip during the scan process, and dissipate the kinetic energy of the moving tip–sample contact into the surface or tip material. This can be due to permanent changes in the surface itself, by scratching or indenting, or also by the excitation of lattice vibration (i.e. phonons) in the material.

2.1.3.6 Chemical Binding Forces

Chemical binding forces arise from the overlap of molecular orbitals, due to specific bonding states between the tip and the surface molecules. These forces are extremely short-ranged, and can be exploited to achieve atomic resolution imaging of surfaces. As these forces are also specific to the chemical identity of the molecules, it is conceivable to identify the chemical character of the surface atoms with AFM scans.

2.1.3.7 Magnetic and Electrostatic Forces

These forces are of long-range character and might be either attractive or repulsive; they are usually measured when the tip is not in contact with the surface (i.e. 'noncontact' mode). For magnetic forces, magnetic materials must be used for tip or tip coating. Well-defined electrical potentials between tip and sample are necessary for the measurement of electrostatic forces.

More detailed information on the intermolecular and surface forces relevant for AFM measurements can be found in the monographs of Israelachvili [18] and Sarid [25]. Details of the most important forces are summarized in Figure 2.7. Although, in principle, every type of force can be measured using the atomic force microscope, the actual sensitivity to a specific force depends on the mode of operation. Hence, the most important modes are introduced in the next section.

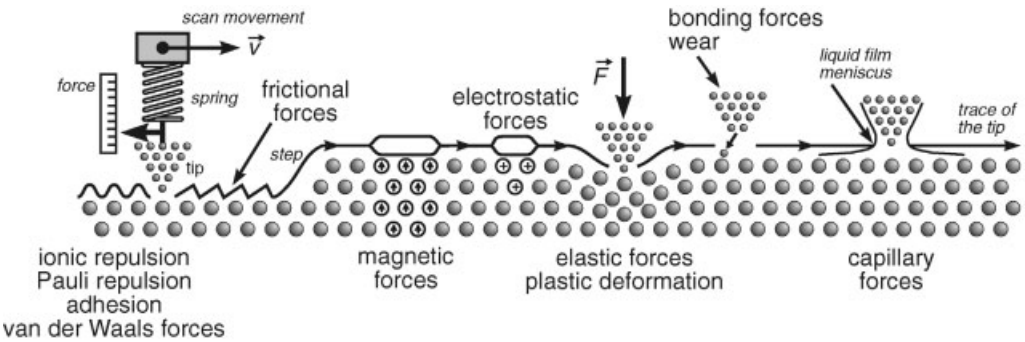


Figure 2.7 Summary of the forces relevant in atomic force microscopy. (Image courtesy of Udo D. Schwarz, Yale University; used with kind permission.)

2.2
Modes of Operation

Although an atomic force microscope can be driven in different modes of operation, we concentrate here on the two most important modes that are widely used to image sample surfaces down to the atomic scale.

2.2.1
Static or Contact Mode

The contact mode, which historically is the oldest, is used frequently to obtain nanometer-resolution images on a wide variety of surfaces. This technique also has the advantage that not only the deflection, but also the torsion of the cantilever, can be measured. As shown by Mate *et al.* [26], the lateral force can be directly correlated to the friction between tip and sample, thus extending AFM to friction force microscopy (FFM).

Some typical applications of an atomic force microscope driven in contact-mode are shown in Figure 2.8a and b. Here, the images represent a measurement of a L-α-

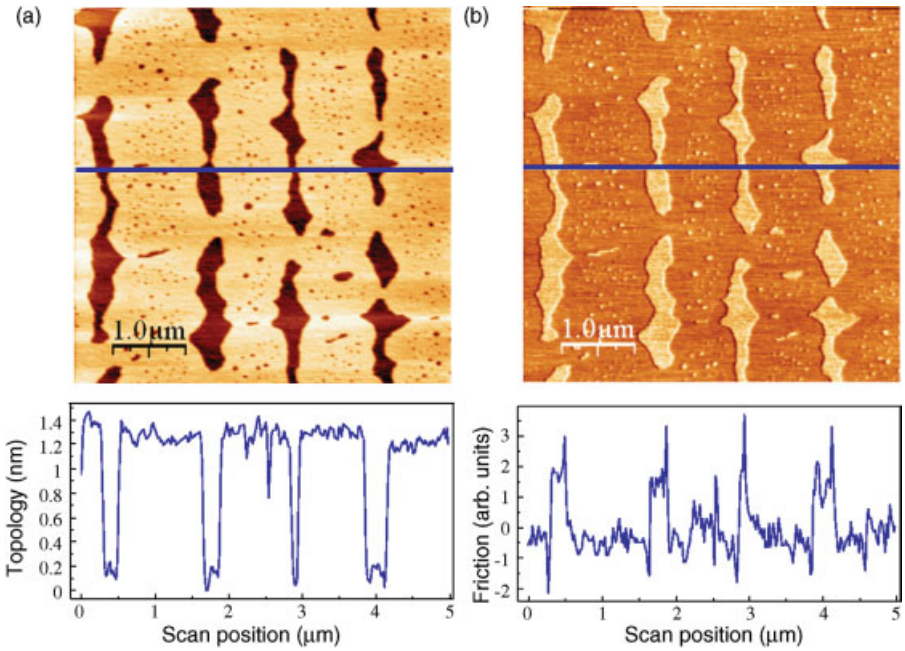


Figure 2.8 (a) Atomic force microscopy image obtained in contact mode of a monomolecular DPPC (L-α-dipalmitoyl-phophatidycholine) film adsorbed onto mica. The image is color-coded; that is, dark areas represent the mica substrate and light areas the DPPC film; (b) The simultaneously recorded friction image shows a lower friction on the film (dark areas) than on the substrate (light areas). The graphs represent single scan lines obtained at the positions marked by a dark line in the above images.

dipalmitoyl-phophatidycholine (DPPC) film adsorbed onto a mica substrate. The lateral force was simultaneously recorded with the topography, and shows a contrast between the DPPC film and the substrate. This effect can be attributed to the different frictional forces on DPPC and the mica substrate, and is frequently used to obtain a chemical contrast on flat surfaces [27, 28].

2.2.1.1 Force versus Distance Curves

So far, we have neglected one important issue for the operation of the atomic force microscope, namely the mechanical stability of the measurement. In static AFM the tip is allowed to approach very slowly towards the surface, and the attractive forces between the tip and sample must be counteracted by the restoring force of the cantilever. However, this fails if the force gradient of the tip–sample forces is larger than the spring constant of the cantilever. Mathematically speaking, an instability occurs if

$$c_z < \frac{\partial F_{ts}(z)}{\partial z}. \tag{2.10}$$

In this case the attractive forces can no longer be sustained by the cantilever and the tip 'jumps' towards the sample surface [29].

This effect has a strong influence on static-mode AFM measurements, as exemplified by a typical force-versus-distance curve shown in Figure 2.9. Here, the force

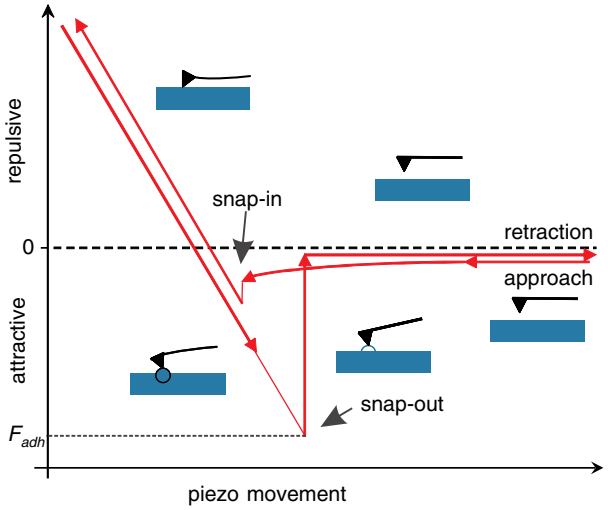


Figure 2.9 A schematic of a typical force versus distance curve obtained in static mode. The cantilever is approached towards the sample surface. Due to strong attractive forces it 'jumps' (snap-in) towards the sample surface at a specific position. During retraction, the tip is strongly attracted by the surface and the 'snap-out' point is considerably behind the 'snap-in' point. This results in an hysteresis between approach and retraction.

acting on the tip recorded during an approach and retraction movement of the cantilever is depicted. Upon approach of the cantilever towards the sample, the attractive forces acting on the tip bend the cantilever towards the sample surface. At a specific point close to the sample surface these forces can be no longer sustained by the cantilever spring, and the tip 'jumps' towards the sample surface. Now, the tip and sample are in direct mechanical contact, and a further approach towards the sample surface pushes the tip into the sample. As the spring constant of the cantilever usually is much softer than the elasticity of the sample, the bending of the cantilever increases almost linearly.

If the cantilever is now retracted from the surface, the tip stays in contact with the sample because it is strongly attracted by the sample due to adhesive forces, and the force F_{adh} is necessary to disconnect the tip from the surface. The 'snap-out' point is always at a larger distance from the surface than the 'snap-in', and this results in an hysteresis between the approach and retraction of the cantilever. This phenomenon of mechanical instability is often referred to as the *jump-to-contact*. Unfortunately, this sudden jump can lead to undesired changes of the tip and/or sample.

2.2.2 Dynamic Modes

Despite the success of contact-mode AFM, the resolution was found to be limited in many cases (in particular for soft samples) by lateral forces acting between tip and sample. In order to avoid this effect, the cantilever can be oscillated in a vertical direction near the sample surface. AFM imaging with vibrating cantilever is often denoted as dynamic force microscopy (DFM).

The historically oldest scheme of cantilever excitation in DFM imaging is the *external* driving of the cantilever at a *fixed excitation frequency* exactly at or very close to the cantilever's first resonance [30–32]. For this driving mechanism, different detection schemes measuring either the change of the oscillation amplitude or the phase shift were proposed. Over the years, the amplitude modulation (AM) or 'tapping' mode, where the oscillation amplitude is used as a measure of the tip–sample distance, has developed into the most widespread technique for imaging under ambient conditions and liquids.

In a vacuum, any external oscillation of the cantilever is disadvantageous. Standard AFM cantilevers constructed from silicon exhibit very high Q-values in vacuum, which results in very long response times of the system. Consequently, in 1991 Albrecht *et al*. [33] introduced the frequency modulation (FM) mode, which works well for high-Q systems and subsequently has developed into the dominant driving scheme for high-resolution DFM experiments in ultra-high vacuum (UHV) [34–37]. In contrast to the AM mode, this approach features a so-called *self-driven* oscillator [38, 39] which, when placed in a closed-loop set-up ('active feedback'), uses the cantilever deflection itself as the driving signal, thus ensuring that the cantilever instantaneously adapts to changes in the resonance frequency. These two driving mechanisms are discussed in more detail in the following section.

2.3
Amplitude Modulation (Tapping Mode)

2.3.1
Experimental Set-Up of AM-Atomic Force Microscopy

As an alternative to the contact mode, the cantilever can be excited to vibrate near its resonant frequency close to the sample surface. Under the influence of tip–sample forces the resonant frequency (and consequently also the amplitude and phase) of the cantilever will change and serve as the measurement parameters. This is known as the *dynamic* mode. If the tip is approached towards the surface, the oscillation parameters of amplitude and phase are influenced by the tip–surface interaction, and can therefore be used as feedback channels. A certain set-point (e.g. the amplitude) is given, whereby the feedback loop will adjust the tip–sample distance so that the amplitude remains constant. The controller parameter is recorded as a function of the lateral position of the tip with respect to the sample, and the scanned image essentially represents the surface topography.

The technical realization of dynamic-mode AFM is based on the same key components as a static AFM set-up. A sketch of the experimental set-up of an atomic force microscope driven in AM mode is shown in Figure 2.10.

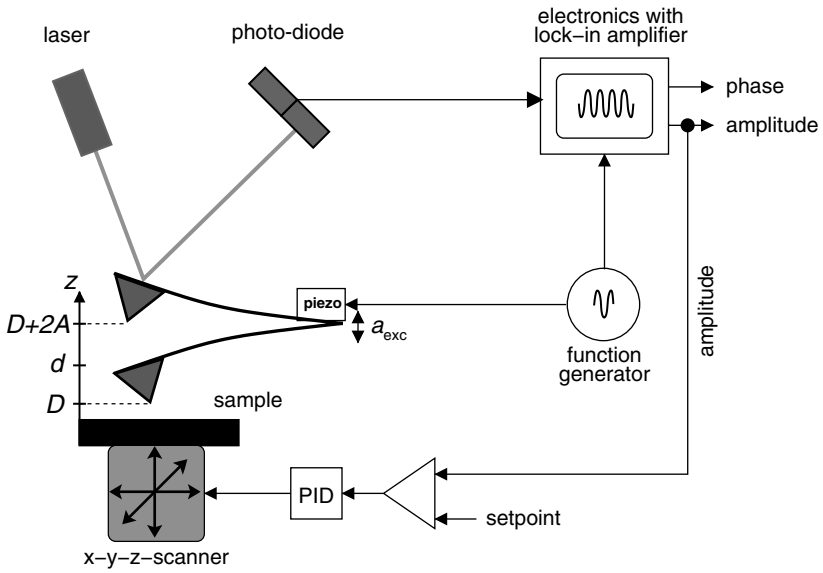


Figure 2.10 Set-up of a dynamic force microscope operated in AM or tapping mode. A laser beam is deflected by the reverse side of the cantilever, with the deflection being detected by a split photo-diode. The cantilever vibration is caused by an external frequency generator driving an excitation piezo. A lock-in amplifier is used to compare the cantilever driving with its oscillation. The amplitude signal is held constant by a feedback loop which controls the cantilever–sample distance.

The deflection of the cantilever is typically measured with the laser beam deflection method, as indicated [8, 9], but other displacement sensors such as *interferometric sensors* [12, 13, 30, 40] have also been applied. During operation in conventional tapping mode, the cantilever is driven at a fixed frequency with a constant excitation amplitude from an external function generator, while the resulting oscillation amplitude and/or the phase shift are detected by a lock-in amplifier. The function generator supplies not only the signal for the dither piezo; its signal serves simultaneously as a reference for the lock-in amplifier.

This set-up can be operated both in air and in liquids. A typical image obtained with this experimental set-up in ambient conditions is shown in Figure 2.11. For a direct comparison with the static mode, the sample is also DPPC-adsorbed onto a mica substrate. In contact mode the frictional forces are measured simultaneously with the topography, whereas in dynamic mode the phase between excitation and oscillation is acquired as an additional channel. The phase image provides information about the different material properties of DPPC and the mica substrate. It can be shown, that the phase signal is closely related to the energy dissipated in the tip–sample contact [41–43].

Due to its technical relevance the investigation of polymers has been the focal point of many studies (see Ref. [44] for a recent review). High-resolution imaging has been extensively performed in the area of materials science; for example, by using specific

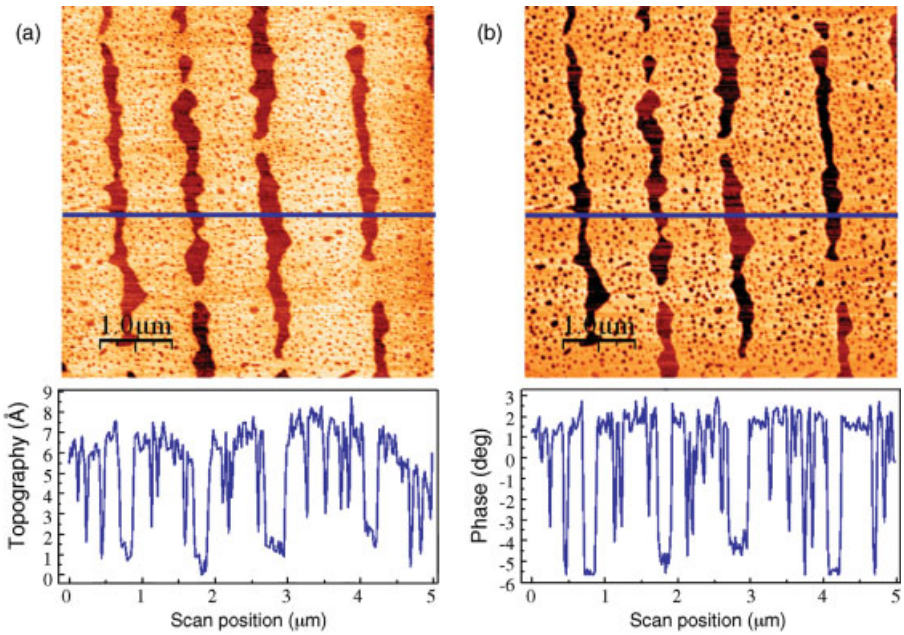


Figure 2.11 (a) A dynamic force microscopy image of a monomolecular DPPC film adsorbed onto mica; (b) The phase contrast is directly related to the topography; that is, the phase is different between the substrate and the DPPC film.

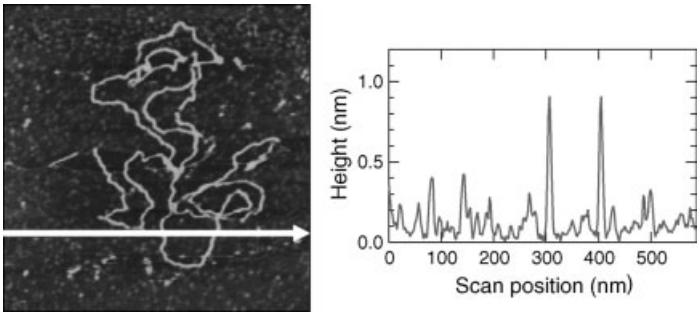


Figure 2.12 Topography of DNA adsorbed onto mica imaged in buffer solution by tapping mode AFM. The graph shows a single scan line obtained at the position marked by a white arrow in the image.

tips with additionally grown sharp spikes, Klinov *et al.* [45] obtained true molecular resolution on a polydiacetylene crystal.

Imaging in liquids opens up an avenue for the investigation of biological samples in their natural environment. For example, Möller *et al.* [46] have obtained high-resolution images of the topography of the hexagonally packed intermediate (HPI) layer of *Deinococcus radiodurans*, using tapping-mode AFM. A typical example of the imaging of DNA in liquid solution is shown in Figure 2.12.

2.3.1.1 Theory of AM-AFM

Based on the above description of the experimental set-up, it is possible to formulate the basic equation of motion describing the cantilever dynamics of AM-AFM:

$$m\ddot{z}(t) + \frac{2\pi f_0 m}{Q_0}\dot{z}(t) + c_z(z(t) - d) + = \underbrace{a_d c_z \cos(2\pi f_d t)}_{\text{external driving force}} + \underbrace{F_{ts}[z(t), \dot{z}(t)]}_{\text{tip-sample force}}. \tag{2.11}$$

Here, $\dot{z}(t)$ is the position of the tip at the time t; c_z, m and $f_0 = \sqrt{(c_z/m)}/(2\pi)$ are the spring constant, the effective mass, and the eigenfrequency of the cantilever, respectively. As a small simplification, it is assumed that the quality factor Q_0 combines the intrinsic damping of the cantilever and all influences from surrounding media, such as air or liquid (if present) in a single value. The equilibrium position of the tip is denoted as d. The first term on the right-hand side of the equation represents the external driving force of the cantilever by the frequency generator. It is modulated with the constant excitation amplitude a_d at a fixed frequency f_d. The (nonlinear) tip–sample interaction force F_{ts} is introduced by the second term.

Before discussing the solutions of this equation, some words of caution should be added with regards to the universality of the equation of motion and the various solutions discussed below. Equation 2.11 disregards two effects, which might become important under certain circumstances. First, we describe the cantilever by a spring-mass-model and neglect in this way the higher modes of the cantilever.

This is justified in most cases, as the first eigenfrequency is by far the most dominant in typical AM-AFM experiments (see Refs [41, 47–50]). Second, we assume in our model equation of motion that the dither piezo applies a sinusoidal force to the spring, but do not consider that the movement of the dither piezo simultaneously also changes the effective position of the tip at the cantilever end by $a_{\mathrm{exc}}(t) = a_d \cos(2\pi f_d t)$ [47, 51, 52]. This effect becomes important when a_d is in the range of the cantilever oscillation amplitude.

In a first step, we assume that the cantilever vibrates far away from the sample surface. Consequently, we can neglect tip–sample forces ($F_{ts} \equiv 0$), resulting in the well-known equation of motion of a driven damped harmonic oscillator.

After some time the external driving amplitude forces the cantilever to oscillate exactly at the driving frequency f_d. Therefore, the steady-state solution is given by the ansatz

$$z(t \gg 0) = d + A\cos(2\pi f_d t + \phi), \tag{2.12}$$

where ϕ is the phase difference between the excitation and the oscillation of the cantilever. With this, we obtain two functions for the amplitude and phase curves:

$$A = \frac{a_d}{\sqrt{\left(1 - \frac{f_d^2}{f_0^2}\right)^2 + \left(\frac{1}{Q_0}\frac{f_d}{f_0}\right)^2}}, \tag{2.13a}$$

$$\tan\phi = \frac{1}{Q_0}\frac{f_d/f_0}{1 - f_d^2/f_0^2}. \tag{2.13b}$$

The features of such an oscillator are well known from introductory physics courses.

If the cantilever is brought closer towards the sample surface, the tip senses the tip–sample interaction force, F_{ts}, which changes the oscillation behavior of the cantilever. However, as the mathematical form of realistic tip–sample forces is highly nonlinear, this fact complicates the analytical solution of the equation of motion Equation 2.11. For the analysis of DFM experiments we need to focus on steady-state solutions of the equation of motion with sinusoidal cantilever oscillation. Therefore, it is advantageous to expand the tip–sample force into a Fourier series

$$\begin{aligned} F_{ts}[z(t), \dot{z}(t)] \approx{}& f_d \int_0^{1/f_d} F_{ts}[z(t), \dot{z}(t)]dt \\ &+ 2f_d \int_0^{1/f_d} F_{ts}[z(t), \dot{z}(t)]\cos(2\pi f_d t + \phi)dt \times \cos(2\pi f_d t + \phi) \\ &+ 2f_d \int_0^{1/f_d} F_{ts}[z(t), \dot{z}(t)]\sin(2\pi f_d t + \phi)dt \times \sin(2\pi f_d t + \phi) \\ &+ \ldots, \end{aligned} \tag{2.14}$$

where $z(t)$ is given by Equation 2.12.

The first term in the Fourier series reflects the averaged tip–sample force over one full oscillation cycle, which shifts the equilibrium point of the oscillation by a small offset Δd from d to d_0. Actual values for Δd, however, are very small. For typical amplitudes used in AM-AFM in air (some nm to some tens of nm), the averaged tip–sample force is in the range of some pN. The resultant offset Δd is less than 1 pm for typical sets of parameters. As this is well beyond the resolution limit of an AM-AFM experiment in air, we neglect this effect in the following and assume $d \approx d_0$ and $D = d - A$.

For further analysis, we now insert the first harmonics of the Fourier series Equation 2.14 into the equation of motion (Equation 2.11), thus obtaining two coupled equations [53, 54]

$$\frac{f_0^2 - f_d^2}{f_0^2} = I_+(d, A) + \frac{a_d}{A}\cos\phi, \tag{2.15a}$$

$$-\frac{1}{Q_0}\frac{f_d}{f_0} = I_-(d, A) + \frac{a_d}{A}\sin\phi, \tag{2.15b}$$

where the following integrals have been defined:

$$\begin{aligned} I_+(d, A) &= \frac{2f_d}{c_z A}\int_0^{1/f_d} F_{ts}[z(t), \dot{z}(t)]\cos(2\pi f_d t + \phi)dt \\ &= \frac{1}{\pi c_z A^2}\int_{d-A}^{d+A} (F_\downarrow + F_\uparrow)\frac{z-d}{\sqrt{A^2-(z-d)^2}}dz, \end{aligned} \tag{2.16a}$$

$$\begin{aligned} I_-(d, A) &= \frac{2f_d}{c_z A}\int_0^{1/f_d} F_{ts}[z(t), \dot{z}(t)]\sin(2\pi f_d t + \phi)dt \\ &= \frac{1}{\pi c_z A^2}\int_{d-A}^{d+A} (F_\downarrow - F_\uparrow)dz \\ &= \frac{1}{\pi c_z A^2}\Delta E(d, A). \end{aligned} \tag{2.16b}$$

Both integrals are functions of the actual oscillation amplitude A and the cantilever–sample distance d. Furthermore, they depend on the sum and the difference of the tip–sample forces during approach ($E_\downarrow$) and retraction ($F_\uparrow$), as manifested by the labels '+' and '−' for easy distinction. The integral I_+ is a weighted average of the tip–sample forces ($F_\downarrow + F_\uparrow$). On the other hand, the integral I_- is directly connected to ΔE, which reflects the energy dissipated during an individual oscillation cycle. Consequently, this integral vanishes for purely conservative tip–sample forces, where $F_\downarrow$ and $F_\uparrow$ are identical. A more detailed discussion of these integrals can be found in Refs. [55, 56].

By combining Equations 1.13b and 1.16b we obtain a direct correlation between the phase and the energy dissipation.[1)]

1) The '−'-sign on the right-hand side of the equation is due to our definition of the phase ϕ in Equation 2.12.

$$\sin\phi = -\left(\frac{A}{A_0}\frac{f_d}{f_0} + \frac{Q_0 \Delta E}{\pi c_z A_0 A}\right). \tag{2.17}$$

This relationship can be also obtained from the conservation of energy principle [41–43].

Equation (2.15) can be used to calculate the resonance curves of a dynamic force microscope, including tip–sample forces. The results are

$$A = \frac{a_d}{\sqrt{\left(1 - \frac{f_d^2}{f_0^2} - I_+(d,A)\right)^2 + \left(\frac{1}{Q_0}\frac{f_d}{f_0} + I_-(d,A)\right)^2}}, \tag{2.18a}$$

$$\tan\phi = \frac{\frac{1}{Q_0}\frac{f_d}{f_0} + I_-(d,A)}{1 - \frac{f_d^2}{f_0^2} - I_+(d,A)}. \tag{2.18b}$$

Equation 2.18a describes the shape of the resonance curve, but it is an implicit function of the oscillation amplitude A, and cannot be plotted directly.

Figure 2.13 contrasts the solution of this equation (solid lines) with numerical solution (symbols). As pointed out by various authors (see Refs [47, 57–63]), the amplitude versus frequency curves are multivalued within certain parameter ranges. Moreover, as the gradient of the analytical curve increases to infinity at specific positions, some branches become unstable. The resulting instabilities are reflected by the 'jumps' in the simulated curves (marked by arrows in Figure 2.13), where only stable oscillation states are obtained. Obviously, they are different for increasing and decreasing driving frequencies. This well-known effect is frequently observed in nonlinear oscillators (see Refs [64, 65]).

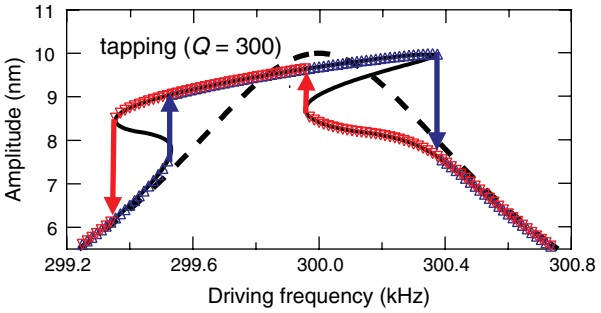


Figure 2.13 Resonance curves for tapping mode operation if the cantilever oscillates near the sample surface with $d = 8.5$ nm and $A_0 = 10$ nm, thereby experiencing the model force field given by Equation 2.9. The solid lines represent the analytical result of Equation 2.21, while the symbols are obtained from a numerical solution of the equation of motion, Equation 2.11. The dashed lines reflect the resonance curves without tip–sample force, and are shown purely for comparison. The resonance curve exhibits instabilities ('jumps') during a frequency sweep; these jumps take place at different positions (marked by arrows), depending on whether the driving frequency is increased or decreased.

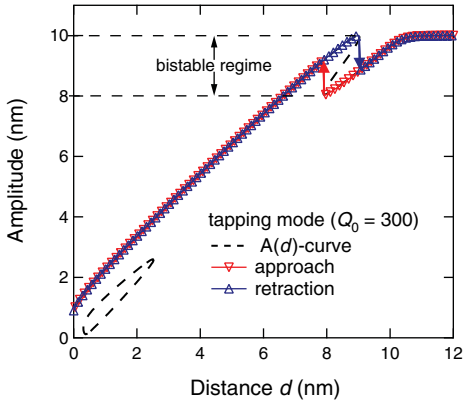


Figure 2.14 Amplitude versus distance curve for conventional ('tapping mode') AM-AFM for $A_0 = 10$ nm, $f_0 = 300$ kHz, and a tip–sample interaction force as given in Figure 2.6. The dashed lines represent the analytical result, while the symbols are obtained from a numerical solution of the equation of motion, Equation 2.11. The overall amplitudes decrease during an approach towards the sample surface in both cases, although instabilities (indicated by red and blue arrows) occur.

In AM-AFM, the cantilever might be oscillated at any frequency around the resonance peak. Here, we restrict ourselves to the situation where the driving frequency is set *exactly* to the eigenfrequency of the cantilever ($f_d = f_0$). With this choice – which is also very common in actual DFM experiments – we have defined imaging conditions leading to handy formulas suitable for further analysis. A discussion on the alternative cases of driving the cantilever slightly above or slightly below resonance can be found in the Refs [66–68].

From Equation (2.18a) we obtain the following relationship between the free oscillation amplitude A_0, the actual amplitude A, and the equilibrium tip position $\bar{d}$:

$$A_0 = A\sqrt{1 + (Q_0 I_+[\bar{d}, A])^2}. \tag{2.19}$$

In order to derive this formula, we used the approximation that the maximal value of the free oscillation amplitude at resonance is given by $A_0 \approx a_d Q_0$.

Solving this equation allows us to study amplitude versus distance curves for different effective Q-factors, as shown in Figure 2.14 for a Q-factor of 300. As observed previously in the resonance curves displayed in Figure 2.13, stable and unstable branches develop, which can be unambiguously identified by a comparison with numerical results (symbols).

Most noticeable, the tapping-mode curve exhibits jumps between unstable branches, which occur at different locations for approach and retraction. The resulting bistable regime then causes a hysteresis between approach and retraction, which has been the focus of numerous experimental and theoretical studies (see Refs [47, 61, 66, 67, 69–72]). As shown by various authors, the instability in conventional AM-AFM divides the tip–sample interaction into two regimes [47, 61, 70, 72]. Before the instability occurs, the tip interacts during an individual oscillation exclusively with the attractive part of

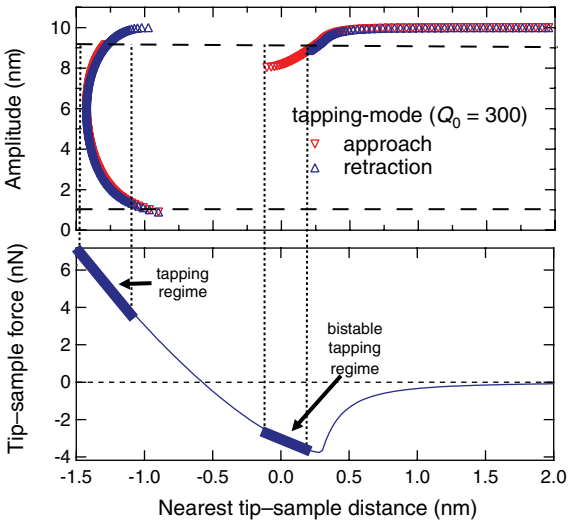


Figure 2.15 A comparison between the maximum tip–sample forces (tip–sample forces acting at the point of closest tip–sample approach/nearest tip–sample position *D*) experienced by conventional 'tapping mode' AM-AFM, assuming the same parameters as in Figure 2.14. The upper graph shows the nearest tip–sample position *D* versus the actual oscillation amplitude *A* for tapping mode. The lower graph shows the force regimes sensed by the tip. The maximal tip–sample forces in tapping mode are on the repulsive (tapping regime) as well as attractive (bistable tapping regime) part of the tip–sample force curve.

the tip–sample force. After jumping to the higher branch, however, the tip senses also the repulsive part of the tip–sample interaction.

In Figure 2.15 the oscillation amplitude is plotted as a function of the nearest tip–sample distance. In addition, the lower graph depicts the corresponding tip–sample force (cf. Figure 2.6). The origin of the nearest tip–sample position *D* is defined by this force curve. As both the amplitude curves and the tip–sample force curve are plotted as a function of the nearest tip–sample position, it is possible to identify the resulting maximum tip–sample interaction force for a given oscillation amplitude.

A closer look at the $A(D)$-curves helps to identify the different interaction regimes in AM-AFM. During the approach of the vibrating cantilever towards the sample surface, this curve shows a discontinuity for the nearest tip–sample position *D* (the point of closest approach during an individual oscillation) between 0 and −1 nm. This gap corresponds to the bistability and the resulting jumps in the amplitude versus distance curve. When the jump from the attractive to the repulsive regime has occurred, the amplitude decreases continuously, but the nearest tip–sample position does not reduce accordingly, remaining roughly between −0.8 nm and −1.5 nm. As a result, larger A/A_0 ratios do not necessarily translate into lower tip–sample interactions – a point which is important to bear in mind while adjusting imaging parameters in tapping mode AM-AFM imaging. In contrast, once the repulsive regime has been reached, the user's ability to influence the tip–sample interaction

strength by modifying the set-point for A is limited, thus also limiting the possibilities of improving the image quality.

For practical applications, it is reasonable to assume that the set-point of the amplitude used for imaging has been set to a value between 90% (= 9 nm) and 10% (= 1 nm) of the free oscillation amplitude. With this condition, we can identify the accessible imaging regimes indicated by the horizontal (dashed) lines and the corresponding vertical (dotted) lines in Figure 2.15. In tapping mode, two imaging regimes are realized: the *tapping* regime (left) and the *bistable tapping* regime (middle). The first can be accessed by any amplitude set-point between 9 nm and 1 nm, and results in a maximum tip–sample forces well within the repulsive regime. The second regime, belonging to the bistable imaging state, is only accessible during approach; here, the corresponding amplitude set-point is between 9 nm and 8 nm. Imaging in this regime is possible with the limitation that the oscillating cantilever might jump into the repulsive regime [68, 70].

2.3.1.2 Reconstruction of the Tip–Sample Interaction

Previously, we have outlined the influence of the tip–sample interaction on the cantilever oscillation, calculated the maximum tip–sample interaction forces based on the assumption of a specific model force, and subsequently discussed possible routes for image optimization. However, during AFM imaging, the tip–sample interaction is not known *a priori*. However, several groups [52, 73–75] have suggested solutions to this inversion problem. Here, we present an approach which is based on the analysis of the amplitude and phase versus distance curves which can easily be measured with most AM-AFM set-ups.

Let us start by applying the transformation $D = d - A$ to the integral I_+ in Equation 2.16a, where D corresponds to the nearest tip–sample distance, as defined in Figure 2.10. Next we note that, due to the cantilever oscillation, the current method intrinsically recovers the values of the force that the tip experiences at its lower turning point, where $F_\downarrow$ necessarily equals $F_\uparrow$. We thus define $F_{ts} = (F_\downarrow + F_\uparrow)/2$, and Equation 2.16a subsequently reads as

$$I_+ = \frac{2}{\pi c_z A^2} \int_D^{D+2A} F_{ts} \frac{z-D-A}{\sqrt{A^2-(z-D-A)^2}} dz. \tag{2.20}$$

The amplitudes commonly used in AM-AFM are considerably larger than the interaction range of the tip–sample force. Consequently, tip–sample forces in the integration range between $D + A$ and $D + 2A$ are insignificant. For this so-called 'large-amplitude approximation' [76, 77], the last term can be expanded at $z \to D$ to $(z-D-A)/\sqrt{A^2-(z-D-A)^2} \approx -\sqrt{A/2(z-D)}$, resulting in

$$I_+ \approx -\frac{\sqrt{2}}{\pi c_z A^{3/2}} \int_D^{D+2A} \frac{F_{ts}}{\sqrt{z-D}} dz. \tag{2.21}$$

By introducing this equation into Equation 2.15a, we obtain the following integral equation:

$$\underbrace{\frac{c_z A^{3/2}}{\sqrt{2}}\left[\frac{a_d\cos(\phi)}{A}-\frac{f_0^2-f_d^2}{f_0^2}\right]}_{\kappa}=\frac{1}{\pi}\int_D^{D+2A}\frac{F_{ts}}{\sqrt{z-D}}dz. \tag{2.22}$$

The left-hand side of this equation contains only experimentally accessible data, and we denote this term as κ. The benefit of these transformations is that the integral equation can be inverted [65, 76] and, as a final result, we find

$$F_{ts}(D)=-\frac{\partial}{\partial D}\int_D^{D+2A}\frac{\kappa(z)}{\sqrt{z-D}}dz. \tag{2.23}$$

It is now straightforward to recover the tip–sample force using Equation 2.23 from a 'spectroscopy experiment' – that is, an experiment where the amplitude and the phase are continuously measured as a function of the actual tip–sample distance $D=d-A$ at a fixed location. With this input, one first calculates κ as a function of D. In a second step, the tip–sample force is computed solving the integral in Equation 2.23 numerically.

Additional information about the tip–sample interaction can be obtained, remembering that the integral I_- is directly connected to the energy dissipation ΔE. By simply combining Equations 1.15b and 1.16b, we get

$$\Delta E=\left(\frac{1}{Q_0}\frac{f_d}{f_0}+\frac{a_d}{A}\sin\phi\right)\pi c_z A^2. \tag{2.24}$$

The same result was found earlier by Cleveland *et al.* [41], using the conservation of energy principle. However, in a further development of Cleveland's investigations we suggest plotting the energy dissipation as a function of the nearest tip–sample distance $D=d-A$ in order to have the same scaling as for the tip–sample force.

An application of the method to experimental data obtained on a silicon wafer is shown in Figure 2.16, where only the data points before the jump were used to reconstruct the tip–sample force and energy dissipation. As a consequence, the experimental force curve showed only the attractive part of the force between tip and sample, with a minimum of -1.8 nN. This result was in agreement with previous studies which stated that the tip sensed only attractive forces before the jump [66, 69, 78].

2.3.2
Frequency-Modulation or Noncontact Mode in Vacuum

In order to obtain high-resolution images with an atomic force microscope, it is very important to prepare clean sample surfaces that are free from unwanted adsorbates. Therefore, these experiments are usually performed in ultra-high vacuum with pressures below 1×10^{-10} mbar. As a consequence, most DFM experiments in

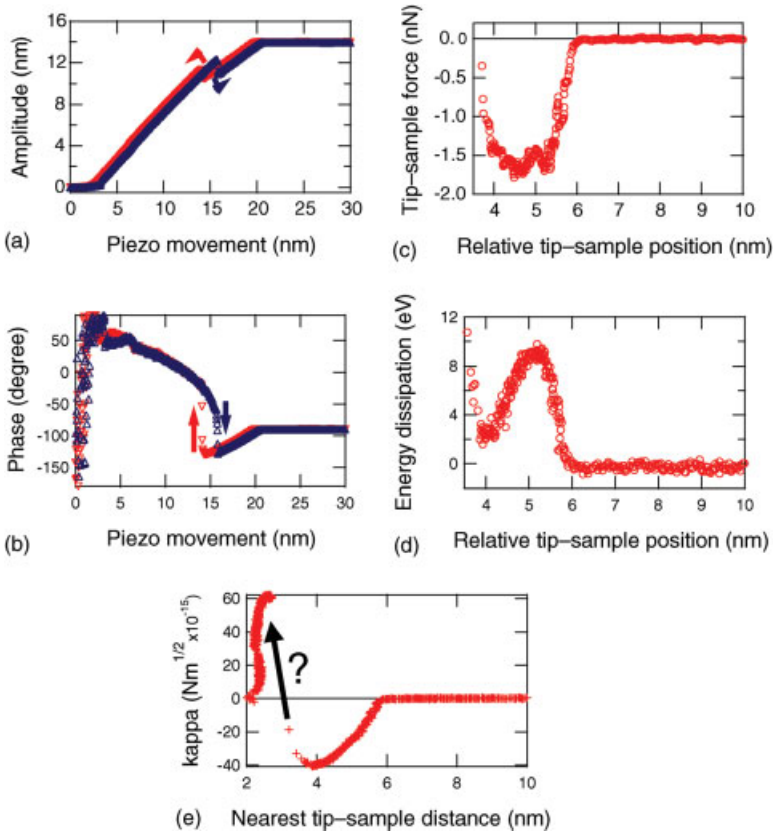


Figure 2.16 Dynamic force spectroscopy experiment on a silicon wafer in air (parameters of the cantilever: $f_d = f_0 = 328.61$ kHz, $c_z = 33.45$ N m^{-1}, $Q_0 = 537$). (a) A measurement of the oscillation amplitude as a function of the oscillation amplitude shows jumps at different positions during approach and retraction; (b) The jumps are also observed in the phase versus distance curves; (c) Using the algorithm described in the text, the tip–sample force can be reconstructed. This curve is calculated from the approach data. Only the data points before the jump are used for reconstruction of the tip–sample force; (d) The energy dissipation per oscillation cycle can be easily obtained from Equation 2.28; (e) This graph shows the $\kappa(D)$-values computed from the amplitude and phase versus distance curves plotted in panels (a) and (b). The jump in these curves results also in a jump in the κ-curve.

vacuum utilize the FM detection scheme introduced by Albrecht *et al.* [33]. In this mode, the cantilever is self-oscillated, in contrast to the AM- or tapping-mode discussed in Section 2.2.3. The FM technique enables the imaging of single point defects on clean sample surfaces in vacuum, and its resolution is comparable with that of the scanning tunneling microscope, while not restricted to conducting surfaces. During the years after the invention of the FM technique the term noncontact atomic force microscopy (NC-AFM) was established, because it is commonly believed that a repulsive, destructive contact between the tip and sample

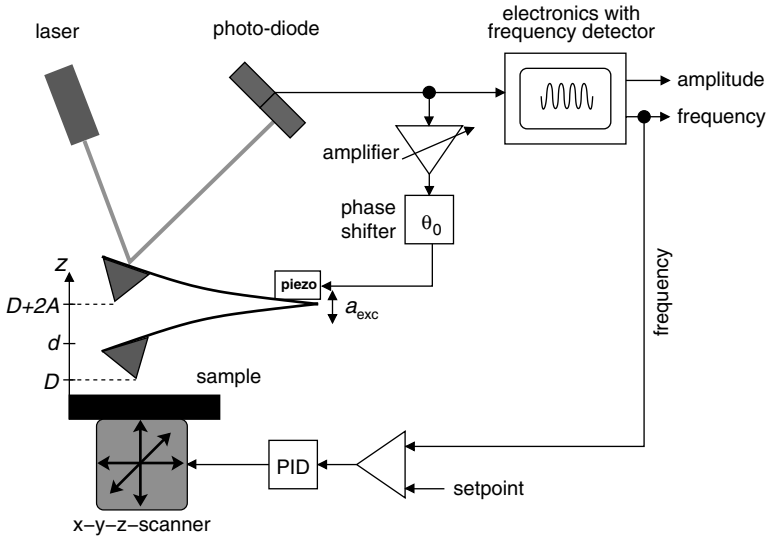


Figure 2.17 The schematic set-up of a dynamic force microscope using the frequency modulation technique. This experimental set-up is often used in UHV. A significant feature is the positive feedback of the self-driven cantilever. The detector signal is amplified and phase-shifted before being used to drive the piezo. The measured quantity is the frequency shift due to tip–sample interaction, which serves as the control signal for the cantilever–sample distance.

is prevented by this technique. In the following subsection we introduce the basic principles of the experimental set-up, explain the origin and calculation of the detected frequency shift, and present applications of this mode.

2.3.2.1 Set-Up of FM-AFM

In vacuum applications, the *Q*-factor of silicon cantilevers is in the range of 10 000 to 30 000. High *Q*-factors, however, limit the acquisition time (bandwidth) of DFM, as the oscillation amplitude of the cantilever requires a long time to adjust. This problem is avoided by the FM-detection scheme based on the specific features of a self-driven oscillator.

The basic set-up of a dynamic force microscope utilizing this driving mechanism is shown schematically in Figure 2.17. The movement of the cantilever is measured with a displacement sensor, after which this signal is fed back into an amplifier with an automatic gain control (AGC); the signal is subsequently used to excite the piezo oscillating the cantilever. The time delay between the excitation signal and cantilever deflection is adjusted by a time ('phase') shifter to a value $t_0 = 1/(4f_0)$, corresponding to $\approx 90^\circ$, as this ensures an oscillation at resonance. Two different modes have been established: (i) the *constant-amplitude mode* [33], where the oscillation amplitude A is kept at a constant value by the AGC; and (ii) the *constant excitation mode* [79], where the excitation amplitude is kept constant. In the following, however, we focus on the constant amplitude mode.

The key feature of the described set-up is the positive feedback loop which oscillates the cantilever always at its resonance frequency, f [39]. The reason for this behavior is that the cantilever serves as the frequency-determining element. This is in contrast to an external driving of the cantilever by a frequency generator, where the driving frequency f_d is not necessarily the resonant frequency of the cantilever.

If the cantilever oscillates near the sample surface, the tip–sample interaction alters its resonant frequency, which is then different from the eigenfrequency f_0 of the free cantilever. The actual value of the resonant frequency depends on the nearest tip–sample distance and the oscillation amplitude. The measured quantity is the *frequency shift* Δf, which is defined as the difference between both frequencies ($\Delta f = f - f_0$). The detection method received its name from the frequency demodulator (FM-detector). The cantilever driving mechanism, however, is independent of this part of the set-up. Other set-ups use a phase-locked loop (PLL) to detect the frequency and to oscillate the cantilever exactly with the frequency measured by the PLL [80].

For imaging, the frequency shift Δf is used to control the cantilever sample distance. Thus, the frequency shift is constant and the acquired data represents planes of constant Δf, which can be related to the surface topography in many cases. The recording of the frequency shift as a function of the tip–sample distance, or alternatively the oscillation amplitude can be used to determine the tip–sample force with high resolution (see Section 2.2.4.5).

2.3.2.2 Origin of the Frequency Shift

Before presenting experimental results obtained in vacuum, we will analyze the origin of the frequency shift. A good insight into the cantilever dynamics is provided by examining the tip potential displayed in Figure 2.18. If the cantilever is far away from the sample surface, the tip moves in a symmetric, parabolic potential (dotted line), and its oscillation is harmonic. In such a case, the tip motion is sinusoidal and the resonance

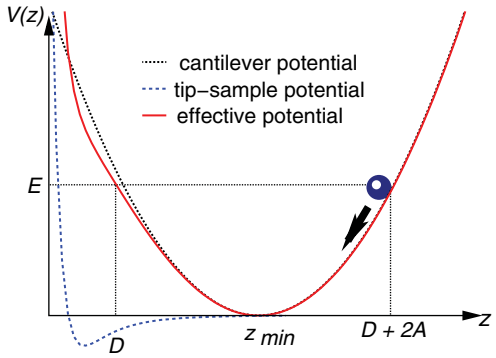


Figure 2.18 The frequency shift in dynamic force microscopy is caused by the tip–sample interaction potential (dashed line), which alters the harmonic cantilever potential (dotted line). Therefore, the tip moves in an anharmonic and asymmetric effective potential (solid line). z_{min} is the minimum position of the effective potential.

frequency is given by the eigenfrequency f_0 of the cantilever. If, however, the cantilever approaches the sample surface, the potential – which determines the tip oscillation – is modified to an effective potential V_{eff} (solid line) given by the sum of the parabolic potential and the tip–sample interaction potential V_{ts} (dashed line). This effective potential differs from the original parabolic potential and shows an asymmetric shape.

As a result of this modification of the tip potential the oscillation becomes anharmonic, and the resonance frequency of the cantilever depends now on the oscillation amplitude A. Since the effective potential experienced by the tip changes also with the nearest distance D, the frequency shift is a functional of both parameters ($\Rightarrow \Delta f := \Delta f(D, A)$).

Figure 2.19 displays some experimental frequency shift versus distance curves for different oscillation amplitudes. These experiments were carried out with

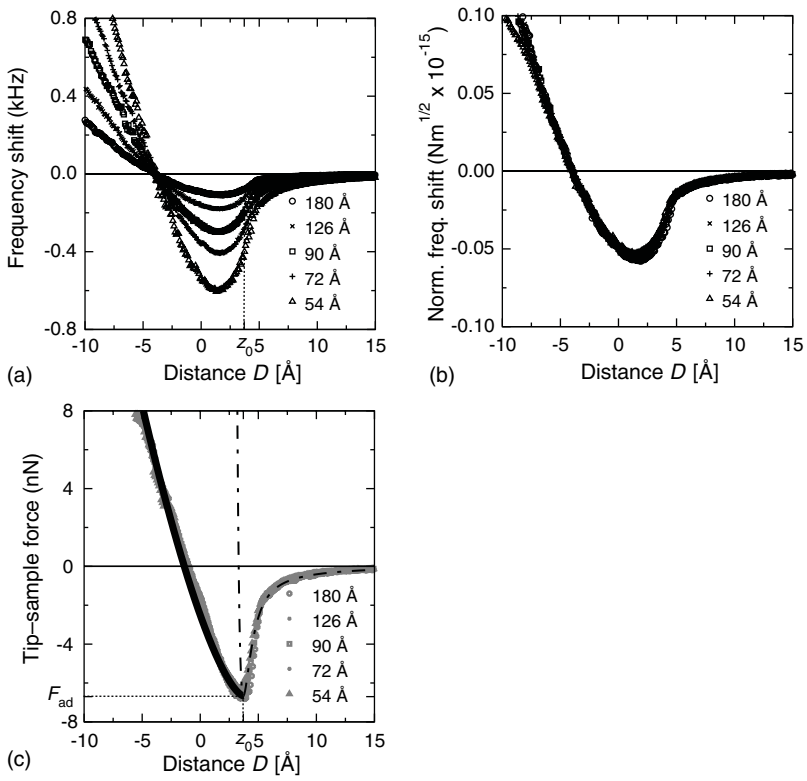


Figure 2.19 (a) Experimental frequency shift versus distance curves acquired with a silicon cantilever ($c_z = 38$ N m^{-1}; $f_0 = 171$ kHz) and a graphite sample for different amplitudes (54–180 Å) in UHV at low temperature ($T = 80$ K). The curves are shifted along the x-axes to make them comparable; (b) Transformation of all frequency shift curves shown in (a) to one universal curve using Equation 2.32. The normalized frequency shift $\gamma(D)$ is nearly identical for all amplitudes; (c) The tip–sample force calculated with the experimental data shown in (a) and (b), using the formula in Equation 2.33. The force F_{ts} (Equation 2.34) is plotted using a dashed-dotted line; the best fit using the force law F_c is displayed by a solid line. The border between 'contact' and 'noncontact' force is marked by the position z_0.

an atomic force microscope designed for operation in UHV and at low temperatures [10].

The obtained experimental frequency shift versus distance curves show a behavior expected from the simple model explained above. All curves show a similar overall shape, but differ in magnitude depending on the oscillation amplitude and the nearest tip–sample distance. During the approach of the cantilever towards the sample surface, the frequency shift decreases and reaches a minimum. With a further reduction of the nearest tip–sample distance, the frequency shift increases again and becomes positive. For smaller oscillation amplitudes, the minimum of the $\Delta f(z)$-curves is deeper and the slope after the minimum is steeper than for larger amplitudes – that is, the overall effect is larger for smaller amplitudes.

This can also be explained by the simple potential model: A decrease in the amplitude A for a fixed nearest distance D moves the minimum of the effective potential closer to the sample surface. Therefore, the relative perturbation of the harmonic cantilever potential increases, which increases also the absolute value of the frequency shift.

2.3.2.3 Theory of FM-AFM

As described in the previous subsection, it is a specific feature of the FM technique that the cantilever is 'self-driven' by a positive feedback loop. Due to this experimental set-up, the corresponding equation of motion is different from the case of the externally driven cantilever discussed in Section 2.2.3. The external driving term must be replaced in order to describe the self-driving mechanism correctly; therefore, the equation of motion is given by

$$m^*\ddot{z}(t) + \frac{2\pi f_0 m^*}{Q}\dot{z}(t) + c_z(z(t)-d) + \underbrace{g c_z(z(t-t_0)-d)}_{\text{driving}} = F_{ts}[z(t), \dot{z}(t)] \tag{2.25}$$

where $z := z(t)$ represents the position of the tip at the time t; and c_z, m and Q are the spring constant, the effective mass and the quality factor of the cantilever, respectively. $F_{ts} = -(\partial V_{ts})/(\partial z)$ is the tip–sample interaction force. The last term on the left describes the active feedback of the system by the amplification of the displacement signal by the *gain factor* g measured at the retarded time $t - t_0$.

The frequency shift can be calculated from the above equation of motion with the ansatz

$$z(t) = d + A\cos(2\pi f t) \tag{2.26}$$

describing the stationary solutions of Equation 2.25. As described in Section 2.2.3, it is assumed that the cantilever oscillations are more or less sinusoidal, such that the tip–sample force F_{ts} is developed into a Fourier-series, as in Equation 2.14. This procedure results in a set of two coupled trigonometric equations [38, 39]:

$$g\cos(2\pi f t_0) = \frac{f^2 - f_0^2}{f_0^2} + I_+ \tag{2.27}$$

$$g\sin(2\pi f t_0) = \frac{1}{Q}\frac{f}{f_0} + I_- \tag{2.28}$$

where the two integrals I_+ Equation 2.16a and I_- Equation 2.16b were defined in accordance to Section 2.2.3.2. These two coupled equations can be solved numerically, if one is interested in the exact dependency of the tip–sample interaction force F_{ts} and the time delay t_0 on the oscillation frequency f and the gain factor g.

Fortunately, a detailed analysis shows that the results of a FM-AFM experiment are mainly determined by the tip–sample force, and only very slightly by the time delay, if t_0 is set to an optimal value before approaching the tip towards the sample surface. These values of the time delay are specific resonance values corresponding to 90° (i.e. $t_0 = 1/4f_0$), and can be easily found by minimizing the gain factor as a function of the time delay. Therefore, it can be assumed that $\cos(2\pi f t_0) \approx 0$ and $\sin(2\pi f t_0) \approx 1$ and the two coupled Equations 1.27 and 1.28 can be decoupled. As a result, an equation for the frequency shift is obtained:

$$\Delta f \cong -\frac{f_0}{2} I_+ = \frac{1}{\pi c_z A^2}\int_{d-A}^{d+A} (F_\downarrow + F_\uparrow)\frac{z-d}{\sqrt{A^2-(z-d)^2}}dz \tag{2.29}$$

and the energy dissipation

$$\Delta E = \left(g - \frac{1}{Q}\frac{f}{f_0}\right)\pi c_z A^2. \tag{2.30}$$

As no assumptions were made about the specific force law of the tip–sample interaction F_{ts}, these equations are valid for any type of interaction as long as the resulting cantilever oscillations remain nearly sinusoidal.

As the amplitudes in FM-AFM are often considerably larger than the distance range of the tip–sample interaction, we can again make the 'large amplitude approximation' [76, 77] and introduce the approximation Equation 2.21 for the integral I_-. This yields the formula

$$\Delta f = \frac{1}{\sqrt{2}\pi}\frac{f_0}{c_z A^{3/2}}\int_D^{D+2A} \frac{F_{ts}(z)}{\sqrt{z-D}}dz \tag{2.31}$$

It is interesting to note that the integral in this equation is virtually independent of the oscillation amplitude. The experimental parameters (c_z, f_0 and A) appear as prefactors. Consequently, it is possible to define the *normalized frequency shift* [77]

$$\gamma(z) := \frac{c_z A^{3/2}}{f_0}\Delta f(z). \tag{2.32}$$

This is a very useful quantity to compare experiments obtained with different amplitudes and cantilevers. The validity of Equation 2.32 is nicely demonstrated by the application of this equation to the frequency shift curves already presented in Figure 2.19a. As shown in Figure 2.19b, all curves obtained for different amplitudes result in one universal γ-curve, which depends only on the actual tip–sample distance, D.

2.3.2.4 Applications of FM-AFM

The initial excitement surrounding the NC-AFM technique in UHV was driven by the first results of Giessibl [81], who was able to image the true atomic structure of the Si(1 1 1)-7 × 7-surface with this technique in 1995. In the same year, Sugawara *et al.* [82] observed the motion of single atomic defects on InP with true atomic resolution. However, imaging on conducting or semi-conducting surfaces is also possible by using scanning tunneling microscopy (STM), and these first NC-AFM images provided no new information on surface properties. The true potential of NC-AFM lies in the imaging of nonconducting surface with atomic precision, which was first demonstrated by Bammerlin *et al.* [83] on NaCl. A longstanding question about the surface reconstruction of the technological relevant material aluminum oxide could be answered by Barth *et al.* [84], who imaged the atomic structure of the high-temperature phase of α-Al_2O_3(0 0 0 1).

The high-resolution capabilities of NC-AFM are nicely demonstrated by the images shown in Figure 2.20. Allers *et al.* [85] resolved atomic steps and defects with atomic resolution on nickel oxide. Today, such a resolution is routinely obtained by various research groups (for an overview, see Refs [3, 34, 35, 86]). Recent efforts have also been concentrated on the analysis of functional organic molecules, since in the field of nanoelectronics it is anticipated that organic molecules in particular will play an important role as the fundamental building blocks of nanoscale electronic device elements. For example, atomic resolution on the highly curved surface of a nanotube [87] was achieved. The analysis of growth properties of thin films [88–90] with respect to their electronic properties has been investigated, while the intramolecular contrast of individual molecules has also been resolved [91], which might be directly related to the internal charge density distribution inside the molecules.

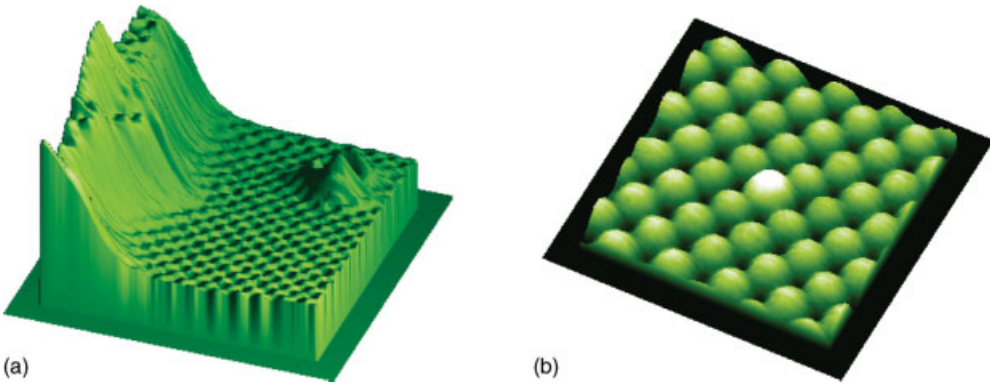


Figure 2.20 Imaging of a NiO(0 0 1) sample surface with a noncontact AFM. (a) Surface step and an atomic defect. The lateral distance between two atoms is 4.17 Å; (b) A dopant atom is imaged as a light protrusion about 0.1 Å higher than the other atoms. (Images courtesy of W. Allers and S. Langkat, University of Hamburg; used with kind permission.)

2.3.2.5 Dynamic Force Spectroscopy

Since its invention in 1986, the atomic force microscope has been used extensively to study tip–sample interactions for various material combinations. Unfortunately, in contact mode such investigations were often hindered close to the sample surface by a 'jump to contact' (see Section 2.2.1.1). In tapping mode, on the other hand, the force analysis is limited due to the instabilities in the amplitude and phase versus distance curves (see Section 2.2.3.3). Such problems, however, are avoided by using large oscillation amplitudes in the FM technique.

In Section 2.2.4.3 it was shown how the frequency shift can be calculated for a given tip–sample interaction law. The inverse problem, however, is even more interesting: *How can the tip-sample interaction be determined from frequency shift data?* Various mathematical solutions to this question have been presented by many research groups [76, 92–97], and this has led to the dynamic force spectroscopy (DFS) technique, which is a direct extension of the FM-AFM mode.

Here, we present the approach of Dürig [76], which is based on the inversion of the integral Equation 2.31 already presented in Section 2.2.4.3. This can be transformed to

$$F_{ts}(D) = \sqrt{2}\frac{c_z A^{3/2}}{f_0}\frac{\partial}{\partial D}\int_D^\infty \frac{\Delta f(z)}{\sqrt{z-D}}dz, \tag{2.33}$$

which allows a direct calculation of the tip–sample interaction force from the frequency shift versus distance curves.

An application of this formula to the experimental frequency shift curves already presented in Section 2.2.4.2 is shown in Figure 2.19c. The obtained force curves are almost identical, despite being obtained with different oscillation amplitudes. As the tip–sample interactions can be measured with high resolution, DFS opens a direct way to compare experiments with theoretical models and predictions.

Giessibl [77] suggested a description of the force between the tip and the sample by combining a long-range (van der Waals) and a short-range (Lennard–Jones) term (see Section 2.1.3). Here, the long-range part describes the van der Waals interaction of the tip, modeled as a sphere with a specific radius, with the surface. The short-range Lennard–Jones term is a superposition of the attractive van der Waals interaction of the last tip apex atom with the surface and the coulombic repulsion. For a tip with radius R, this assumption results in the tip–sample force:

$$F_{nc}(z) = -\frac{A_H R}{6z^2} + \frac{12E_0}{r_0}\left(\left(\frac{r_0}{z}\right)^{13} - \left(\frac{r_0}{z}\right)^{7}\right). \tag{2.34}$$

As this approach does not explicitly consider elastic contact forces between tip and sample, we call this the 'noncontact' force law in the following sections.

A fit of this equation to the experimental tip–sample force curve is shown in Figure 2.19c by a solid line; the obtained parameters are $A_H R = 2.4 \times 10^{-27}$ Jm, $r_0 = 3.4$ Å, and $E_0 = 3$ eV [98]. The regime on the right from the minimum fits well to the experimental data, but the deep and wide minimum of the experimental curves cannot be described accurately with the noncontact force. This is caused by the steep increase in the Lennard–Jones force in the repulsive regime ($\Rightarrow F_{ts} \propto 1/r^{13}$ for $z < r_0$).

The elastic contact behavior can be described with the assumption of the above-described DMT-M model (see Section 2.1.3), that the overall shape of tip and sample changes only slightly until point contact is reached and that, after the formation of this point contact, the tip–sample forces are described by the Hertz theory. A fit of the Hertz model to the experimental data is shown in Figure 2.19 by a solid line. The experimental force curves agree quite well with the contact force law for distances $D < z_0$. This shows that the overall behavior of the experimentally obtained force curves can be described by a combination of long-range (van der Waals), short-range (Lennard–Jones), and contact (Hertz/DMT) forces.

As Equation 2.31 was derived under the assumption that the resonance amplitude is considerably larger than the decay length of the tip–sample interaction, the same restriction applies for Equation 2.33. However, by using more advanced algorithms it is also possible to determine forces from DFS experiments without the large amplitude restriction. The numerical approach of Gotsmann *et al.* [94], as well as the semi-empirical methods of Dürig [92], Giessibl [93] and Sader and Jarvis [97], are applicable in all regimes.

The resolution of DFS can be driven down to the atomic scale. Lantz *et al.* [99] measured frequency shift versus distance curves at different lattice sites of the Si (1 1 1)-(7 × 7) surface, and in this way were able to distinguish differences in the bonding forces between inequivalent adatoms of the 7 × 7 surface reconstruction of silicon.

The concept of DFS can be also extended to three-dimensional (3-D)-force spectroscopy by mapping the complete force field above the sample surface [100]. A schematic of the measurement principle is shown in Figure 2.21a. Frequency shift versus distance curves are recorded on a matrix of points perpendicular to the sample

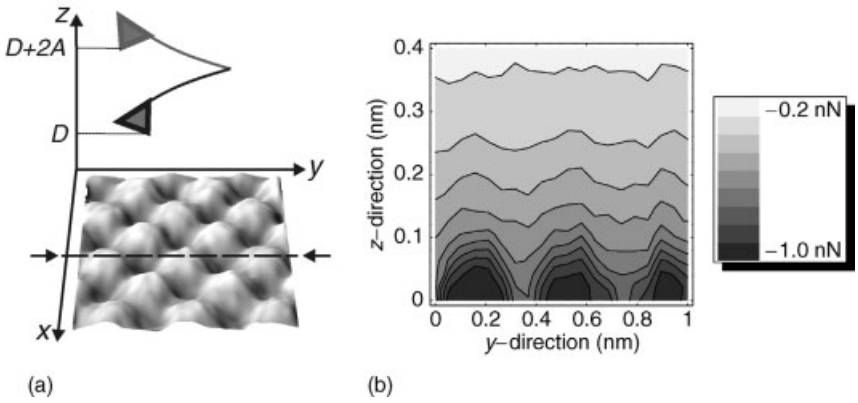


Figure 2.21 (a) Principle of 3-D force spectroscopy. The cantilever oscillates near the sample surface and measure the frequency shift in a *x*–*y*–*z*-box. The 3-D surface shows the topography of the sample (image size: 10 Å × 10 Å) obtained immediately before recording of the spectroscopy field; (b) The reconstructed force field of NiO(0 0 1) shows atomic resolution. The data are recorded along the dotted line shown in (a).

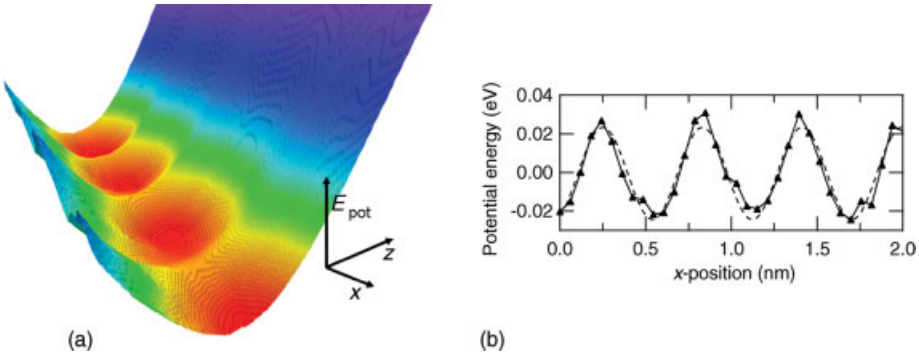


Figure 2.22 (a) A 3-D representation of the interaction energy map determined from 3-D force spectroscopy experiments on a NaCl(1 0 0) crystal surface. The red circular depressions represent the local energy minima; (b) Potential energy profile obtained from (a) by collecting the energy minimum values along the *x*-axis. This curve thus directly reveals the potential energy barrier of $\Delta E_{barrier} = 48$ meV which separates the local energy minima.

surface. By using Equation 2.33, the complete 3-D force field between the tip and sample can be recovered with atomic resolution. Figure 2.21b shows a cut through the force field as a two-dimensional (2-D) map.

The 3D-force technique has been applied also to a NaCl(1 0 0) surface, where not only conservative but also the dissipative tip–sample interaction could be measured in full space [101]. Initially, the forces were measured in the attractive as well as repulsive regime, allowing for the determination of the local minima in the corresponding potential energy curves (Figure 2.22). This information is directly related to the atomic energy barriers responsible for a multitude of dynamic phenomena in surface science, such as diffusion, faceting and crystalline growth. The direct comparison of conservative with the simultaneously acquired dissipative processes furthermore allowed determining atomic-scale mechanical relaxation processes.

If the NC-AFM is capable of measuring forces between single atoms with sub-nN precision, why should it not be possible to also exert forces with this technique? In fact, the new and exciting field of nanomanipulation would be driven to a whole new dimension, if defined forces could be reliably applied to single atoms or molecules. In this respect, Loppacher *et al.* [102] were able to exert pressure on different parts of an isolated Cu–TBBP molecule, which is known to possess four rotatable legs. Here, the force–distance curves were measured while one of the legs was pushed by the AFM tip and turned by 90°, and hence were able to measure the energy which was dissipated during the 'switching' of this molecule between different conformational states. The manipulation of single silicon atoms with NC-AFM was demonstrated by Oyabu *et al.* [103], who removed single atoms from a Si(1 1 1)-7 × 7 surface with the AFM tip and were able subsequently to re-deposit atoms from the tip onto the surface. This approach was driven to its limits by Sugimoto *et al.*, who manipulated single Sn-

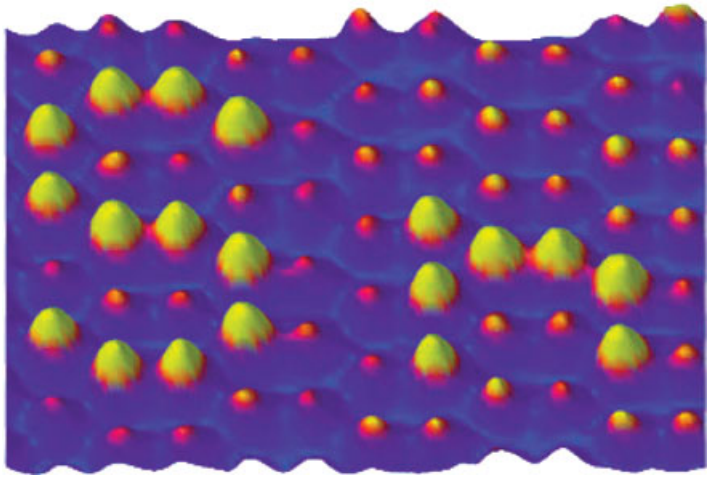


Figure 2.23 Final topographic NC-AFM image of the process of rearranging single atoms at room temperature. The image was acquired with a cantilever oscillation amplitude of 15.7 nm, using a Si cantilever. (Reproduced from Ref. [104].)

atoms on the Ge(1 1 1)-c(2 × 8) semiconductor surface. By pushing single Sn-atoms from one lattice site to another, they finally succeeded in writing the term 'Sn' with single atoms (Figure 2.23).

2.4 Summary

In summary, we have presented an overview over the basic principles and modern applications of AFM. This versatile technique can be categorized into two operational modes, static and dynamic. The static mode allows the simultaneous measurement of normal and lateral forces, thus yielding direct information about friction mechanisms of nanoscale contacts. The main advantage of the dynamic mode is the possibility to control tip–sample distances while avoiding the undesirable and destructive 'jump-to-contact' phenomenon. Two different excitation schemes for dynamic force microscopy were introduced, where the amplitude-modulation or tapping mode are in particular well-suited to high-resolution imaging under ambient or liquid conditions. The ultimate 'true' atomic resolution, however, is limited to vacuum conditions using FM or noncontact techniques. Nonetheless, the impact of AFM reaches far beyond the high-resolution imaging of surface topography: DFS allows the quantification of tip–sample forces, through the systematic acquisition of parameters such as amplitude, phase and oscillation frequency as a function of the relative tip–sample distance. Based on this approach, not only the bonding force of single interatomic chemical bonds can be measured, but also the full 3-D force field can be determined, at atomic resolution. Finally, the finding that atomic forces can

not only be measured but also exerted with atomic precision will open up the new and exciting field of nanomanipulation.

Acknowledgments

The authors would like to thank all colleagues who contributed to these studies with their images and experimental results, including Boris Anczykowski, Daniel Ebeling, Jan-Erik Schmutz, Domenique Weiner (University of Münster), Wolf Allers, Shenja Langkat, Alexander Schwarz (University of Hamburg) and Udo D. Schwarz (Yale University).

References

1 Binnig, G., Quate, C.F. and Gerber, Ch. (1986) Atomic force microscopy. *Physical Review Letters*, **56**, 930–933.

2 Binnig, G., Rohrer, H., Gerber, C. and Weibel, E. (1982) Surface studies by scanning tunneling microscopy. *Physical Review Letters*, **49**, 57–61.

3 Meyer, E., Hug, H.-J. and Bennewitz, R. (2004) *Scanning Probe Microscopy – The Lab on a Tip*, Springer-Verlag.

4 Bhushan, B. and Marti, O.(eds) (2005) Scanning probe microscopy – principle of operation, instrumentation, and probes, in *Nanotribology and Nanomechanics – An Introduction*, (ed. B. Bhushan) Springer-Verlag, Berlin Heidelberg, pp. 41–115.

5 Lüthi, R., Meyer, E., Haefke, H., Howald, L., Gutmannsbauer, W., Guggisberg, M., Bammerlin, M. and Güntherodt, H.-J. (1995) Nanotribology: an UHV-SFM study on thin films of C60 and AgBr. *Surface Science*, **338**, 247–260.

6 Neumeister, J.M. and Ducker, W.A. (1994) Lateral, normal, and longitudinal spring constants of atomic force microscopy cantilevers. *Review of Scientific Instruments*, **65**, 2527–2531.

7 Sader, J.E. (1995) Parallel beam approximation for V-shaped atomic force microscope cantilevers. *Review of Scientific Instruments*, **66**, 4583–4587.

8 Alexander, S., Hellemans, L., Marti, O., Schneir, J., Elings, V. and Hansma, P.K. (1988) An atomic-resolution atomic-force microscope implemented using an optical lever. *Journal of Applied Physics*, **65**, 164–167.

9 Meyer, G. and Amer, N.M. (1988) Novel optical approach to atomic force microscopy. *Applied Physics Letters*, **53**, 1045–1047.

10 Allers, W., Schwarz, A., Schwarz, U.D. and Wiesendanger, R. (1998) A scanning force microscope with atomic resolution in ultrahigh vacuum and at low temperatures. *Review of Scientific Instruments*, **69**, 221–225.

11 Kawakatsu, H., Kawai, S., Saya, D., Nagashio, M., Kobayashi, D., Toshiyoshi, H. and Fujita, H. (2002) Towards atomic force microscopy up to 100 MHz. *Review of Scientific Instruments*, **73** (6), 2317–2320.

12 Moser, A., Hug, H.-J., Jung, T., Schwarz, U.D. and Guntherodt, H.-J. (1993) A miniature fibre optic force microscope scan head. *Measurement Science and Technology*, **4**, 769–775.

13 Rugar, D., Mamin, H.J. and Guethner, P. (1989) Improved fiber-optic interferometer for atomic force microscopy. *Applied Physics Letters*, **55**, 2588–2590.

14 Linnemann, R., Gotszalk, T., Rangelow, I.W., Dumania, P. and Oesterschulze, E. (1996) Atomic force microscopy and lateral force microscopy using piezoresistive cantilevers. *Journal of Vacuum Science & Technology B*, **14** (2), 856–860.

15 Tortonese, M., Barrett, R.C. and Quate, C.F. (1993) Atomic resolution with an atomic force microscope using piezoresistive detection. *Applied Physics Letters*, **62**, 834–836.

16 Yuan, C.W., Batalla, E., Zacher, M., de Lozanne, A.L., Kirk, M.D. and Tortonese, M. (1994) Low temperature magnetic force microscope, utilizing a piezoresistive cantilever. *Applied Physics Letters*, **65**, 1308–1310.

17 Stahl, U., Yuan, C.W., de Lozanne, A.L. and Tortonese, M. (1994) Atomic force microscope using piezoresistive, cantilevers and combined with a scanning electron microscope. *Applied Physics Letters*, **65**, 2878–2880.

18 Israelachvili, J.N. (1992) *Intermolecular and Surface Forces*, Academic Press, London.

19 Stifter, Th., Marti, O. and Bhushan, B. (2000) Theoretical investigation of the distance dependence of capillary and van der Waals forces in scanning force microscopy. *Physical Review B - Condensed Matter*, **62**, 13667–13673.

20 Johnson, K.L. (1985) *Contact Mechanics*, Cambridge University Press, Cambridge, UK.

21 Landau, L.D. and Lifschitz, E.M. (1991) *Lehrbuch der theoretischen Physik VII: Elastizitätstheorie*, Akademie-Verlag, Berlin.

22 Johnson, K.L., Kendall, K. and Roberts, A.D. (1971) Surface energy and contact of elastic solids. *Proceedings of the Royal Society of London Series A - Mathematical, Physical and Engineering Sciences*, **324**, 301.

23 Derjaguin, B.V., Muller, V.M. and Toporov, Y.P. (1975) Effect of contact deformations on the adhesion of particles. *Journal of Colloid and Interface Science*, **53**, 314–326.

24 Schwarz, U.D. (2003) A generalized analytical model for the elastic deformation of an adhesive contact between a sphere and a flat surface. *Journal of Colloid and Interface Science*, **261**, 99–106.

25 Sarid, D. (1994) *Scanning Force Microscopy – With Applications to Electric, Magnetic, and Atomic Forces*, Oxford University Press.

26 Mate, C.M., McClelland, G.M., Erlandsson, R. and Chiang, S. (1987) Atomic-scale friction of a tungsten tip on a graphite surface. *Physical Review Letters*, **59**, 1942–1945.

27 McKendry, R., Theoclitou, M.-E., Rayment, T. and Abell, C. (1998) Chiral discrimination by chemical force microscopy. *Nature*, **391**, 566–568.

28 Overney, R.M., Meyer, E., Frommer, J., Brodbeck, D., Lüthi, R., Howald, L., Güntherodt, H.-J., Fujihira, M., Takano, H. and Gotoh, Y. (1992) Friction measurements on phase-separated thin films with a modified atomic force microscope. *Nature*, **359**, 133–135.

29 Burnham, N.A. and Colton, R.J. (1989) Measuring the nanomechanical properties and surface forces of materials using an atomic force microscopy. *Journal of Vacuum Science and Technology A - Vacuum Surfaces and, Films*, **7**, 2906.

30 Martin, Y., Williams, C.C. and Wickramasinghe, H.K. (1987) Atomic force microscope–force mapping and profiling on a sub 100-Å scale. *Journal of Applied Physics*, **61**, 4723–4729.

31 Putman, C.A.J., Vanderwerf, K.O., Degrooth, B.G., Vanhulst, N.F. and Greve, J. (1994) Tapping mode atomic force microscopy in liquid. *Applied Physics Letters*, **64**, 2454–2456.

32 Zhong, Q.D., Inniss, D., Kjoller, K. and Elings, V.B. (1993) Fractured polymer/silica fiber surface studied by tapping mode atomic force microscopy. *Surface Science Letters*, **290**, L688–L692.

33 Albrecht, T.R., Grütter, P., Horne, D. and Rugar, D. (1991) Frequency modulation detection using high-Q cantilevers for enhanced force microscope sensitivity. *Journal of Applied Physics*, **69**, 668–673.

34 Garcia, R. and Pérez, R. (2002) Dynamic atomic force microscopy methods. *Surface Science Reports*, **47**, 197–301.

35 Giessibl, F.-J. (2003) Advances in atomic force microscopy. *Reviews of Modern Physics*, **75**, 949–983.

36 Hölscher, H. and Schirmeisen, A. (2005) Dynamic force microscopy and spectroscopy, in *Advances in Imaging and Electron Physics*, (ed. P.W. Hawkes), Academic Press Ltd, London, pp. 41–101.

37 Morita, S., Wiesendanger, R. and Meyer, E.(eds) (2002) *Noncontact Atomic Force Microscopy*, Springer-Verlag, Berlin.

38 Hölscher, H., Gotsmann, B., Allers, W., Schwarz, U.D., Fuchs, H. and Wiesendanger, R. (2001) Measurement of conservative and dissipative tip-sample interaction forces with a dynamic force microscope using the frequency modulation technique. *Physical Review B - Condensed Matter*, **64**, 075402.

39 Hölscher, H., Gotsmann, B., Allers, W., Schwarz, U.D., Fuchs, H. and Wiesendanger, R. (2002a) Comment on "damping mechanism in dynamic force microscopy". *Physical Review Letters*, **88**, 019601.

40 Schönenberger, C. and Alvarado, S.F. (1989) A differential interferometer for force microscopy. *Review of Scientific Instruments*, **60**, 3131–3134.

41 Cleveland, J.P., Anczykowski, B., Schmid, A.E. and Elings, V.B. (1998) Energy dissipation in tapping-mode atomic force microscopy. *Applied Physics Letters*, **72**, 2613.

42 Garcia, R., Gómez, C.J., Martinez, N.F., Patil, S., Dietz, C. and Magerle, R. (2006) Identification of nanoscale dissipation processes by dynamic atomic force microscopy. *Physical Review Letters*, **97**, 016103.

43 Tamayo, J. and García, R. (1998) Relationship between phase shift and energy dissipation in tapping-mode scanning force microscopy. *Applied Physics Letters*, **73**, 2926–2928.

44 Maganov, S. (2004) Visualization of polymer structures with atomic force microscopy, in *Applied Scanning Probe Methods*, (eds H. Fuchs M. Hosaka and B. Bhushan) Springer-Verlag, pp. 207–250.

45 Klinov, D. and Maganov, S. (2004) True molecular resolution in tapping-mode atomic force microscopy with high-resolution probes. *Applied Physics Letters*, **84**, 2697–2698.

46 Möller, C., Allen, M., Elings, V., Engel, A. and Müller, D.J. (1999) Tapping-mode atomic force microscopy produces faithful high-resolution images of protein surfaces. *Biophysical Journal*, **77**, 1150–1158.

47 Lee, S.I., Howell, S.W., Raman, A. and Reifenberger, R. (2002) Nonlinear dynamics of microcantilevers in tapping mode atomic force microscopy: a comparison between theory and experiment. *Physical Review B - Condensed Matter*, **66**, 115409.

48 Rodríguez, T.R. and García, R. (2002) Tip motion in amplitude modulation (tapping-mode) atomic-force microscopy: Comparison between continuous and point-mass models. *Applied Physics Letters*, **80**, 1646–1648.

49 Stark, R.W. and Heckl, W. (2000) Fourier transformed atomic force microscopy: tapping mode atomic force microscopy beyond the hookian approximation. *Surface Science*, **457**, 219–228.

50 Stark, R.W., Schitter, G., Stark, M., Guckenheimer, R. and Stemmer, A. (2004) State-space model of freely vibrating and surface-coupled cantilever dynamics in atomic force microscopy. *Physical Review B - Condensed Matter*, **69**, 085412.

51 Legleiter, J. and Kowalewski, T. (2005) Insight into fluid tapping-mode atomic force microscopy provided by numerical

simulations. *Applied Physics Letters*, **87**, 163120.

52 Legleiter, J., Park, M., Cusick, B. and Kowalewski, T. (2006) Scanning probe acceleration microscopy (SPAM) in fluids: Mapping mechanical properties of surfaces at the nanoscale. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 4813–4818.

53 Hölscher, H., Ebeling, D. and Schwarz, U.D. (2006) Theory of Q-controlled dynamic force microscopy in air. *Journal of Applied Physics*, **99**, 084311.

54 Sahin, O., Quate, C.F., Solgaard, O. and Atalar, A. (2004) Resonant harmonic response in tapping-mode atomic force microscopy. *Physical Review B - Condensed Matter*, **69**, 165416.

55 Dürig, U. (2000a) Interaction sensing in dynamic force microscopy. *New Journal of Physics*, **2**, 5.1.

56 Sader, J.E., Uchihashi, T., Farrell, A., Higgins, M.J., Nakayama, Y. and Jarvis, S.P. (2005) Quantitative force measurements using frequency modulation atomic force microscopy – theoretical foundations. *Nanotechnology*, **16**, S94–101.

57 Aimé, J.P., Boisgard, R., Nony, L. and Couturier, G. (1999) Nonlinear dynamic behavior of an oscillating tip-microlever system and contrast at the atomic scale. *Physical Review Letters*, **82**, 3388–3391.

58 Gleyzes, P., Kuo, P.K. and Boccara, A.C. (1991) Bistable behavior of a vibrating tip near a solid surface. *Applied Physics Letters*, **58**, 2989–2991.

59 Kühle, A., Sorensen, A. and Bohr, J. (1998a) Role of attractive forces in tapping tip force microscopy. *Journal of Applied Physics*, **81**, 6562–6569.

60 Nony, L., Boisgard, R. and Aimé, J.-P. (2001) Stability criterions of an oscillating tip-cantilever system in dynamic force microscopy. *European Physical Journal B*, **24**, 221–229.

61 San Paulo, A. and García, R. (2002) Unifying theory of tapping-mode atomic-force microscopy. *Physical Review B - Condensed Matter*, **66**, 041406.

62 Sasaki, Naruo and Tsukada, Masaru (1999) Theory for the effect of the tip-surface interaction potential on atomic resolution in forced vibration system of noncontact AFM. *Applied Surface Science*, **140**, 339–343.

63 Wang, L. (1998) Analytical descriptions of the tapping-mode atomic force microscopy response. *Applied Physics Letters*, **73**, 3781–3783.

64 Hoppenstaedt, Frank C. (1993) *Analysis and Simulation of Chaotic Systems*, Springer-Verlag, New York.

65 Landau, L.D. and Lifschitz, E.M. (1990) *Lehrbuch der Theoretischen Physik: Mechanik*, Akademie-Verlag, Berlin.

66 Anczykowski, B., Krüger, D. and Fuchs, H. (1996) Cantilever dynamics in quasinoncontact force microscopy: Spectroscopic aspects. *Physical Review B - Condensed Matter*, **53**, 15485–15488.

67 Haugstad, G. and Jones, R.R. (1999) Mechanisms of dynamic force microscopy on polyvinyl alcohol: region-specific non-contact and intermittent contact regimes. *Ultramicroscopy*, **67**, 77–86.

68 Stark, R.W., Schitter, G. and Stemmer, A. (2003) Tuning the interaction forces in tapping mode atomic force microscopy. *Physical Review B - Condensed Matter*, **68**, 085401.

69 García, R. and Paulo, A.S. (2000) Dynamics of a vibrating tip near or in intermittent contact with a surface. *Physical Review B - Condensed Matter*, **61**, R13381–R13384.

70 San Paulo, A. and García, R. (2000) High-resolution imaging of antibodies by tapping-mode atomic force microscopy: Attractive and repulsive tip-sample interaction regimes. *Biophysical Journal*, **78**, 1599–1605.

71 Tamayo, J. and García, R. (1996) Deformation, contact time, and phase contrast in tapping mode scanning force microscopy. *Langmuir*, **12**, 4430–4435.

72 Zitzler, L., Herminghaus, S. and Mugele, F. (2002) Capillary forces in tapping mode atomic force microscopy. *Physical Review B - Condensed Matter*, **66**, 155436.

73 Hölscher, H. (2006) Quantitative measurement of tip-sample interactions in amplitude modulation atomic force microscopy. *Applied Physics Letters*, **89**, 123109.

74 Lee, M. and Jhe, W. (2006) General solution of amplitude-modulation atomic force microscopy. *Physical Review Letters*, **97**, 036104.

75 Stark, M., Stark, R.W., Heckl, W.M. and Guckenberger, R. (2002) Inverting dynamic force microscopy: From signals to time,-resolved interaction forces. *Proceedings of the National Academy of Sciences of the United States of, America*, **99**, 8473–8478.

76 Dürig, U. (1999) Relations between interaction force and frequency shift in large-amplitude dynamic force microscopy. *Applied Physics Letters*, **75**, 433–435.

77 Giessibl, F.J. (1997) Forces and frequency shifts in atomic-resolution dynamic-force microscopy. *Physical Review B - Condensed Matter*, **56**, 16010–16015.

78 Kühle, A., Sørensen, A.H., Zandbergen, J.B. and Bohr, J. (1998b) Contrast artifacts in tapping tip atomic force microscopy. *Applied Physics A: Materials Science & Processing*, **66**, S329–S332.

79 Ueyama, H., Sugawara, Y. and Morita, S. (1998) Stable operation mode for dynamic noncontact atomic force microscopy. *Applied Physics A: Materials Science & Processing*, **66**, S295–S297.

80 Loppacher, Ch., Bammerlin, M., Battiston, F., Guggisberg, M., Müller, D., Lüthi, R., Hidber, H.R., Meyer, E. and Güntherodt, H.-J. (1998) Fast digital electronics for application in dynamic force microscopy using high-q cantilevers. *Applied Physics A: Materials Science & Processing*, **66**, S215–S218.

81 Giessibl, F.-J. (1995) Atomic resolution of the silicon (111)-(7 × 7) surface by atomic force microscopy. *Science*, **267**, 68.

82 Sugawara, Y., Otha, M., Ueyama, H. and Morita, S. (1995) Defect motion on an InP (110) surface observed with noncontact atomic force microscopy. *Science*, **270**, 1646.

83 Bammerlin, M., Lüthi, R., Meyer, E., Baratoff, A., Lü, J., Guggisberg, M., Gerber, Ch., Howald, L. and Güntherodt, H.-J. (1997) True atomic resolution on the surface of an insulator via ultrahigh vacuum dynamic force microscopy. *Probe Microscopy*, **1**, 3.

84 Barth, C. and Reichling, M. (2001) Imaging the atomic arrangement on the high-temperature reconstructed α-Al_2O_3(0001) surface. *Nature*, **414**, 54–57.

85 Allers, W., Langkat, S. and Wiesendanger, R. (2001) Dynamic low-temperature scanning force microscopy on nickel oxide (001). *Applied Physics A: Materials Science & Processing*, **72**, S27.

86 Morita, S. and Sugawara, Y. (2002) *Noncontact Atomic Force Microscopy*, (eds S. Morita, R. Wiesendanger and E. Meyer), Springer-Verlag, Heidelberg, Germany. pp. 47–77.

87 Ashino, M., Schwarz, A., Behnke, T. and Wiesendanger, R. (2004) Atomic-resolution dynamic force microscopy and spectroscopy of a single-walled carbon nanotube: Characterization of interatomic van der Waals forces. *Physical Review Letters*, **93**, 136101.

88 Burke, S.A., Mativetsky, J.M., Hoffmann, R. and Grütter, P. (2005) Nucleation and submonolayer growth of C60 on KBr. *Physical Review Letters*, **94**, 096102.

89 Kunstmann, T., Schlarb, A., Fendrich, M., Wagner, Th., Möller, R. and Hoffmann, R. (2005) Dynamic force microscopy study of 3,4,9,10-perylenetetracarboxylic dianhydride on KBr(001). *Physical Review B - Condensed Matter*, **71**, 121403.

90 Loppacher, Ch., Zerweck, U., Eng, L.M., Gemming, S., Seifert, G., Olbrich, C., Morawetz, K. and Schreiber, M. (2006)

Adsorption of PTCDA on a partially KBr covered Ag(111) substrate. *Nanotechnology*, **17**, 1568.

91 Such, B., Weiner, D., Schirmeisen, A. and Fuchs, H. (2006) Influence of the local adsorption environment on the intramolecular contrast of organic molecules in non-contact atomic force microscopy. *Applied Physics Letters*, **89**, 093104.

92 Dürig, U. (2000b) Extracting interaction forces and complementary observables in dynamic probe microscopy. *Applied Physics Letters*, **76**, 1203–1205.

93 Giessibl, F.-J. (2001) A direct method to calculate tip-sample forces from frequency shifts in frequency-modulation atomic force microscopy. *Applied Physics Letters*, **78**, 123–125.

94 Gotsmann, B., Ancykowski, B., Seidel, C. and Fuchs, H. (1999) Determination of tip-sample interaction forces from measured dynamic force spectroscopy curves. *Applied Surface Science*, **140**, 314–319.

95 Hölscher, H., Gotsmann, B. and Schirmeisen, A. (2003) On dynamic force spectroscopy using the frequency modulation technique with constant excitation. *Physical Review B - Condensed Matter*, **68**, 153401.

96 Hölscher, H., Schwarz, U.D. and Wiesendanger, R. (1999) Calculation of the frequency shift in dynamic force microscopy. *Applied Surface Science*, **140**, 344–351.

97 Sader, J.E. and Jarvis, S.P. (2004) Accurate formulas for interaction force and energy in frequency modulation force spectroscopy. *Applied Physics Letters*, **84**, 1801–1803.

98 Hölscher, H., Schwarz, A., Allers, W., Schwarz, U.D. and Wiesendanger, R. (2000) Quantitative analysis of dynamic force spectroscopy data on graphite(0001) in the contact and non-contact regime. *Physical Review B - Condensed Matter*, **61**, 12678–12681.

99 Lantz, M.A., Hug, H., Hoffmann, R., van Schendel, P.J.A., Kappenberger, P., Martin, S., Baratoff, A. and Güntherodt, H.-J. (2001) Quantitative measurement of short-range, chemical bonding forces. *Science*, **291**, 2580–2583.

100 Hölscher, H., Langkat, S.M., Schwarz, A. and Wiesendanger, R. (2002b) Measurement of three-dimensional force fields with atomic resolution using dynamic force spectroscopy. *Applied Physics Letters*, **81**, 4428–4430.

101 Schirmeisen, A., Weiner, D. and Fuchs, H. (2006) Single-atom contact mechanics: From atomic scale energy barrier to mechanical relaxation hysteresis. *Physical Review Letters*, **97**, 136101.

102 Loppacher, Ch., Guggisberg, M., Pfeiffer, O., Meyer, E., Bammerlin, M., Luthi, R., Schlittler, R., Gimzewski, J.K., Tang, H. and Joachim, C. (2003) Direct determination of the energy required to operate a single molecule switch. *Physical Review Letters*, **90**, 066107.

103 Oyabu, N., Custance, O., Yi, I., Sugawara, Y. and Morita, S. (2003) Mechanical vertical manipulation of selected single atoms by soft nanoindentation using near contact atomic force microscopy. *Physical Review Letters*, **90**, 176102.

104 Sugimoto, Y., Abe, M., Hirayama, S., Oyabu, N., Custance, O. and Morita, S. (2005) Atom inlays performed at room temperature using atomic force microscopy. *Nature Materials*, **4**, 156–159.

3
Probing Hydrodynamic Fluctuations with a Brownian Particle

Sylvia Jeney, Branimir Lukic, Camilo Guzman, and László Fórró

3.1
Introduction

The observation of Brownian motion has been a subject of interest since the invention of optical microscopy during the seventeenth century [1]. From then on, the understanding of its origin remained a subject of debate until 1905, when Einstein described a convincing model which, assumed that the fluctuations of a small-sized particle floating in a fluid were caused by momentum transfer from thermally excited fluid molecules. Einstein identified the mean square displacement of the particle as *the* characteristic experimental observable of Brownian motion, and showed that it grows linearly with time as $\langle \Delta x^2(t) \rangle = 2Dt$, thereby introducing the diffusion coefficient D [2]. In 1908, Langevin reformulated Newton's force balance equation by adding to the instantaneous Stokes' friction [3] a stochastic force term, representing the random impacts of surrounding medium molecules on the Brownian particle [4]. At the same time, Henri pointed out the limited nature of Stokes' formula for the friction force, when applied to neutrally buoyant, micron-sized particles [5], which is the case for most Brownian particles used in experiments. In such cases, correlations between friction and velocity are non-instantaneous, and the positions of the particle are expected to be correlated up to longer times. The origin of this effect comes from the non-negligible inertia of the fluid, and this must be accounted for in the description of Brownian motion. The expression for the mean square displacement using the noninstantaneous friction force was introduced by Vladimirsky and Terletzky in 1945 [6], but remained largely ignored, as their contribution was published in Russian. In 1967, Alder and Wainwright discovered, in numerical simulations, that the particle's velocity correlations (another characteristic observable of Brownian motion) display a power-law decay [7] instead of an exponential relaxation, as expected for instantaneous friction. These simulations led theoreticians during the 1970s to reconsider the contribution of fluid mechanics to Brownian motion [8–14], and to address its relevance in experiments. The idea of using a particle subjected to Brownian motion as a reporter of its local environment was settled. With this approach, any deviation

Nanotechnology: Volume 6: Nanoprobes. Edited by Harald Fuchs

ISBN: 978-3-527-31733-2

from the normal diffusive behavior of the particle could be interpreted as a response to the material properties of its complex environment [15, 16]. To measure such behavior, a high spatial resolution down to the nanometer scale is needed. Experiments using dynamic light scattering in colloidal suspensions confirmed that the diffusion of colloidal particles is influenced by fluid mechanics, and hence is time-dependent [17–21]. However, in order to achieve a sufficiently high resolution, averaging over an ensemble of different particles was necessary. Nowadays, tracking a *single* particle in a fluid on time scales sufficiently short to detect hydrodynamic contributions can be realized by using optical tracking interferometry (OTI). This allows a *direct* measurement of Brownian motion at the same resolution as techniques averaging over many particles, and an individual particle comes to be the local Brownian probe. OTI utilizes a weak optical trap [22] and interferometric particle position detection. The trapping laser provides a light source for the position detection of the particle, and at the same time ensures that the particle remains within the detector range.

In this chapter we provide a complete picture of the measurements of a Brownian particle immersed in a viscous fluid and held by an optical trap. First, relevant theoretical insights are exposed and the different timescales of Brownian motion are summarized. Next, the technical aspects of OTI are described, and methodologies on data acquisition, analysis and interpretation provided. The influences of experimentally relevant parameters, such as the trapping force constant, the fluid properties and the Brownian particle itself, are presented. Finally, the overlap of the different measurable time scales of Brownian motion is quantified, and the consequences are discussed.

3.2
Theoretical Model of Brownian Motion in an Optical Trap

In general, the Brownian motion of a particle in a fluid is the result of thermal fluctuations of the surrounding fluid molecules. Collisions between a bath of fluid molecules at temperature T and the particle lead to an exchange of energy that allows the establishment of a thermal equilibrium between the particle and its environment. A simple model system to describe the phenomenon quantitatively in terms of statistical mechanics *as well as* hydrodynamics consists of a micrometer-sized sphere immersed in a viscous, so-called Newtonian fluid. Such a system can typically be observed in OTI. In this section, theoretical predictions on the motion of a Brownian sphere in a harmonic potential are discussed, starting from the Langevin equation, and including effects arising from hydrodynamic interactions with the viscous fluid. The section ends with an overview of different relaxation times related to the Brownian particle, the fluid and the optical trap.

3.2.1
The General Langevin Equation for a Brownian Sphere in an Incompressible Fluid

There are principally two, counteracting forces that govern the motion of a Brownian particle. First, the particle is driven through the thermal force $F_{th}(t)$, that arises from

random fluctuations of the fluid molecules excited by the thermal energy k_BT. Second, $F_{th}(t)$ is resisted by the friction force $F_{fr}(t)$, which (over-)damps the motion of the fluctuating particle. $F_{fr}(t)$ is the force exerted on the Brownian sphere by the surrounding viscous fluid, when the fluid is perturbed through the particle's fluctuations. Following from Newton's second law, the equation of motion can be written as the generalized Langevin equation:

$$m_s\ddot{x}(t) = F_{th}(t) + F_{fr}(t) + F_{ex}(t), \tag{3.1}$$

where m_s is the inertia of the Brownian sphere (s referring to the sphere's parameters), and $F_{ex}(t)$ represents the sum of any external forces, such as gravity or the force of the optical trap. For simplicity, we will discuss only the one-dimensional case for the axis x, even though OTI measurements give access to all three directions x, y and z.

3.2.1.1 The Random Thermal Force $F_{th}(t)$

The random force $F_{th}(t)$ represents the collisions of the fluid molecules on the particle. Its contribution in a homogeneous and isotropic medium varies so rapidly compared to the observable time scales, that $F_{th}(t)$ should be, on average, zero; $\langle F_{th}(t)\rangle = 0$.

Furthermore, as a very large number of collisions occurs during two successive measurements at times t and t', the correlation time of $F_{th}(t)$ is much shorter than the time interval between the two measurements [10]:

$$\langle F_{th}(t)F_{th}(t')\rangle = 2k_BT\gamma\langle\xi(t)\xi(t')\rangle$$

where k_B is the Boltzmann constant, γ is the viscous drag of the fluid on the particle and $\zeta(t)$ is a white noise term with no finite correlation time: $\langle\xi(t)\xi(t')\rangle = \delta(t-t')$.

$F_{th}(t)$ obeys the fluctuation–dissipation theorem, and has the expression:

$$F_{th}(t) = \sqrt{2k_BT\gamma}\xi(t) \tag{3.2}$$

In real systems the correlation time will not typically vanish instantaneously, because of the finite-size and finite-scale interactions which also exist between the fluid molecules themselves. Viscous and thermal forces will then become spatially and temporally correlated, through a time-dependent viscous drag γ (this is discussed next). It is worthy of note that $F_{th}(t)$ can only be described in terms of its statistical properties, and as its effect has already vanished on experimentally accessible time scales, $F_{th}(t)$ has never been measured [23].

3.2.1.2 The Friction Force $F_{fr}(t)$

An incompressible isotropic fluid with a viscosity $\eta_f(t)$ and density ρ_f (f refers to fluid parameters) generates a viscous drag $\gamma(t)$ on the thermally excited Brownian particle as it moves through the fluid, giving rise to the friction force $F_{fr}(t)$. A correct

expression of $\gamma(t)$ and the resulting $F_{fr}(t)$ is given by solving the Navier–Stokes equation, describing the hydrodynamic properties of the fluid [24]. Here, the molecular character of the viscous fluid is neglected, and the fluid is treated as a continuum, which is valid when the Brownian sphere with radius a_s and density ρ_s is much larger than the fluid molecules. Then, the average free path length of the molecules which compose the fluid is small compared to the dimension of, for example, the sphere immersed in it. Furthermore, we will only consider the sphere moving far away from any boundary, like an obstacle placed in its trajectory, which would make the fluid anisotropic. Two experimentally relevant solutions of the Navier–Stokes equation can then be distinguished, the first being a commonly used approximation of the second:

(i) $\rho_s \gg \rho_f$: If the sphere has a high inertia m_s, and hence a density much higher than the fluid's density ρ_f, it will move steadily and at very slow speeds through the fluid. The fluid's response to the particle's presence can then be considered as instantaneous, and the solution of the Navier–Stokes equation is simply the constant Stokes' drag [3]:

$$\gamma = 6\pi\eta_f a_s \tag{3.3}$$

and the friction force F_{fr} follows Stokes' law, which states that it is instantaneously linear with the sphere's velocity $\dot{x}$:

$$F_{fr} = -6\pi\eta_f a_s \dot{x} \tag{3.4}$$

It must be noted here that η_f, the dynamic viscosity of the fluid, is considered as time-independent, thus implying that correlations between successive collisions of the fluid molecules on the Brownian particle have already vanished. Motion can hence be observed as a Markovian process – that is, a random walk.

(ii) $\rho_s \approx \rho_f$: As noted by Lorentz [25], Equation 3.1, which includes Stoke's law (Equation 3.4), is only consistent with hydrodynamics when $m_s \gg m_f = (4/3)\pi a_s^3 \rho_f$. When the sphere has a density similar to its surrounding medium – which is usually the case for the neutrally buoyant particles used in optical trapping – the sphere's motion will be determined not only by its own inertia but also by the inertia m_f of the surrounding fluid. Then, Brownian motion theory needs to include frequency-dependent effects, and the time dependence of $\dot{x}$ should be accounted for when computing the drag γ.

As the particle receives momentum from the fluctuating fluid molecules, it displaces the fluid in its immediate vicinity. Although the fluid can still be considered as a continuum, and even with the conditions of a low Reynold's number, the fluid's flow field will be perturbed. The non-negligible inertia of entrained fluid $m_f = (4/3)\pi a_s^3 \rho_f$ will act back on the sphere. As a consequence, correlations in the fluid's fluctuations will persist up to time scales observable by OTI and become experimentally relevant. The Brownian sphere will move with a non-constant velocity and perform a non-random walk, which will depend heavily on the nature of the surrounding medium. This phenomenon is commonly called hydrodynamic memory. Such perturbations give rise to the Stokes–Boussinesq friction force that is

derived from the Navier–Stokes equation accounting for the inertia of the fluid, and given as [9]:

$$F_{fr} = -6\pi\eta_f a_s \dot{x} - \frac{2}{3}\pi a_s^3 \rho_f \ddot{x} - 6a_s^2 \sqrt{\pi\eta_f \rho_f} \int_0^t (t-t')^{-1/2} \ddot{x}(t')\mathrm{d}t' \tag{3.5}$$

The first term is the ordinary Stokes' friction from Equation 3.4, while the second term is connected to the mass m_f of the incompressible fluid displaced by the Brownian particle. In principle, this term defines an effective mass $M = m_s + m_f/2$ that should replace m_s on the left-hand side of the Langevin equation (Equation 3.1). The third term is time-dependent, stating that the friction force at time t is determined by the penetration depth of the viscous, unsteady flow around the sphere at all preceding times. Equation 3.5 confirms that, for a fluid with a density similar to the density of the Brownian particle, the terms containing ρ_f cannot be neglected, as they are of the same order of magnitude as the inertial term.

Consequently, an instantaneous disturbance of the fluid from the thermally excited Brownian sphere will spread, and its initial momentum will be shared with all of the molecules in a small volume around the sphere. The velocity field of this moving volume then grows by vorticity diffusion. In an incompressible liquid, this enforces a back flow at short times, which creates a vortex ring in three dimensions [12]. The diffusive spreading of this vortex carries the momentum into the fluid on a time scale $\tau_f = a_s^2 \rho_f / \eta_f$ – the time needed for vorticity to diffuse over the distance of one particle radius. Figure 3.1 shows a simplified scheme of the characteristic double-vortex structure of the velocity field caused by the initial displacement of the Brownian sphere in a simple liquid.

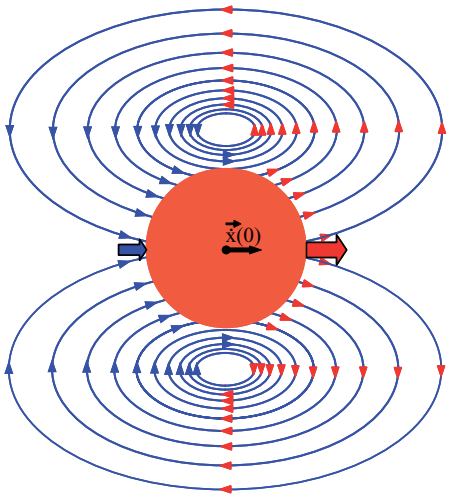


Figure 3.1 Schematic visualization of the velocity field of the fluid after the colloidal particle has been set in motion. (Inspired from computer simulations by Ref. [26].)

3.2.1.3 The External Force $F_{ex}(t)$

Only two different cases for $F_{ex}(t)$ will be considered for deriving the solutions of the Langevin equation:

(i) $F_{ex}(t) = 0$: When no additional force acts on the sphere and its motion is considered as free.

(ii) $F_{ex}(t) = -kx(t)$: Corresponding to the harmonic trapping potential with a force constant k created by the optical trap, which retains the sphere within the observation volume of the detector. The sphere's motion is then qualified as optically confined.

3.2.2 Solutions to the Different Langevin Equations for Cases Observable by OTI

From the solution of the Langevin equation for a Brownian sphere, the following measurable quantities of physical interest are derived for further studies in experiments (see Sections 3.3 and 3.4) the velocity autocorrelation function (VAF) $\langle \dot{x}(t)\dot{x}(0)\rangle$; the mean square displacement (MSD) $\langle \Delta x^2(t)\rangle$, which is related to the velocity autocorrelation function through:

$$\langle \Delta x^2(t)\rangle = 2\int_0^t (t-t')\langle \dot{x}(t')\dot{x}(0)\rangle \mathrm{d}t'; \tag{3.6}$$

and also the power spectral density (PSD) $\langle |\tilde{x}(f)|^2\rangle$, which mirrors the MSD through its Fourier transform as

$$\langle \Delta x^2(t)\rangle = 4\int_0^\infty \cos(2\pi f t)\langle |\tilde{x}(f)|^2\rangle \mathrm{d}f. \tag{3.7}$$

The following listing of all three measurables derived from the four discussed Langevin equations is meant to provide a summarizing overview of the theoretical models of Brownian motion derived in the literature by various authors. Each model can be picked accordingly to fit the data acquired by OTI, as discussed in Section 3.4. The most accurate expressions are also the most complex; however, a good understanding of the problem of Brownian motion in a viscous fluid is already gained by only considering the characteristic limiting behaviors in each situation.

3.2.2.1 Free Brownian Motion

Solving the Langevin Equation using Stokes Friction Solving the Langevin equation using the Stokes friction of Equation 3.4 and $F_{ex}(t) = 0$ results in a VAF:

$$\langle \dot{x}(t)\dot{x}(0)\rangle_{free} = \frac{k_B T}{m_s} e^{-t/\tau_s} \tag{3.8}$$

and in a MSD:

$$\langle \Delta x^2(t) \rangle_{free} = 2Dt\left[1 + \frac{\tau_s}{t}(e^{-t/\tau_s} - 1)\right] \tag{3.9}$$

that both decay exponentially with a characteristic time constant $\tau_s = m_s/\gamma = 2a_s^2\rho_s/9\eta_f$. This implies that, for a larger/heavier particle and/or a less viscous fluid, correlations will last longer. If $D = k_BT/\gamma$ is the diffusion constant, then the PSD will be:

$$\langle |\tilde{x}(f)|^2 \rangle_{free} = \frac{D}{\pi^2 f^2[(\gamma f/2\pi m_s)^2 + 1]} = \frac{D}{\pi^2 f^2[(f/\phi_s)^2 + 1]} \tag{3.10}$$

with $\phi_s = 1/2\pi\tau_s$, the corresponding characteristic frequency.

For short times ($t \to 0$), and respectively, high frequencies ($f \to \infty$), the particle moves with its initial velocity:

$$\dot{x}(0) = \sqrt{k_BT/m_s} \tag{3.11}$$

Then, the motion is ballistic with:

$$\langle \Delta x^2(t) \rangle_{free} = (k_BT/m_s)t^2 \tag{3.12}$$

and:

$$\langle |\tilde{x}(f)|^2 \rangle_{free} = (D\phi_s^2/\pi^2)f^{-4} \tag{3.13}$$

At long times ($t \to \infty$), and respectively, low frequencies ($f \to 0$), velocity correlations vanish exponentially with τ_s, the relaxation time of the particle's initial momentum. The particle has then lost all information about its initial velocity, and diffuses randomly with

$$\langle \Delta x^2(t) \rangle_{free} = 2Dt \tag{3.14}$$

and respectively

$$\langle |\tilde{x}(f)|^2 \rangle_{free} = (D/\pi^2)f^{-2} \tag{3.15}$$

Solving the Langevin Equation using the Stokes–Boussinesq Friction Force Solving the Langevin equation using the Stokes–Boussinesq friction force of Equation 3.5 and $F_{ex}(t) = 0$ results in more complex expressions [6, 11]:

$$\langle \dot{x}(t)\dot{x}(0) \rangle_{free} = \frac{k_BT}{2\pi a_s^3 \rho_f \sqrt{5 - 8\rho_s/\rho_f}}\left[\alpha_+ e^{\alpha_+^2 t} erfc(\alpha_+\sqrt{t}) - \alpha_- e^{\alpha_-^2 t} erfc(\alpha_-\sqrt{t})\right]$$

$$\text{with} \quad \alpha_\pm = \frac{3}{2}\frac{3 \pm \sqrt{5 - 36\tau_s/\tau_f}}{\sqrt{t}(1 + 9\tau_s/\tau_f)} \tag{3.16}$$

now also including the hydrodynamic effect, decaying with a fluid-dependent time constant $\tau_f = a_s^2\rho_f/\eta_f$, which represents the time needed by the perturbed fluid flow field to diffuse over the distance of one particle radius. For an incompressible fluid, the initial velocity is now given by $\dot{x}(0) = \sqrt{k_BT/(m_s + m_f/2)} = \sqrt{k_BT/M}$, and

τ_s should in principle be expressed as $\tau_s = (m_s + m_f/2)/\gamma = a_s^2(2\rho_s + \rho_f)/9\eta_f$. The MSD is given by:

$$\langle \Delta x^2(t) \rangle_{free} = 2Dt\left[1 - 2\sqrt{\frac{\tau_f}{\pi t}} + 4\frac{\tau_f}{t} - \frac{\tau_s}{t} + \Xi\left(\frac{\tau_s}{\tau_f}, \frac{t}{\tau_f}\right)\right]$$

$$\text{with} \quad \Xi\left(\frac{\tau_s}{\tau_f}, \frac{t}{\tau_f}\right) = \frac{3}{t\sqrt{5 - 36\tau_s/\tau_f}}\left[\frac{1}{\alpha_+^3} e^{\alpha_+^2 t} erfc\left(\alpha_+\sqrt{t}\right) - \frac{1}{\alpha_-^3} e^{\alpha_-^2 t} erfc\left(\alpha_-\sqrt{t}\right)\right]. \tag{3.17}$$

Equation 3.17 depends on the particle's inertia through τ_s and on the fluid's inertia through τ_f. The two times are connected by the relationship $\frac{\tau_s}{\tau_f} \approx \frac{2\rho_s}{9\rho_f}$.

The corresponding characteristic frequency $\phi_f = 1/2\pi\tau_f$ appears in the PSD as [27]:

$$\langle |\tilde{x}(f)|^2 \rangle_{free} = \frac{D}{\pi^2 f^2} \frac{\left(1 + \sqrt{f/2\phi_f}\right)}{\left[(f/\phi_s) + \sqrt{f/2\phi_f} + (f/9\phi_f)\right]^2 + \left(1 + \sqrt{f/2\phi_f}\right)^2} \tag{3.18}$$

The behaviors in the time and frequency limits of all three functions remain similar to the previously discussed case, meaning that, for very short times, the motion is ballistic and, for very long times, the motion is diffusive. However, the transition between the two regimes is algebraic and delayed to significantly longer times compared to the case of simple Stokes' friction. This translates into a slow algebraic decay in the MSD (Equation 3.12) and results in a VAF which is governed by a power-law rather than by an exponential tail [7]:

$$\langle \dot{x}(t)\dot{x}(0) \rangle_{free} \propto (t/\tau_f)^{-3/2}, \quad \text{for} \quad t \geq \tau_f \tag{3.19}$$

This power law is usually referred to in the literature as the 'long-time tail', and arises from the fluid vortices observed around the colloidal particle, as sketched in Figure 3.1.

The log-log plot in Figure 3.2a compares the VAF given by Equation 3.8 (red line) with that given by Equation 3.16 (blue line), both normalized by their respective initial velocity, for a sphere with radius $a_s = 1\,\mu m$, density $\rho_s = 1.51\,kg\,dm^{-3}$ immersed in water with viscosity $\eta_f = 10^{-3}$ Pa·s at $T = 22\,^\circ C$. It can be seen that the exponential relaxation resulting from the Stokes' friction changes to a power-law decay when the fluid's inertia is accounted for. In the same way, the log-log representations in Figures 3.2b and 3.2c show a comparison between Equations 3.9 and 3.17, as well as between Equations 3.10 and 3.18. Differences in the MSD and PSD are less visible in this representation, but the respective common limiting behaviors, translating to characteristic slopes are indicated. In Figure 3.2c, the discrepancies visible at high frequencies above 2 MHz (arrow) between both functions, arise from the differences in the displaced masses; m_s, and respectively M. The green bars on the abscissa highlight the time, and respectively, frequency regions accessible by OTI, as introduced in Section 3.3.

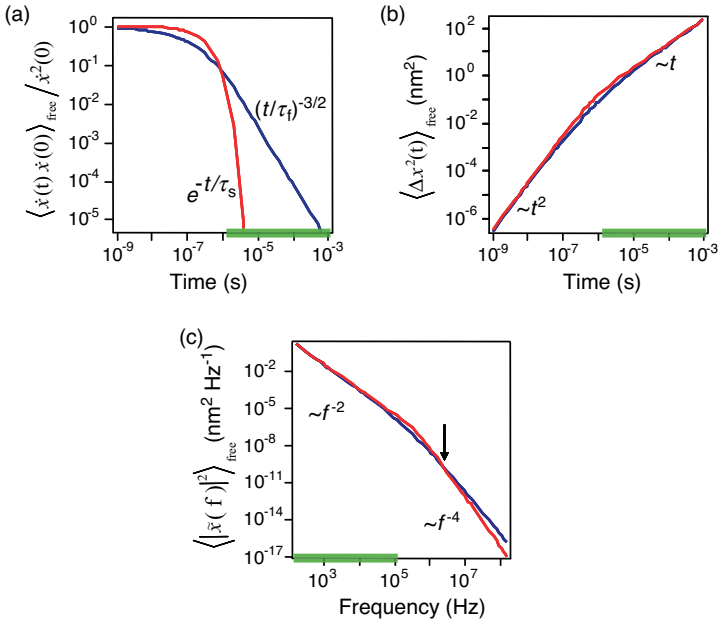


Figure 3.2 Log-log plots of (a) the normalized VAF given by Equation 3.8 (red line) and Equation 3.16 (blue line); (b) the MSD given by Equation 3.9 (red line) and Equation 3.17 (blue line); and (c) the PSD given by Equation 3.10 (red line) and Equation 3.18 (blue line) for a 2 µm-sized sphere in water. The chosen time span ranges from the short-time to the long-time limits of each function, which are indicated in each graph. The green bars on the abscissa highlight the regions accessible by OTI.

3.2.2.2 Optically Confined Brownian Motion

The Case of a Nonfree Particle In the case of a nonfree particle the situation becomes more complex. The diffusion of such a particle in an unbounded fluid was first described by Ornstein and Uhlenbeck [28]. The Langevin equation using the Stokes friction of Equation 3.4 gives then a velocity autocorrelation function:

$$\langle \dot{x}\ (t)\dot{x}(0)\rangle = \frac{k_B T}{m_s(\varsigma_+ - \varsigma_-)}\left[\varsigma_+ e^{-\varsigma_+ t} - \varsigma_- e^{-\varsigma_- t}\right]$$

$$\text{with} \quad \varsigma_\pm = \frac{1}{2\tau_s}\left(1 \pm \sqrt{1 - 4\tau_s/\tau_k}\right) \tag{3.20}$$

The VAF now has a positive part which decays exponentially to zero as $\langle \dot{x}(t)\dot{x}(0)\rangle \propto e^{-t/\tau_s}$ for $t \approx \tau_s$, and a negative part which decays exponentially as:

$$\langle \dot{x}(t)\dot{x}(0)\rangle \propto -e^{-t/\tau_k} \tag{3.21}$$

for $t \approx \tau_s$, a new characteristic time constant $\tau_k = 6\pi\eta_f a_s/k$ determined by the trap stiffness k. The MSD follows as:

$$\langle \Delta x^2(t) \rangle = \frac{2k_B T}{k}\left[1 + \frac{\varsigma_-}{\varsigma_+ - \varsigma_-}e^{-\varsigma_+ t} + \frac{\varsigma_+}{\varsigma_- - \varsigma_+}e^{-\varsigma_- t}\right] \tag{3.22}$$

and the PSD becomes Lorentzian [29]:

$$\langle |\tilde{x}(f)|^2 \rangle = \frac{D}{\pi^2 f^2 [1 + (\phi_k/f)^2]} \tag{3.23}$$

showing that the motion is also influenced by the trapping potential with the characteristic frequency $\phi_k = 1/2\pi\tau_k = k/12\pi^2\eta_f a_s$, corresponding to the corner frequency of the Lorentzian function.

In the short time limits, the behavior remains the same as introduced above, but in the long time limit it is now governed by the confining potential – that is, the optical trapping constant. Then, the velocity correlations still disappear, but Equation 3.22 approaches the time-independent limit:

$$\langle \Delta x^2(t) \rangle_{t \to \infty} = 2k_B T/k \tag{3.24}$$

and Equation 3.23 the limit:

$$\langle |\tilde{x}(f)|^2 \rangle_{f \to 0} = 4k_B T\gamma/\pi k^2 \tag{3.25}$$

The sphere's motion is now governed by the potential's restoring force with stiffness k.

Using the Stokes–Boussinesq Friction Force The most complete solution considered in this work is when the Stokes–Boussinesq friction force of Equation 3.5 and $F_{ex}(t) = -kx(t)$ are used to set up the Langevin equation. Then, the VAF is [14]:

$$\begin{aligned}\langle \dot{x}(t)\dot{x}(0) \rangle = {} & \frac{\zeta_1^3 e^{\zeta_1^2 t} erfc(\zeta_1\sqrt{t})}{(\zeta_1-\zeta_2)(\zeta_1-\zeta_3)(\zeta_1-\zeta_4)} + \frac{\zeta_2^3 e^{\zeta_2^2 t} erfc(\zeta_2\sqrt{t})}{(\zeta_2-\zeta_1)(\zeta_2-\zeta_3)(\zeta_2-\zeta_4)} \\ & + \frac{\zeta_3^3 e^{\zeta_3^2 t} erfc(\zeta_3\sqrt{t})}{(\zeta_3-\zeta_1)(\zeta_3-\zeta_2)(\zeta_3-\zeta_4)} + \frac{\zeta_4^3 e^{\zeta_4^2 t} erfc(\zeta_4\sqrt{t})}{(\zeta_4-\zeta_1)(\zeta_4-\zeta_2)(\zeta_4-\zeta_3)}\end{aligned} \tag{3.26}$$

where the coefficients ζ_1, ζ_2, ζ_3 and ζ_4 are the four roots of the equation:

$$\left(\tau_s + \frac{1}{9}\tau_f\right)\zeta^4 + \sqrt{\tau_f}\zeta^3 + \zeta^2 + \frac{1}{\tau_k} = 0.$$

Correspondingly, the MSD results in:

$$\langle \Delta x^2(t)\rangle = 2D\tau_k + \frac{2D}{\tau_s+\frac{\tau_f}{9}}\left[\frac{e^{\zeta_1^2 t}erfc(\zeta_1\sqrt{t})}{\zeta_1(\zeta_1-\zeta_2)(\zeta_1-\zeta_3)(\zeta_1-\zeta_4)}+\frac{e^{\zeta_2^2 t}erfc(\zeta_2\sqrt{t})}{\zeta_2(\zeta_2-\zeta_1)(\zeta_2-\zeta_3)(\zeta_2-\zeta_4)}\right] + \frac{2D}{\tau_s+\frac{\tau_f}{9}}\left[\frac{e^{\zeta_3^2 t}erfc(\zeta_3\sqrt{t})}{\zeta_3(\zeta_3-\zeta_1)(\zeta_3-\zeta_2)(\zeta_3-\zeta_4)}+\frac{e^{\zeta_4^2 t}erfc(\zeta_4\sqrt{t})}{\zeta_4(\zeta_4-\zeta_1)(\zeta_4-\zeta_2)(\zeta_4-\zeta_3)}\right] \tag{3.27}$$

and the PSD is given by [27]:

$$\langle|\tilde{x}(f)|^2\rangle = \frac{D}{\pi^2 f^2}\left[\frac{\left(1+\sqrt{f/2\phi_f}\right)}{\left[(\phi_k/f)-(f/\phi_s)-\sqrt{f/2\phi_f}-\left(f/9\phi_f\right)\right]^2+\left(1+\sqrt{f/2\phi_f}\right)^2}\right] \tag{3.28}$$

which is by far the most convenient formula of all three quantities to use for fitting data acquired in OTI experiments. Fortunately, despite their complexity, the limiting behaviors of all three expressions simplify in the same way as in the above-discussed cases.

3.2.3 Time Scales of Brownian Motion

All of the above considerations show that many different time scales (as outlined in Table 3.1) govern the physics of Brownian motion. To better appreciate these times, we can consider a sphere with radius $a_s = 1\,\mu m$ in water. The sphere is set into motion by random collisions with fluid molecules within $\tau_{col} = \bar{d}_{mol}\bar{v}_{mol} \approx 0.1$ ps, the correlation time of $F_{th}(t)$, estimated by the ratio of the mean solvent particle separation $\bar{d}_{mol}$ and fluctuation speed $\bar{v}_{mol}$ [30]. The momentum is then transferred from the particle to the fluid on very different time scales. If compressibility effects of the fluid are taken into account, one-third of the initial momentum is carried off

Table 3.1 Overview of the characteristic time scales of a Brownian particle in a harmonic potential.

Time constant	Origin	Typical value for our model system
$\tau_{col} = \bar{d}_{mol}/\bar{v}_{mol}$	Random molecular collisions	$\sim 10^{-13}$ s
$\tau_{sw} = a_s/c$	Sound wave propagation over the distance of a sphere's radius	$\sim 10^{-9}$ s
$\tau_s = m_s/\gamma = 2a_s^2\rho_s/9\eta_f$	Inertia of the Brownian particle	$\sim 10^{-6}$ s
$\tau_f = a_s^2\rho_f/\eta_f$	Inertia of the perturbed fluid surrounding the Brownian particle	$\sim 10^{-6}$ s
$\tau_k = 6\pi\eta_f a_s/k.$	Harmonic potential of the optical trap	$\sim 10^{-4}$ s

Typical values are calculated for a resin sphere ($a_s = 1\,\mu m$, $\rho_s = 1.51\,kg\,dm^{-3}$) in water ($\rho_f = 1\,kg\,dm^{-3}$, $\eta = 10^{-3}$ Pa·s).

rapidly by a spherical sound wave, the front of which leaves the sphere after a time $\tau_{sw} = a_s/c \approx 0.7$ ns, where c is the speed of sound in the fluid [31]. Equation 3.5, which was set for an incompressible fluid, has simply to be corrected to include a rapid change of the particle's inertial mass m_s to the combined mass $M = m_s + m_f/2$ [13]. Consequently, the velocity correlation function starts with the initial value given by the equipartition theorem, $\langle \dot{x}(0)\dot{x}(0)\rangle = k_B T/m_s$ and, after a short time, on the order of τ_{sw}, decays from $k_B T/m_s$ to $k_B T/M$ due to acoustic damping of the particle's velocity. When the sound wave has separated and the particle has relaxed with $\tau_s = m_s/\gamma \approx 0.9\,\mu$s or $\tau_{s+f} = (m_s + m_f/2)/\gamma \approx 0.7\,\mu$s, the vortex ring develops around the particle. The region of vorticity (see Figure 3.1) grows diffusively, as the remaining momentum is distributed over increasingly larger volumes, with the disturbance taking a time of order $\tau_f = a_s^2 \rho_f/\eta_f$ to leave the particle. Finally, the optical trapping force, $F_{ex}(t)$, sets in after a time $\tau_k = 6\pi\eta_f a_s/k$ and slows down the sphere, confining its motion around the potential minimum. The stronger the trap, the earlier optical confinement will reduce the particle's free motion.

For the model of a sphere in a harmonic potential with a typical spring constant $k = 10\,\mu\text{N m}^{-1}$, τ_s and τ_f are in the microsecond range, whereas τ_k is in the millisecond range. The relaxation time of the optical trapping potential is thus well separated from the others, and its separation can be adjusted by choosing suitable experimental parameters, that is, a_s, ρ_s, ρ_f, η_f and k. The diffusion constant D is on the order of $1\,\mu\text{m}^2\,\text{s}^{-1}$, and hence the time for the sphere to diffuse over its radius is on the order of 1 s. Correspondingly, within 1 μs, the sphere will have diffused about 1 nm, which is the time and distance range accessible by OTI. This will allow study of the interplay between τ_f and τ_k, as shown in detail in Section 3.4.

3.3
Experimental Aspects of Optical Trapping Interferometry

In OTI, optical trapping is combined with high-resolution interferometric position detection. The optical trap provides the linear external force $F_{ex}(t) = -kx(t)$ in the Langevin equation, which will maintain the studied microsphere within the detection range of the system. In general, the principles of optical trapping, as well as instrumental aspects, are well known and have been reported extensively in the literature [29, 32]. Therefore, at this point we will present only briefly the opto-mechanical and electronic components of the set-up. However, the data acquisition and analysis strategy deserves greater emphasis, as it allows the fine characterization of interactions between the Brownian probe and its environment.

3.3.1
Experimental Set-Up

The apparatus consists mainly of a custom-built inverted light microscope with a 3-D sample positioning stage, an infrared laser for trapping, and a quadrant photodiode for high-resolution 3-D and time-resolved position detection. The two main custom-made

Figure 3.3 Photograph of the experimental set-up on a vibration isolation table. The mechanical frame is made from titanium, steel and aluminum to achieve maximal mechanical and temperature stability. For details of abbreviations, see text and Figure 3.4. (Photograph courtesy of Daniel Gutierrez.)

circular base plates are made from the titanium alloy Ti-6AL-4V, on the basis of its high tensile strength, light weight, low thermal conductivity, low thermal expansion coefficient and corrosion resistance compared to steel or aluminum. The commercially available mechanical pieces are either made of aluminum or steel. In order to minimize mechanical vibrations, the whole set-up is mounted on a table with tuned damping (RS-4000, Newport, UK), supported by pneumatic isolators (I-2000, Newport). The main features of the instrument are shown in Figure 3.3.

3.3.1.1 Optical Trapping Interferometry and Microscopy Light Path

The optical paths can be divided into an infrared (IR) trapping and detection light path, and a visible illumination and imaging light path, as shown in Figure 3.4. The trapping beam is emitted by a diode-pumped, ultra-low-noise Nd:YaG laser with a wavelength of $\lambda = 1064\,\text{nm}$ (IRCL-500-1064-S; CrystaLaser, USA) and a maximal light power of 500 mW in continuous-wave mode. The choice of the near-IR wavelength satisfies the requirement of minimal water absorption to avoid heating of the sample in the laser focus [29]. A high-intensity gradient for good trapping efficiency is achieved by over-illuminating the high numerical back-aperture of the focusing lens (OBJ). Therefore, the effective laser beam diameter is expanded 20-fold by a telecentric lens system (EXP; Beam Expander, Sill Optics, Germany). In order to

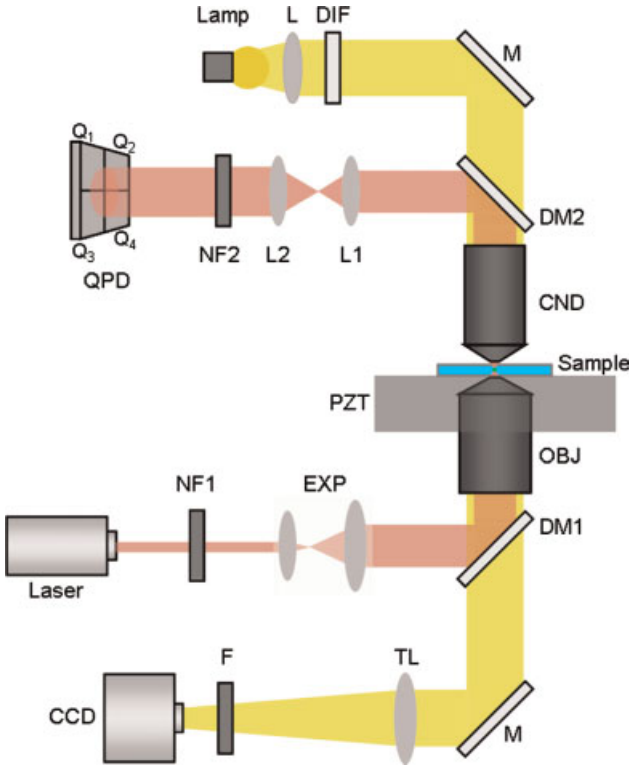


Figure 3.4 Schematic layout of the infrared (IR; red) and visible (yellow) light paths. The laser beam is expanded 20-fold by a beam expander (EXP), attenuated if necessary by a neutral density filter (NF1), and then reflected by a dichroic mirror (DM1) and focused by the objective lens (OBJ) into the sample chamber, which is mounted on a piezostage (PZT). The scattered IR light is collected by a condenser (CND), and directed by a second dichroic mirror (DM2) onto the quadrant photodiode (QPD). A second neutral density filter (NF2) is placed in front of the QPD, to avoid possible saturation of the detector. A 50 W halogen light source (Lamp) illuminating the object plane, through a lens (L) and diffuser (DIF), is reflected by the first mirror M, but transmitted through both dichroic mirrors (DM1 and 2). The image created by the CND and OBJ is reflected by the second lower mirror (M) and the 180 mm tube lens (TL) onto the charge-coupled device camera (CCD).

minimize noise, the laser is operated at high power, and its intensity is adapted after expansion by neutral density filters (NF1) with variable transmission coefficients ($T = 0.25$, 0.1, 0.01 or 0.001; OWIS, Germany). Increasing the transmission coefficient of NF1 will decrease the trapping stiffness, as this depends linearly on the laser power [33]. Next, the IR-beam is reflected by a dichroic mirror (DM1; AHF analysentechnik AG, Germany) into the high numerical aperture (NA = 1.2) of a $\times 60$ water-immersion objective (OBJ; UPLapo/IR, Olympus, Japan), which focuses the laser down to its refraction limit into the object plane of the microscope and creates the optical trap. The choice of a water-immersion objective lens offers a longer working distance of up to 280 µm compared to oil-immersion lenses. Such a long

working distance guarantees a stable space-invariant trap through the entire sample chamber. This is particularly essential when studies on Brownian particles far away from any surface boundary are of interest (as will be the case for the experiments discussed below).

The sample is mounted onto an *xyz*-piezo scanning table (PZT; P-561, Physikalishe Instrumente, Germany) for manipulation and positioning. The PZT with controller (E-710 Digital PZT Controller; Physikalishe Instrumente, Germany) has a travel range of 100 µm along all three dimensions, with a precision of 1 nm. The laser light focused by the objective lens is collected with a condenser (CND; 63X, Achroplan, NA = 0.9, water-immersion; Zeiss, Germany), and projected by a second dichroic mirror (DM2) and two lenses (L1 and L2, with focal lengths $f_1 = 30$ mm and $f_2 = 50$ mm) with a magnification of $f_1/f_2 \approx 1.67$ onto an InGaAs quadrant photodiode (QPD; G6849, Hamamatsu Photonics, Japan). The QPD is placed in the back focal plane of the condenser and fixed on an *x*–*y* translation stage (OWIS, Germany) for manual centering of the detector relative to the IR-beam. In order to avoid possible saturation of the QPD, a second neutral density filter (NF2) can be optionally placed in front of the QPD. For illumination in the visible range, a 50 W halogen lamp is diffused (DIF) and projected by a mirror (M) through the condenser, objective and a 180 mm tube lens (TL) that creates an image of the object plane onto a charge-coupled device camera (CCD; ORCA ER S5107, Hamamatsu Photonics, Japan). The image of the object plane is digitized (Hamamatsu Digital Controller ORCA ER), and can be further processed (HiPic, Hamamatsu Photonics, Japan).

3.3.1.2 Sample Preparation

The sample chamber consists of a custom-made flow cell. A coverslip (thickness ~130 µm) is stuck to a standard microscope slide using two pieces of double-sided tape, arranged in such a way as to form a ~5 mm-wide and ~70 µm-thick channel with a volume of ~20 µl. After loading with a suspension of microspheres, the flow-cell is mounted upside down on the 3-D piezo-stage. In the experiments presented in Section 3.4, either polystyrene ($\rho_s = 1.05\,\mathrm{kg\,dm^{-3}}$; Bangs Laboratories, USA), melamine resin ($\rho_s = 1.51\,\mathrm{kg\,dm^{-3}}$; Sigma-Aldrich, USA) or silica spheres ($\rho_s = 1.96\,\mathrm{kg\,dm^{-3}}$; Bangs Laboratories, USA) were used with radii (a_s) varying from 0.27 to 2 µm. To guarantee the manipulation and analysis of exclusively one particle, a particle concentration of 10^6 spheres per milliliter was used to maximize the average distance between two neighboring particles and minimize their mutual influence on their motions [34].

3.3.2 Position Signal Detection and Acquisition

When following the 3-D Brownian motion of the trapped particle, the scattering of the strongly focused trapping laser on the particle is measured by the QPD. The InGaAs Quadrant Photodiode (G6849, Hamamatsu Photonics, Japan) is 2.0 mm in diameter with a dead zone of 0.1 mm between the quadrants. The photosensitivity is 0.67 A W^{-1} at 1064 nm. The QPD signals are fed into a custom-built preamplifier (Pre-AMP;

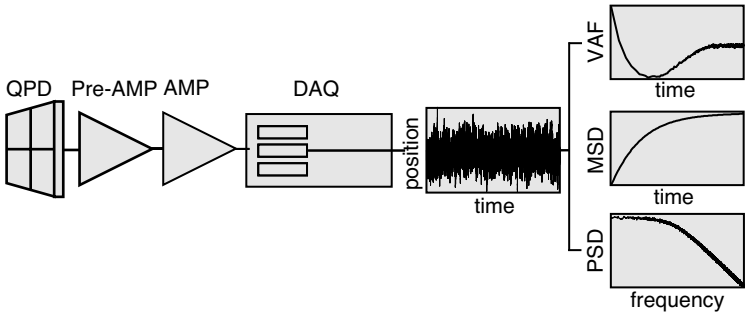


Figure 3.5 Position signal acquisition and data processing. Intensity fluctuations are recorded on the quadrant photodiode (QPD) and converted to volts (Pre-AMP). The signal is amplified (AMP) and digitized using the acquisition card (DAQ). The VAF, MSD and PSD are then calculated from the recorded position time trace.

Öffner MSR-Technik, Germany) which provides two differential signals between the segments and one signal that is proportional to the total light intensity (Figure 3.5). Preamplification of the quadrant photodiode signals at $20\,\mathrm{V\,mA^{-1}}$ with $0.67\,\mathrm{A\,W^{-1}}$ photosensitivity leads to a voltage of $13.4\,\mathrm{V\,mW^{-1}}$. Subsequently, differential amplifiers (AMP; Öffner MSR-Technik) adjust the preamplifier signals for optimal digitalization by the data acquisition board (DAQ) with a dynamic range of 12 bits (NI-6110, National Instruments, USA). Amplification of the QPD signal is chosen to span the maximal dynamic range of the acquisition card. The amplifier (Öffner MSR-Technik), with a maximal gain of 500, has a cut-off frequency around 1 MHz. The particle's position can be detected in all three dimensions. On the QPD, scattered and unscattered light generate an interference pattern that corresponds to the probe's position. A displacement of the particle near the beam focus modulates the optical power collected by the QPD. When the sphere moves perpendicularly to the optical axis – that is, along the x-direction – the current signal $S_x = (Q_1 + Q_2) - (Q_3 + Q_4)$ measured between both top segments (Q_1,Q_2) and both bottom segments (Q_3,Q_4) of the QPD changes correspondingly. The same holds for movements in the y-direction. Displacements along the z-axis instead affect the sum-signal $S_z = Q_1 + Q_2 + Q_3 + Q_4$ of the QPD, and so z displacements can be determined by the changes in total intensity. The full 3-D position information of the probe is thus encoded in the interference pattern of the forward-scattered and transmitted laser light recorded by the QPD. A detailed analysis of the detector response is given for fluctuations perpendicular to the optical axis in Ref. [35], and along the optical axis in Refs [36] and [37]. For small displacements from the trap center, the differential signals from the QPD are proportional to the lateral displacements of the particle in the optical trap, and the sum signal to the axial displacement.

The scanning stage, CCD camera and data acquisition are controlled and coordinated by a custom-made program written in VEE (Agilent, USA). Data are saved in binary format and analyzed with Igor 6.0 (WaveMetrics, USA).

The 3-D position–time traces of the Brownian motion of the probe can be acquired up to maximally $N = 10^7$ points (this limitation is set by the working memory of the

VEE). Data analysis also becomes very slow when the data files exceed 3.10^7 points per channel, and therefore the combination between data acquisition rate f_{acq} and total recording time t_{tot}, must be adjusted according to $N = f_{acq}t_{tot}$. A schematic overview of the signal acquisition and data processing schemes is shown in Figure 3.5.

3.3.3
Position Signal Processing

The three quantities of VAF, MSD and PSD, which were introduced in Section 3.2 can be calculated from the same experimental time trace $x(t)$, which is recorded in volts and converted to nanometers after fitting to the suitable theory of Brownian motion (as described in Section 3.4.1). An example of such a time trace, as well as the resulting VAF, MSD and PSD is shown on the right-hand side of Figure 3.5 for a 2 μm resin bead. For the sake of clarity, only one-dimensional time traces $x(t)$ will be presented through the remainder of this chapter, even though the developed data analysis strategy also applies to the y and z directions.

The velocity $\dot{x}(t) = \Delta x(t)/\Delta t$ of the studied sphere is derived from the steps $\Delta x(t) = x(t + \Delta t) - x(t)$ it performed, where Δt is the lag time related to the acquisition frequency as $f_{acq} = 1/\Delta t$. For the total number of acquired points N, the total recording time is expressed as $t_{tot} = N\Delta t$.

The velocity autocorrelation function $\langle \dot{x}(t)\dot{x}(0)\rangle$ is then defined as:

$$\langle \dot{x}(t)\dot{x}(0)\rangle = \frac{1}{(\Delta t)^2}\langle \Delta x(t)\Delta x(0)\rangle = \frac{f_{acq}^2}{N}\sum_i \Delta x(t + i\Delta t)\Delta x(i\Delta t) \tag{3.29}$$

with N the total number of acquired points.

$\langle \dot{x}(t)\dot{x}(0)\rangle$ can be normalized by its initial value in an incompressible fluid:

$$\langle \dot{x}^2(0)\rangle = k_B T/(m_s + m_f/2) \text{ at } t = 0.$$

The MSD is calculated from $x(t)$ as:

$$\langle \Delta x^2(t)\rangle = \frac{1}{N}\sum_i [x(t + i\Delta t) - x(i\Delta t)]^2 \tag{3.30}$$

The discrete Fourier transform of $x(t)$ is: $\tilde{x}(f) = (1/f_s)\sum_i e^{i2\pi tf/N}x(t)$, where $t = i\Delta t$ and $i = 0, \pm 1, \ldots, \pm N/2$ and the power spectral density follows as:

$$\langle |\tilde{x}(f)|^2\rangle = \frac{|\tilde{x}(f)|^2}{t_{tot}} \tag{3.31}$$

3.3.4
Temporal and Spatial Resolution of the Instrument

Apart from the signal arising from the particle's thermal fluctuations, anything that changes the intensity recorded by the quadrant photodiode will limit the resolution of the system. The main noise sources are the mechanical instabilities

of the microscope, and electronic noise combined with laser intensity fluctuations and pointing instabilities. Mechanical instabilities mainly cause low-frequency noise (drift). Laser intensity fluctuations may lead to temporal variations in the spring constants of the optical trapping potential, and pointing instabilities to unwanted drifting of the trapping focus in the specimen plane. To quantify the contribution of laser noise to the measured signal $x(t)$, it is decomposed in:

$$x(t) = x_s(t) + x_n(t)$$

where $x_s(t)$ is the particle's thermal fluctuations and $x_n(t)$ is the noise contribution to the signal. A subtraction of its contribution can increase the quality of the signal. With the assumption that position fluctuations of the sphere and the noise are uncorrelated – that is, $\langle x_s(t)x_n(t')\rangle = 0$ – the velocity autocorrelation function can be written as:

$$\langle \dot{x}(t)\dot{x}(0)\rangle = \langle \dot{x}_s(t)\dot{x}_s(0)\rangle + \langle \dot{x}_n(t)\dot{x}_n(0)\rangle$$

the MSD of the acquired signal as:

$$\langle \Delta x^2(t)\rangle = \langle \Delta x_s^2(t)\rangle + \langle \Delta x_n^2(t)\rangle$$

and similarly,

$$|\tilde{x}(f)|^2 = |\tilde{x}_s(f)|^2 + |\tilde{x}_n(f)|^2,$$

where $\tilde{x}_s(f)$ and $\tilde{x}_n(f)$ are the Fourier transforms of $x_s(t)$ and $x_n(t)$ respectively.

The calibrated (see Section 3.4.1) position fluctuations as a function of time of a trapped sphere (resin, radius $a_s = 1\,\mu\text{m}$ in water, $k \approx 5\,\mu\text{N m}^{-1}$, $f_{acq} = 5\,\text{MHz}$, $t_{tot} = 2\,\text{s}$) are shown in Figure 3.6a (red line). The blue line represents the recordings of the empty trap's noise signal $x_n(t)$ after the sphere has been released. $x_n(t)$ can be minimized by optimizing the illumination pattern of the incident laser spot on the QPD. The comparison between $x_s(t)$ and $x_n(t)$ indicates that the laser noise contribution in this configuration is very small, and its influence on the VAF, MSD and PSD are shown in Figure 3.6b–d, respectively. As expected, the velocity fluctuations of the laser spot without a scattering particle are uncorrelated and fluctuate around zero (Figure 3.6b, blue line). Correspondingly, $\langle \Delta x^2(t)\rangle_n$ is small and constant (Figure 3.6c, blue line), whereas $\langle \Delta x^2(t)\rangle$ increases with time (Figure 3.6c, red line). The resolution of the MSD of the sphere's position fluctuations can then be enhanced by subtracting $\langle \Delta x^2(t)\rangle_n$ from $\langle \Delta x^2(t)\rangle$. The spatial resolution of $\sim$8 Å achieved by the apparatus can be read from the first time points of the MSD (inset of Figure 3.6c, indicated by brackets).

In the frequency domain (Figure 3.6d), we define f_N as the frequency at which the power spectrum of the trapped bead (red line) drops to the noise level given by the power spectrum of the empty trap (blue line). The amplifier has a Butterworth-type low-pass filter with a cut-off frequency at 1 MHz, which is slightly above the frequency $f_N \approx 0.8\,\text{MHz}$. Therefore, in the following section we will analyze data in the frequency range up to a maximum of 0.5 MHz, setting the time resolution

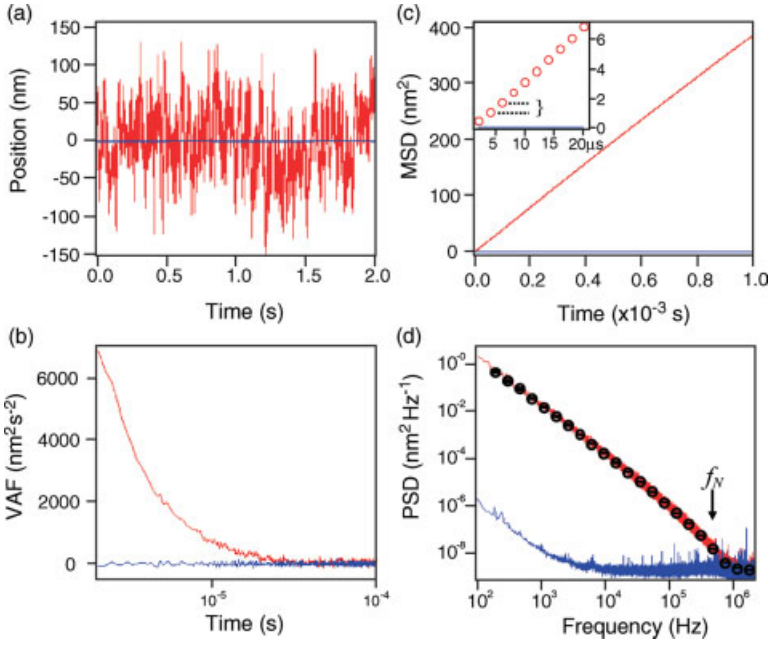


Figure 3.6 (a) Position signal of the sphere ($a_s = 2\,\mu m$) in the trap (red) and of the empty trap (blue) acquired during $t_{tot} = 2$ s at $f_{acq} = 5$ MHz; (b) VAF, (c) MSD and (d) PSD calculated from the position signal (the black circles represent the PSD blocked in five bins per decade). The frequency f_N, where the signal reaches the noise floor, is indicated by an arrow.

of the OTI system at 2 µs. Additionally, in order to avoid aliasing artifacts [38] from the data acquisition card, the acquired signal is oversampled by a factor of 2; $f_{acq} > 2f_N$.

When plotting the VAF and PSD on a log-log scale, further data processing can be made to enhance noisy data. As both functions are distributed exponentially, the number of points in a log-log plot will increase with, respectively, time or frequency. Therefore, data are commonly averaged from a 'block' of consecutive data points [39], resulting in equidistantly distributed points on the logarithmic scale. In the example discussed above, data were blocked in five bins per decade (Figure 3.6d, black circles). The data scattering around their average value gives the standard error, which remains within the size of the black circles. Together with improving the visibility of data, blocking allows fitting the data by the least-squares method, which assumes that the analyzed data points are statistically independent and conform to a Gaussian distribution [27]. Whilst the second assumption is satisfied by the VAF and PSD (as defined in Equations 3.29 and 3.31), the first assumption is satisfied only after blocking.

3.4 High-Resolution Analysis of Brownian Motion

After having characterized the performance of the OTI set-up, we can now present measurements on details in the Brownian motion of the model sphere. We demonstrate that the theory presented in Section 3.2 may be used for fitting and calibrating the data. The influence of experimentally relevant parameters on Brownian motion will be demonstrated, and the consequences of the presence of inertial effects in the data discussed.

3.4.1 Calibration of the Instrument

Position signal calibration consists of determining the detector sensitivity β and the spring constant k of the optical trapping potential. The term β has units of $\mathrm{V\,nm^{-1}}$, as the acquired position signal is recorded in volts, and the position of the particle is expressed in nanometers. Both, experimental and theoretical investigations [40] have shown that, close to the trap center, the optical trapping forces are well approximated by three orthogonal forces derived from a harmonic trapping potential. For a given wavelength of the laser beam, the spring constant along each direction depends mainly on the particle's size, the difference between refractive indices of particle and fluid, and the intensity of the trapping laser light. When the sphere's radius is known, the physics of Brownian motion in a harmonic potential (see Section 3.2.2.2) can be exploited to calibrate the optical trap [38]. Either of three quantities presented in Section 3.3.3 – namely the VAF, MSD or PSD – can be used to obtain β and k simultaneously. An overview is provided in Figure 3.7 for measurements of a resin sphere with radius $a_s = 0.5\,\mu\mathrm{m}$, $f_{acq} = 0.5\,\mathrm{MHz}$, $t_{tot} = 20\,\mathrm{s}$. For least-square fitting, the VAF (top graph) is normalized by its initial value, blocked in 10 bins per decade, and represented in a linear-log plot, whereas the PSD (bottom graph) is blocked in 10 bins per decade and plotted on a log-log scale. As can be seen by comparing Figures 3.2 and 3.6, the bandwidth of OTI allows the measurement of Brownian motion within a time range, during which it is greatly influenced by the hydrodynamic memory effect. Hence, the calibration must be made by using a theory that accounts for the fluid's inertia – that is, using Equation 3.26 instead of Equation 3.8 for fitting the VAF, Equation 3.27 instead of Equation 3.9 for fitting the MSD, and Equation 3.28 instead of Equation 3.10 for fitting the PSD. The black continuous line in each graph of Figure 3.7 therefore corresponds to Equations 3.26, 3.27and 3.28, respectively, being fitted to the data (red circles) with the two fitting parameters β and k. All three fits generally provide an accuracy of better than 6%, depending on the acquisition frequency and total acquisition time. The relative difference for all fitted values of β acquired by either of each equation is less than a small percentage [41]. The trapping force constant k (see Table 3.2 and next section) is obtained from the long-time and respectively low-frequency, limits of each function, according to Equations 3.21, 3.24 and 3.25, or from the corner frequency ϕ_k in Equation 3.23. The two main sources of error are uncertainties in the determination of the bead size and of the temperature

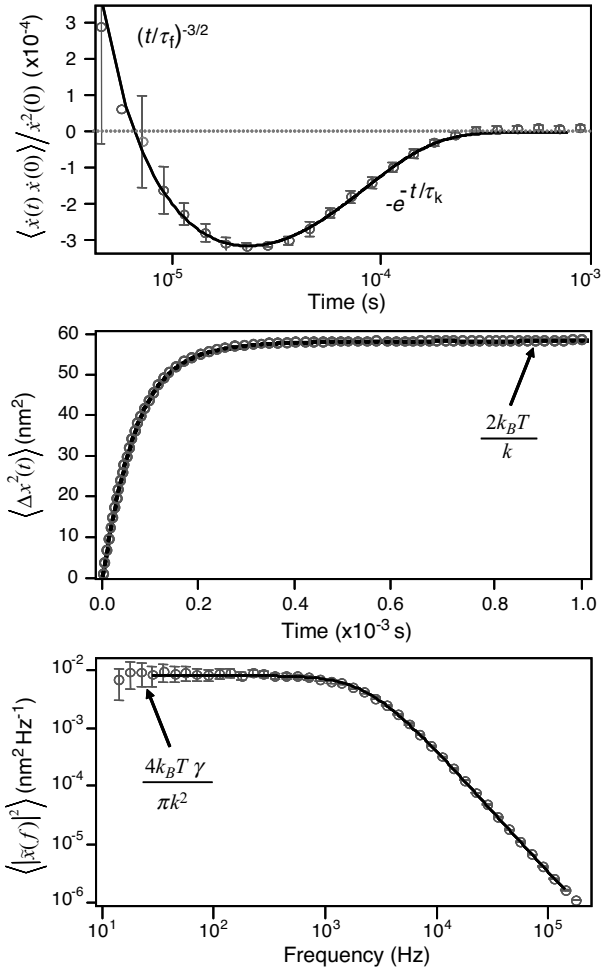


Figure 3.7 Measured VAF (raw data as red line, blocked values as red circles), MSD (circles, data points were removed for clarity) and PSD (only blocked values are shown as circles). The respective fits calculated from Equations 3.26, 3.27or 3.28 are plotted as black lines.

inside the laser focus. The latter can lead, in particular, to unwanted fluctuations in the fluid's viscosity and density [42].

3.4.2
Influence of Different Parameters on Brownian Motion

In this section we describe the influences of the trapping potential, the surrounding fluid's properties and the particle's properties on Brownian motion.

Table 3.2 Comparison of the three values k_1, k_2 and k_3 obtained from fitting the exponential relaxation time, the corner frequency, or the long time limits of the MSD and PSD.

	k_1	k_2	k_3
$\tau_k = \frac{6\pi\eta_f a_s}{k}$	$69\,\mu s \rightarrow 136\,\mu N\,m^{-1}$	$293\,\mu s \rightarrow 33\,\mu N\,m^{-1}$	$798\,\mu s \rightarrow 12\,\mu N\,m^{-1}$
$\phi_k = \frac{k}{12\pi^2\eta_f a_s}$	$2361\,Hz \rightarrow 140\,\mu N\,m^{-1}$	$547\,Hz \rightarrow 32\,\mu N\,m^{-1}$	$197\,Hz \rightarrow 12\,\mu N\,m^{-1}$
$\langle \Delta x^2(t)\rangle_{t\rightarrow\infty} = \frac{2k_BT}{k}$	$58.6\,nm^2 \rightarrow 139\,\mu N\,m^{-1}$	$243.3\,nm^2 \rightarrow 33\,\mu N\,m^{-1}$	$998\,nm^2 \rightarrow 8\,\mu N\,m^{-1}$
$\langle \lvert\tilde{x}(f)\rvert^2\rangle_{f\rightarrow 0} = \frac{96k_BT\eta_f a_s}{k^2}$	$10^{-2}\,nm^2\,Hz^{-1} \rightarrow 135\,\mu N\,m^{-1}$	$1.6\times10^{-2}\,nm^2\,Hz^{-1} \rightarrow 35\,\mu N\,m^{-1}$	$10^{-2}\,nm^2\,Hz^{-1} \rightarrow 7\,N\,m^{-1}$

3.4.2.1 Changing the Trap Stiffness

The influence of k on each quantity is shown in Figure 3.8 for the example of a resin sphere, with $a_s = 0.5\,\mu m$, held in three different potentials $k_1 = 140\,\mu N\,m^{-1}$ (red line), $k_2 = 32\,\mu N\,m^{-1}$ (blue line) and $k_2 = 12\,\mu N\,m^{-1}$ (green line). The variation of the trapping potential was achieved experimentally by changing the neutral density filter in front of the laser (NF1 in Figure 3.4). The acquisition frequency and time were chosen as: $f_{acq} = 0.5$ MHz, $t_{tot} = 20$ s. Changing k varies only τ_k, while τ_p and τ_f remain constant, as the particle's and the fluid's properties are fixed ($\tau_p = 0.084\,\mu s$, $\tau_p = 0.25\,\mu s$).

In the VAF (top graph, Figure 3.8), a distinction can be made between the three regimes in which the velocity correlations are either positive ($t < 10^{-5}$ s), negative ($10^{-5}\,s < t < 10^{-3}$ s), or vanish ($t > 10^{-3}$ s). As stated by Equation 3.16 and shown in Figure 3.2 (top, blue line), the velocity correlations of a free particle are always positive. However, a particle in an optical trap experiences an additional drift towards the trap center due to $F_{ext}(t)$. As the direction of the potential's restoring force $F_{ext}(t) = -kx(t)$ is opposite to the direction of the initial velocity $\dot{x}(0)$, the harmonic potential eventually introduces negative correlations, so-called anti-correlations in the VAF. As expected, the anti-correlations increase with k. The relaxation time τ_k describes the time scale for which $F_{ext}(t)$ makes the particle return from a displaced position to the trap center. For the stiffer trap in this example we find, from fitting, $\tau_{k_1} = 69\,\mu s$, while for the softer trap $\tau_{k_3} = 798\,\mu s$. As seen in the three data sets in Figure 3.8, $\langle \Delta x^2(t)\rangle_{t\rightarrow\infty}$ and, respectively, $\langle \lvert\tilde{x}(f)\rvert^2\rangle_{f\rightarrow 0}$ approach a constant value that is inversely proportional to k (Equation 3.24) or correspondingly k^2 (Equation 3.25). In Table 3.2 it can be seen that the values for k_1, k_2 and k_3 obtained from each relationship are consistent with each other, and vary by a maximum of 10%, except for k_3. A better precision for characterizing such soft trapping potentials can be obtained by acquiring data over longer times, as this increases the number of points in the time range of τ_k (data not shown).

The data in Figure 3.8 show that for $t < 10^{-5}$ s and, respectively, $f > 10^{-4}$ Hz, there is a time range during which the particle is free from any influence of the potential. As this time range is longer for weaker traps, in order to study the influences of the fluid and the particle on Brownian motion separately from the influence of an external

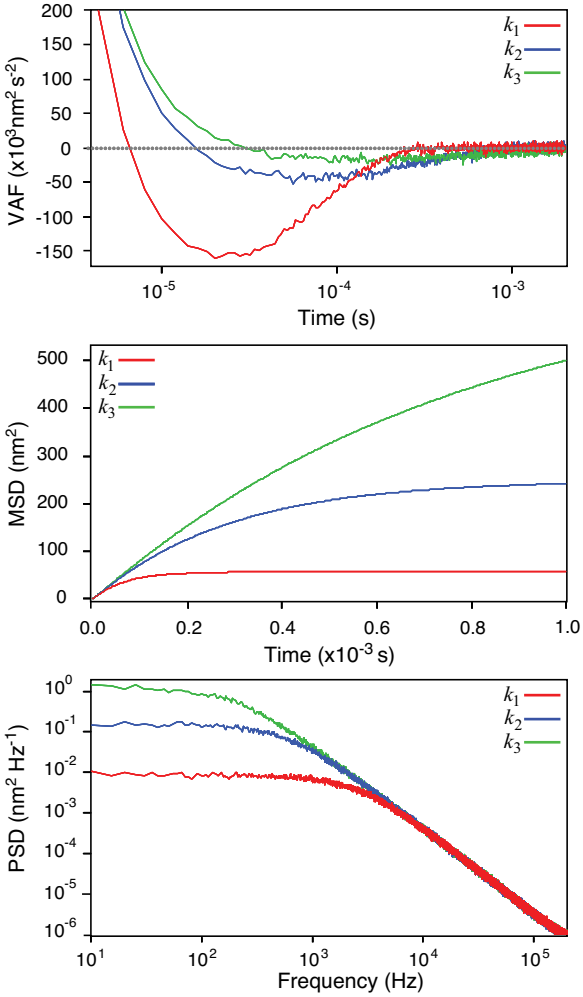


Figure 3.8 Measured VAF, MSD and PSD for three different force constants: $k_1 = 140\,\mu N\,m^{-1}$ (red line), $k_2 = 32\,\mu N\,m^{-1}$ (blue line) and $k_2 = 32\,\mu N\,m^{-1}$ (green line).

force, the trapping stiffness is minimized and the OTI configuration is used solely as a position detector for single particle tracking.

3.4.2.2 Changing the Fluid

In order to study the influence of the fluid's inertia and detect the long-time tail in the normalized VAF (see Figure 3.1 and Section 3.2.2.1), we track the Brownian motion of a larger resin sphere ($a_s = 1.5\,\mu m$, $\rho_s = 1.51\,kg\,dm^{-3}$, $f_{acq} = 0.5\,MHz$, $t_{tot} = 20\,s$) in three different solvents having different viscosities η_f and, unavoidably, different densities ρ_f (Figure 3.9). The resin beads are suspended either in a more viscous

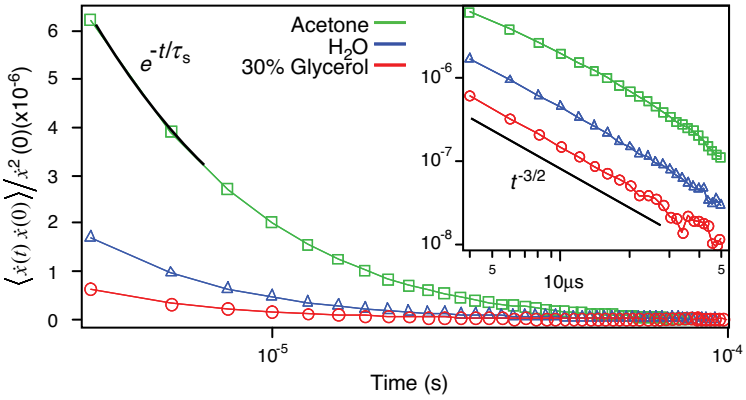


Figure 3.9 Log-linear representation of the measured normalized VAF of a resin sphere ($a_s = 1.5\,\mu m$, $\rho_s = 1.51\,kg\,dm^{-3}$, $f_{acq} = 0.5$ MHz, $t_{tot} = 20$ s) with $k \approx 10\,\mu N\,m^{-1}$, for three different fluids; 30% glycerol in H_2O ($\rho_f = 1.18\,kg\,dm^{-3}$, $\eta_{30\%glycerol} = 2.11 \times 10^{-3}$ Pa·s, ○), H_2O ($\rho_f = 1.118\,kg\,dm^{-3}$, $\eta_{H2O} = 10^{-3}$ Pa·s, △) and acetone ($\rho_f = 0.790\,kg\,dm^{-3}$, $\eta_{acetone} = 0.306 \times 10^{-3}$ Pa·s, □). Fits correspond to the continuous line with the respective color. Inset: log-log plot of the same VAFs. The $t^{-3/2}$ power-law is indicated by the black line.

solution of 30% glycerol in water ($\rho_{30\%Glycerol} = 1.18\,kg\,dm^{-3}$, $\eta_{30\%Glycerol} = 2.11 \times 10^{-3}$Pa·s, ○), in pure water with an intermediate viscosity ($\rho_{H_2O} = 1\,kg/dm^3$, $\eta_{H_2O} = 10^{-3}$ Pa · s, △), or in pure acetone ($\rho_{acetone} = 0.790\,kg\,dm^{-3}$, $\eta_{acetone} = 0.306 \times 10^{-3}$ Pa·s, □), the fluid with the lowest viscosity we could find. Decreasing η_f results in increasing not only τ_f as: $\tau_{30\%Glycerol} = 1.26\,\mu s$, $\tau_{H_20} = 2.25\,\mu s$ and $\tau_{acetone} = 5.8\,\mu s$, but also τ_s as: $\tau_{s_{30\%Gl}} = 0.36\,\mu s$, $\tau_{s_{H_20}} = 0.76\,\mu s$ and $\tau_{s_{ac}} = 2.48\,\mu s$. As $\tau_f$and τ_s depend on a_s^2, the use of a larger Brownian particle increases both time scales and places them closer to (or even within) the detection bandwidth of the instrument. The choice of a resin sphere in this particular experiment is motivated by the fact that resin, unlike polystyrene, is not soluble in acetone. Furthermore, the difference between the refractive indices of resin ($n = 1.68$) and acetone ($n = 1.36$) is still high enough to provide a good contrast and allow visualization of the particle by optical microscopy, which is not the case for example, with silica ($n = 1.37$).

In order to obtain anticorrelations that are negligible compared to the detection limit, the trap stiffness is reduced to $k \approx 10\,\mu N\,m^{-1}$, and hold this constant from one fluid to the other. However, this does not prevent τ_k from changing, as it is dependent on η_f as: $\tau_{k_{30\%Glyc}} \approx 5.9$ ms, $\tau_{k_{H2O}} \approx 2.84$ ms, $\tau_{k_{aceton}} \approx 950\,\mu s$. With such a weak trap the persistent positive correlations arising from the hydrodynamic back-flow introduced in Section 3.2.1.2 can be readily identified up to 50 μs. The data are best fitted by the normalized Equation 3.26 (continuous lines). In Figure 3.9, it can be seen that, as expected, for lower viscosities the correlations last longer (green line). Furthermore, the log-log representation in the inset allows recognizing of the $t^{-3/2}$ power law decay (black line) that is followed by all three fluids. In acetone, as $\tau_{s_{ac}} \approx 3\,\mu s$ is also within the detection range, it is possible to detect between 2 and 6 μs an exponential tendency arising from the combined fluid's and particle's inertia (Figure 3.9, green line).

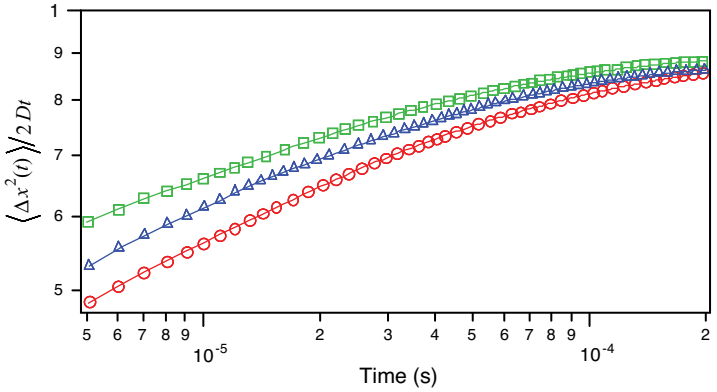


Figure 3.10 Comparison of the motion of particles with the same radius but different densities. Log-log representation of the measured normalized MSD for a silica sphere ($a_s = 1.97\,\mu m$, $\rho_s = 1.96\,kg\,dm^{-3}$, ○), of a resin sphere ($a_s = 2\,\mu m$, $\rho_s = 1.51\,kg\,dm^{-3}$, △) and a polystyrene sphere ($a_s = 1.94\,\mu m$, $\rho_s = 1.05\,kg\,dm^{-3}$, □). Fits correspond to the continuous line with the respective color.

3.4.2.3 Changing the Particle Density

The direct influence of the particle's inertia can be determined by comparing the motion of particles with approximately the same size but different densities. The MSD was calculated from the measured position fluctuations ($f_{acq} = 1\,MHz$, $t_{tot} = 10\,s$) of a silica sphere ($a_s = 1.97\,\mu m$, $\rho_s = 1.96\,kg\,dm^{-3}$, red line), of a resin sphere ($a_s = 2\,\mu m$, $\rho_s = 1.51\,kg\,dm^{-3}$, blue line) and a polystyrene sphere ($a_s = 1.94\,\mu m$, $\rho_s = 1.05\,kg\,dm^{-3}$, green line) in water with $\tau_f \approx 3.9\,\mu s$. The influence of the trap was again kept minimal at $k \approx 14\,\mu N\,m^{-1}$ for the three different beads. Hence, the MSD can be normalized by its long-time limit $\langle \Delta x^2(t) \rangle_{free} = 2Dt$ in the free regime, when $F_{ex}(t) \approx 0$, as shown in Figure 3.10. Equation 3.27 (continuous lines) corresponds to the best fit of the data (silica ○, resin △, polystyrene □). Here, the inertia of the perturbed fluid does not change, so that τ_f stays constant, while the inertias of the three particles are different and lead to different values of τ_s; $\tau_{s_{silica}} = 1.69\,\mu s$, $\tau_{s_{resin}} = 1.34\,\mu s$ and $\tau_{s_{PS}} = 0.88\,\mu s$. Even though differences in τ_s are in fact small compared to the time resolution of $2\,\mu s$, the contribution expected from the particle's mass can still be detected.

Changing the particle's radius a_s is equivalent to varying both τ_f and τ_s, while keeping τ_s/τ_f constant (data not shown) [34].

3.4.3 Implications of the Existence of Long-time Tails in Nanoscale Experiments

Having demonstrated agreement between Brownian motion theory and OTI data, and studied the influence of parameters which can be varied experimentally, we can derive a rule of thumb to estimate the time range during which the particle's motion

can be considered as effectively free from the influence of the trap. Furthermore, we can determine for how long the inertia of a Newtonian fluid will influence the Brownian probe's motion and prevent it from performing a diffusive random walk inside an optical trapping potential. This sets the bandwidth of OTI for high-resolution single particle tracking to probe locally many different media.

3.4.3.1 Single Particle Tracking by OTI

We define the time τ_{free} starting from which the optically confined MSD given by Equation 3.27 begins to deviate by at least 2% from the free MSD described by Equation 3.17. We record the motion of different polystyrene spheres of sizes $a_s = 0.27$, 0.33, 0.50, 0.74, 1.00 and 1.25 μm confined by potentials with a spring constant ranging from 1 to 100 μN m^{-1}. For stronger traps ($\tau_k < 1$ ms), the data are recorded at 500 kHz during 20 s and calibrated in the range between 2 μs and 1 ms. For softer traps ($\tau_k > 1$ ms), the data are recorded at 200 kHz for 50 s and calibrated between 5 μs and 10 ms. In Figure 3.11, τ_{free} is represented as a function of τ_k in a log-log scale, which allows us to formulate an approximate empirical relation between both time scales [41]:

$$\tau_{free} = \tau_k/20 \tag{3.32}$$

Hence, in the time range [2 μs, τ_{free}], which is limited on one side by the noise floor (see Figure 3.6) and on the other side by $\tau_k/20$, OTI can track the motion of a single free Brownian particle. Under these conditions, the sphere probes only its local environment, for up to three decades in time.

3.4.3.2 Diffusion in OTI

The next issue to address is the question of whether or not the Brownian probe has time to reach a purely diffusive behavior before it becomes confined by the potential of the trap. This is equivalent to studying the influence of inertial effects on the particle's motion in the optical trap at times shorter than τ_{free}. Therefore, we investigate the diffusion of a small sphere, which is expected to perturb the fluid less, and hence τ_f and the hydrodynamic memory effect are smaller. Furthermore, we adjust the trapping potential to be the softest possible, as the time region where the particle's motion is free from the influence of the trap lasts longer for high τ_k.

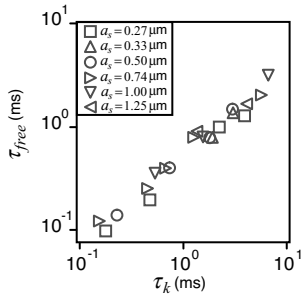


Figure 3.11 Log-log representation of τ_{free} versus τ_k for polystyrene spheres of different radii $a_s = 0.27$, 0.33, 0.50, 0.74, 1.00 and 1.25 μm.

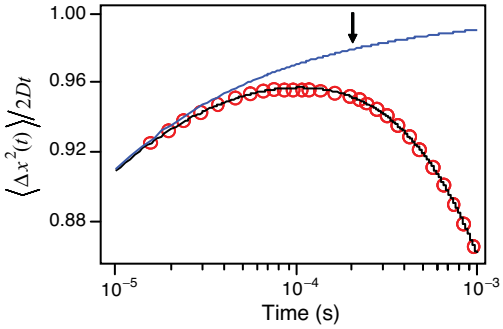


Figure 3.12 $\langle \Delta x^2(t) \rangle / 2Dt$ for the sphere with $a_s = 0.27\,\mu m$ and $k = 1.5\,\mu N\,m^{-1}$, $f_{acq} = 0.2$ MHz, $t_{tot} = 50$ s, fitted by Equation 3.34 (black line). The theory for the free particle is given by the blue line corresponding to Equation 3.24. The arrow indicates the time when $\langle \Delta x^2(t) \rangle_{free} / 2Dt$ reaches diffusive motion within 2%.

We introduce the dimensionless representation $\langle \Delta x^2(t) \rangle / 2Dt$ to distinguish between the free diffusive motion when $\langle \Delta x^2(t) \rangle / 2Dt = 1$ (see Section 3.2.2.1) and the motion influenced by inertial effects when the particle is either free or optically confined (see Section 3.2.2.1, parts (i) and (ii), respectively). In both cases motion is nondiffusive as $\langle \Delta x^2(t) \rangle / 2Dt < 1$.

The measured $\langle \Delta x^2(t) \rangle / 2Dt$ for a polystyrene sphere ($a_s = 0.27\,\mu m$, $k = 1.5\,\mu N\,m^{-1}$) is shown in Figure 3.12 (red circles), fitted to Equation 3.27 (black line) and compared to $\langle \Delta x^2(t) \rangle_{free} / 2Dt$ given by Equation 3.17 (blue line). Here, it can be seen that $\langle \Delta x^2(t) \rangle / 2Dt$ reaches a maximum of ~0.96, but never 1.

For the free particle, $\langle \Delta x^2(t) \rangle_{free} / 2Dt = 1$ would occur within 2% error after approximately 200 μs (Figure 3.12, arrow). Thus, in order to observe free diffusive Brownian motion, the optical trap would have to be so weak that $\tau_{free} > 0.2$ ms is satisfied, or equivalently $\tau_k > 4$ ms according to Equation 3.32. However, for all the combinations of particle sizes and spring constants studied, we could never adjust such a long relaxation time. In the particular case of the sphere with $a_s = 0.27\,\mu m$, a spring constant $k < 1$ nN/m would be needed to observe free diffusive motion for at least one decade in time. However, such a low spring constant does not allow us to trap and observe the particle for a sufficiently long period of time. Hence, in experiments using optical traps, the motion of a particle is influenced by either memory effects and/or by the harmonic potential [41]. This is in contradiction with assumptions commonly made in optical trapping experiments, where a normal diffusive behavior of the trapped particle is assumed and inertial effects from the fluid are ignored [38, 43].

The time-dependent diffusion coefficient can be derived from the VAF or the MSD as:

$$D(t) = \int_0^t \langle \dot{x}(t') \dot{x}(0) \rangle \mathrm{d}t' = \frac{1}{2} \frac{\mathrm{d}}{\mathrm{d}t} \langle \Delta x^2(t) \rangle \tag{3.33}$$

and approaches the diffusion constant D in the infinite time limit when $F_{ex}(t) = 0$:

$$D = \int_0^\infty \langle \dot{x}(t')\dot{x}(0)\rangle \mathrm{d}t' = \frac{k_B T}{6\pi\eta_f a_s} \tag{3.34}$$

3.4.3.3 Thermal Noise Statistics

According to our findings, the process of optically confined Brownian motion as observed by OTI is non-Markovian up to times $t > \tau_k$, when it becomes dominated by the trapping potential. Only from then on does motion become uncorrelated, $\langle \dot{x}(t)\dot{x}(0)\rangle = 0$, and can position data points be considered as statistically independent. It has been proposed that optical trapping data can be calibrated by thermal noise analysis using Boltzmann statistics [44]:

$$p(x)dx = ce^{-E(x)/k_B T} \tag{3.35}$$

which describes the probability $p(x)dx$ of finding the Brownian particle in the potential $E(x)$ (c normalizes the probability distribution).

From the calibrated time traces acquired with $f_{acq} = 0.5$ MHz, the probability distribution is represented in Figure 3.13 for two different resin spheres with $a_s = 0.5\,\mu$m (red histogram) and $a_s = 2\,\mu$m (green histogram) trapped in a similar potential with $k \approx 12.5\,\mu\mathrm{N\,m^{-1}}$ but obviously different τ_k: $\tau_{k_{small}} = 69\,\mu$s and $\tau_{k_{big}} = 798\,\mu$s. The position histograms, with a bin width of 1 nm, are compared for both spheres, and contain either $N = 400\,000$ data points, corresponding to an acquisition time $t_{tot} = 0.08$ s (top graph), $N = 400\,000$ points, corresponding to an acquisition time $t_{tot} = 0.8$ s (middle graph) or $N = 4\,000\,000$ points, corresponding to an acquisition time $t_{tot} = 8$ s (bottom graph).

Even though in each histogram there are apparently enough data points to perform statistical analysis, these points are clearly not statistically independent. Indeed, the upper histogram features only ~100 statistically independent points for the larger sphere and ~1000 for the smaller sphere. The smaller sphere will sample the trapping potential well more rapidly than the larger, which is translated in the differences between $\tau_{k_{small}}$ and $\tau_{k_{big}}$; therefore, the larger sphere will take ~10-fold longer to explore the potential and, consequently, for statistical analysis the trajectory should be acquired over longer times. The temporal resolution of the Boltzmann statistics method is determined by the time required to record uncorrelated data, and is heavily dependent on the particle's size and τ_k. The high acquisition rates used throughout these studies are not needed in this case.

3.5 Summary and Outlook

In this chapter we have shown how OTI can be used to study Brownian motion down to hydrodynamic time scales, where the response of the surrounding fluid becomes dominant. This is only possible due to the high bandwidth (up to 500 kHz) and

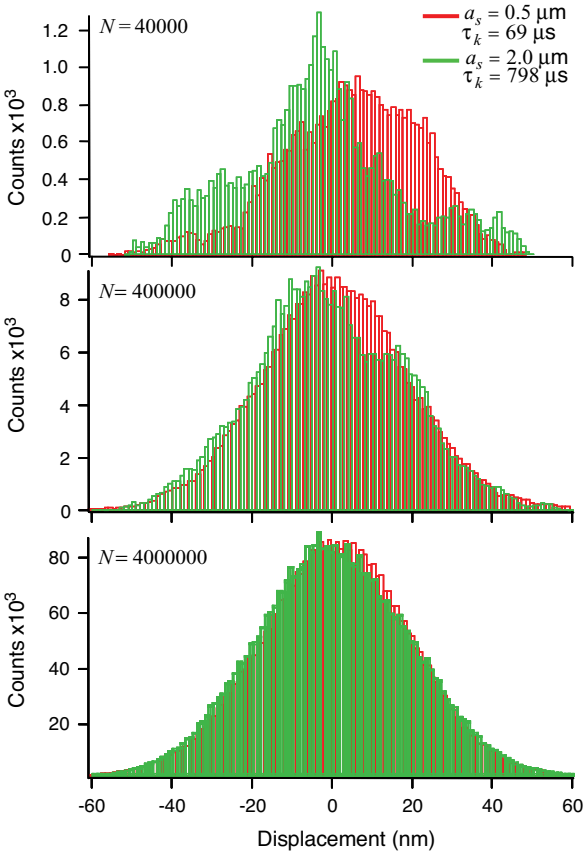


Figure 3.13 Histogram of position fluctuations (bin width of 1 nm) acquired with $f_{acq} = 0.5$ MHz for a sphere with $a_s = 0.5$ µm and $a_s = 2.0$ µm. The number of data points N in the histograms increases from top to bottom from 40 000 ($t_{tot} = 0.08$ s) to 4 000 000 ($t_{tot} = 0.08$ s). The trapping stiffness $k \approx 12.5$ µN m^{-1} is similar for both bead sizes.

subnanometer spatial resolution of the position detection configuration. The precision achieved allows us to detect not only the effects of the fluid's inertia but also the more subtle effects of the sphere's inertia.

The details in the motion of the trapped model sphere provide insight into the interplay between inertial effects and the optical trapping potential, as summarized by Figure 3.14. This shows in particular, the overlap of τ_f, the characteristic time of the

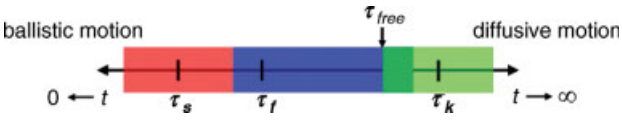


Figure 3.14 Overview of the characteristic times of a Brownian particle confined by a harmonic potential.

viscous fluid, with τ_k, the relaxation time of the restoring force of the trapping potential. The time τ_{free}, which separates τ_f from τ_k, was determined empirically and corresponds to $\sim\tau_k/20$. Below this time, OTI can be used solely as a position detector for *single particle tracking with unprecedented spatial and temporal resolution*.

At these time scales, the sphere performs a non-random walk, dominated by the nature of the surrounding medium [45]. The presented method is capable of providing new insights into the behavior of media that are more complex than just a simple viscous fluid, thus extending the bandwidth of microrheology by two decades in time [46]. For example, the high-frequency response of a viscoelastic polymer solution should provide information on the nanomechanical properties of the polymer molecules. In particular, highly dynamic polymers, such as those encountered in a living cell, should transmit their mechanical and dynamic signatures to the Brownian particle. Furthermore, an obstacle in the particle's trajectory such as a surface, with for example, various chemical functionalities, or a cell membrane, should influence Brownian motion in a characteristic way. Such studies for a variety of biopolymers and surfaces are currently in progress in our laboratory.

Acknowledgments

The authors are grateful to J. Lekki for help in data acquisition, to P. De Los Rios, H. Flyvbjerg, T. Franosch for discussions, and to D. Alexander for reading the manuscript. B.L. and C.G. acknowledge the financial support of the Swiss National Science Foundation and its NCCR. S.J. acknowledges the support of the Gebert Rüf Foundation. The authors also thank EPFL for funding the experimental equipment.

References

1 Haw, M.D. (2002) *Journal of Physics - Condensed Matter*, **14**, 7769.
2 Einstein, A. (1905) *Annalen Der Physik*, **17**, 549.
3 Stokes, G.G. (1851) *Transactions of the Cambridge Philosophical Society*, **9**, 8.
4 Langevin, P. (1908) *Comptes Rendus Hebdomadaires des Seances de L'Academie des Sciences*, **146**, 530.
5 Henri, V. (1908) *Comptes Rendus Hebdomadaires des Seances de L' Academie des Sciences*, **146**, 1024.
6 Vladimirsky, V. and Terletzky, Y. (1945) *Zhurnal Eksperimentalnoi i Teoreticheskoi Fiziki (in Russian)*, **15**, 259.
7 Alder, B.J. and Wainwright, T.E. (1967) *Physical Review Letters*, **18**, 988.
8 Widom, A. (1971) *Physical Review A*, **3**, 1394.
9 Zwanzig, R. and Bixon, M. (1970) *Physical Review A*, **2**, 2005.
10 Bedeaux, D. and Mazur, P. (1974) *Physica*, **76**, 247.
11 Hinch, E.J. (1975) *Journal of Fluid Mechanics*, **72**, 499.
12 Pomeau, Y. and Resibois, P. (1975) *Physics Reports*, **19**, 63.
13 Zwanzig, R. and Bixon, M. (1975) *Journal of Fluid Mechanics*, **69**, 21.
14 Clercx, H.J.H. and Schram, P. (1992) *Physical Review A*, **46**, 1942.

15 Frey, E. and Kroy, K. (2005) *Annalen Der Physik*, **14**, 20.

16 Gittes, F., Schnurr, B., Olmsted, P.D., MacKintosh, F.C. and Schmidt, C.F. (1997) *Physical Review Letters*, **79**, 3286.

17 Boon, J.P. and Bouiller, A. (1976) *Physics Letters A*, **55**, 391.

18 Paul, G.L. and Pusey, P.N. (1981) *Journal of Physics A - Mathematical and General*, **14**, 3301.

19 Ohbayashi, K., Kohno, T. and Utiyama, H. (1983) *Physical Review A*, **27**, 2632.

20 Weitz, D.A., Pine, D.J., Pusey, P.N. and Tough, R.J.A. (1989) *Physical Review Letters*, **63**, 1747.

21 Kao, M.H., Yodh, A.G. and Pine, D.J. (1993) *Physical Review Letters*, **70**, 242.

22 Ashkin, A., Dziedzic, J.M., Bjorkholm, J.E. and Chu, S. (1986) *Optics Letters*, **11**, 288.

23 Berg-Sorensen, K. and Flyvbjerg, H. (2005) *New Journal of Physics*, **7**, 38.

24 Landau, L.D. and Lifshitz, E.M. (1987) *Fluid Mechanics*, Vol. 6, 2nd edn, Butterworth-Heinemann, Oxford.

25 Lorentz, H.A. (1921) *Lessen over Theoretische Natuurkunde, Vol. V*, E.J. Brill, Leiden.

26 Vanderhoef, M.A., Frenkel, D. and Ladd, A.J.C. (1991) *Physical Review Letters*, **67**, 3459.

27 Berg-Sorensen, K. and Flyvbjerg, H. (2004) *Review of Scientific Instruments*, **75**, 594.

28 Uhlenbeck, G.E. and Ornstein, L.S. (1930) *Physical Review*, **36**, 0823.

29 Svoboda, K. and Block, S.M. (1994) *Annual Review of Biophysics and Biomolecular Structure*, **23**, 247.

30 Reif, F. (1985) *Fundamentals of Statistical and Thermal Physics*, McGraw-Hill, Singapore.

31 Henderson, S., Mitchell, S. and Bartlett, P. (2002) *Physical Review Letters*, **88**, 088302.

32 Sterba, R.E. and Sheetz, M.P. (1998) *Methods in Cell Biology*, **55**, 29.

33 Rohrbach, A., Tischer, C., Neumayer, D., Florin, E.L. and Stelzer, E.H.K. (2004) *Review of Scientific Instruments*, **75**, 2197.

34 Lukic, B., Jeney, S., Tischer, C., Kulik, A.J., Forro, L. and Florin, E.L. (2005) *Physical Review Letters*, **95**, 160601.

35 Gittes, F. and Schmidt, C.F. (1998) *Optics Letters*, **23**, 7.

36 Pralle, A., Prummer, M., Florin, E.L., Stelzer, E.H.K. and Horber, J.K.H. (1999) *Microscopy Research and Technique*, **44**, 378.

37 Rohrbach, A. and Stelzer, E.H.K. (2002) *Journal of Applied Physics*, **91**, 5474.

38 Neuman, K.C. and Block, S.M. (2004) *Review of Scientific Instruments*, **75**, 2787.

39 Press, W.H., Flannery, B.P., Teukolsky, S.A. and Vetterling, W.T. (1992) *Numerical Recipes in C*, Cambridge University Press, Cambridge.

40 Rohrbach, A. (2005) *Physical Review Letters*, **95**, 168102.

41 Lukic, B., Jeney, S., Sviben, Z., Kulik, A.J., Florin, E.L. and Forro, L. (2007) *Physical Review E*, **76**, 011112.

42 Guzmán, C., Flyvbjerg, H., Köszali, R., Ecoffet, C., Forró, L. and Jeney, S. (2008) *Applied Physics Letters*, **93**, 184102.

43 Gittes, F. and Schmidt, C.F. (1998) *Methods in Cell Biology*, **55**, 129.

44 Florin, E.L., Pralle, A., Stelzer, E.H.K. and Horber, J.K.H. (1998) *Applied Physics A: Materials Science and Processing*, **66**, S75.

45 Liverpool, T.B. and MacKintosh, F.C. (2005) *Physical Review Letters*, **95**, 208303.

46 Mason, T.G., Ganesan, K., vanZanten, J.H., Wirtz, D. and Kuo, S.C. (1997) *Physical Review Letters*, **79**, 3282.

4
Nanoscale Thermal and Mechanical Interactions Studies using Heatable Probes

Bernd Gotsmann, Mark A. Lantz, Armin Knoll, and Urs Dürig

4.1
Introduction

Thermal properties such as thermal conductivity and diffusivity, although rather difficult to measure, are important properties for many applications. For example, in microelectronics, where power densities can be very high, the local generation of heat and its conduction away from the heated region are a major design issue. In general, the interplay between thermal and mechanical properties of solids is a fascinating topic of research, and is of immediate practical relevance in numerous applications. For example, the mechanical properties of soft matter – namely organic polymers – exhibit such a strong temperature dependence that in standard analysis techniques, such as dynamic mechanical thermal analysis (DMTA), both mechanical stress and temperature are varied. Often, the time scale is also varied, making the analysis rather complicated [1]. The local heating of materials is also used as a micromanufacturing technique. In all of these fields, there is a clear trend to extend research down to the nanoscale, and this is further nurtured by the necessity to understand nanoscale properties in order to tailor materials for nanoscale applications. The trend towards nanoscale opens up research fields and applications beyond conventional materials science. The very definition of temperature, which is a thermodynamic (i.e. statistical) concept based on local equilibrium, becomes vague on length scales smaller than the mean free path of heat carriers (typically in the range of 10–100 nm) [2–4]. On the nanoscale, it is commonly observed that interfaces become more predominant in determining materials properties. This is also true of thermal properties in general, and leads to fascinating concepts such as phonon engineering that promise tailored thermal conductance in nanostructures [5].

In order to study and exploit microscale and nanoscale thermal phenomena, a variety of scanning probe microscopy (SPM) -based techniques have been developed, most of which are based on contact scanning force microcopy (SFM). Two examples of exciting technological applications of these techniques are in the areas of nanoscale data storage [6, 7] and lithography [8]. In this chapter, experimental procedures and

Nanotechnology: Volume 6: Nanoprobes. Edited by Harald Fuchs

ISBN: 978-3-527-31733-2

results from the broad field of heated-probe SFM are addressed. Although the applications of these techniques are rather diverse, two common elements are the use of a sharp, temperature-sensitive tip and the use of SFM techniques to scan this tip over the surface and simultaneously measure the surface topography. Many of these applications also require a means to heat the tip, in turn to heat the sample, on a highly local scale. In the following sections we first describe the various types of probe that have been developed for thermal scanning probe microscopy, and outline the basics of probe-based imaging of thermal properties. Later, we analyze the various heat-loss mechanisms that play a role in the interpretation of thermal SPM data. Specific applications are discussed thereafter, including thermomechanical nanoindentation, data storage and nanopatterning and lithography.

4.2
Heated Probes

At the heart of all the techniques described in this chapter is a heatable probe with an integrated means of sensing the temperature of the probe tip. As with all scanning probe techniques, the resolution is limited at least in part by the geometry of the tip apex and the area of contact between the tip and the surface. The most widely used heated probe is a Wollaston wire probe. In this technique, a thin, bent platinum/rhodium wire is used to produce heat and detect temperature. For SPM-based applications, the wire is bent into the shape of a cantilever, as illustrated in Figure 4.1. Often, also a mirror is glued onto the back of the cantilever to improve optical detection of the cantilever bending. The temperature of the wire can be determined by measuring its electrical resistance and using the

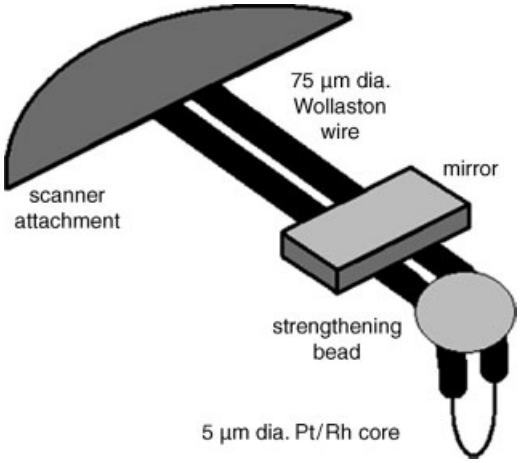


Figure 4.1 Schematic diagram of a thermal probe made from a Wollaston wire process. From Ref. [9].

known temperature dependence of the electrical resistance of the material. Although such probes are accurate and easy to handle, their spatial resolution is limited by the dimensions of the bent wire at the end that acts as the probing tip. To date, the spatial resolution reported using such probes is limited to about ~100 nm (see for example Figure 4.5).

The spatial resolution can be improved by using microfabrication techniques to produce cantilevers with very sharp, temperature-sensitive tips. Shi *et al.* [10] have made such probes using silicon nitride for the cantilever body and microfabricating a platinum tip with a junction to chromium near the tip apex (see Figure 4.2). This junction acts as a thermoelement and can be used to measure the temperature of the tip as it is being scanned over a heated surface, or to investigate local heat sources on a sample. Majumdar *et al.* have used such tips to image hot spots in a very-large-scale integration (VLSI) chip [11] and to image the heat generated by the current flowing through a carbon nanotube [12]. In the latter experiment, an impressive lateral resolution of 50 nm was demonstrated.

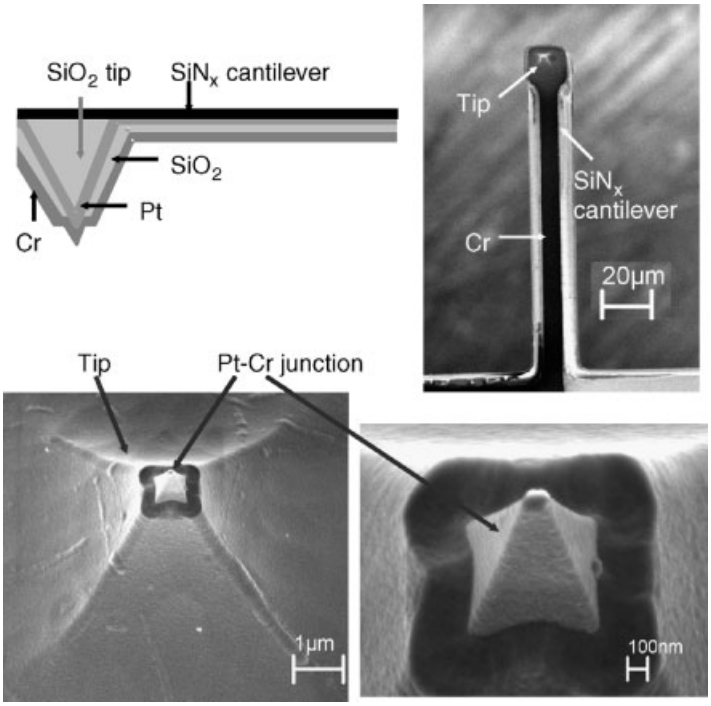


Figure 4.2 Microfabricated probes for scanning thermal microscopy (SThM). Cross-section (upper left) and scanning electron microscopy images of a SThM probe (upper right), the probe tip (lower left) and the Pt-Cr junction (lower right) at the end of the tip. (Reprinted from Ref. [10]; © 2002, American Society of Mechanical Engineering.)

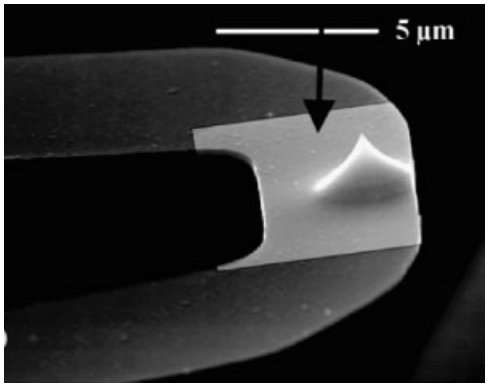


Figure 4.3 Scanning electron microscopy image of a cantilever with integrated heater (artificially contrasted) and tip. (Similar probes were used to obtain the image in Figure 4.7b). (Reprinted from Ref. [14]; © Springer-Verlag.)

Another approach, developed by IBM, uses an all-silicon microfabricated cantilever with an integrated heater and tip (see Figure 4.3) [13]. Here, the largest part of the two-legged cantilever is made from highly doped silicon ($10^{20}\,\mathrm{cm}^{-3}$ As), whereas the part of the cantilever that supports the tip is lower-doped ($5 \times 10^{17}\,\mathrm{cm}^{-3}$) and serves as both heater and sensor. With dimensions of $4\,\mu\mathrm{m} \times 6\,\mu\mathrm{m}$ and a thickness of $\sim$200 nm, the heater has a resistance of few kΩ. The known temperature dependence of the resistivity of doped silicon can be exploited to sense the heater temperature. This type of thermal probe has been used in the majority of the experiments described in this chapter.

The temperature calibration of the heater can be carried out by measuring a current–voltage response curve (I–V curve) of the cantilever (Figure 4.4). For this, a resistor is typically placed in series with the cantilever. The measured voltage drop

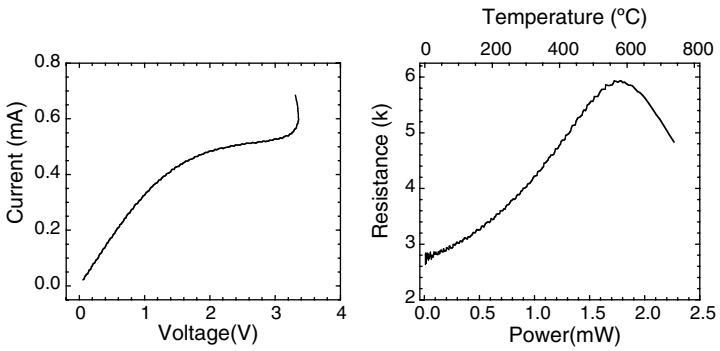


Figure 4.4 Current–voltage (I–V) curve and derived power–resistance (P–R) and temperature–resistance (T–R) curves of an integrated resistive heater.

across the resistor is used to determine the current flowing through the cantilever and, in combination with the measured voltage drop across the cantilever, can be used to calculate the electrical power dissipated in the cantilever. Initially, as the current flowing through the heater is increased, the dissipated power results in an increase in resistance due to increased scattering of the carriers. As the temperature rises, the number of thermally generated carriers also increases, which tends to reduce the rate at which the resistance increases with temperature. When the number of thermally generated carriers equals the number of dopants, the resistance reaches its maximum value, and begins to decrease with further increases in power and temperature. The temperature at which the maximum resistance occurs is a function of the doping density and is known from the literature [15]. The power needed to reach the maximum resistance, $P_{R\max}$, is determined from the measured I–V data. It is assumed that all of the power dissipated in the cantilever contributes to heating of the heater structure, and that the temperature change of the heater is a linear function of the dissipated power. We can then rescale a plot of resistance versus power to a plot of resistance versus temperature using two known values, namely, the resistance at room temperature measured at low very low power, and the temperature at which the maximum resistance occurs. For the doping values used here, the maximum resistance occurs at 550 °C, and the heater temperature can thus be calculated using

$$T_{\text{heater}} = \text{RT} + P(550^{\circ}C - \text{RT})/P_{R\max} = \text{RT} + R_{\text{th}}P,$$

where RT is room temperature.

The implicit assumption here is that the thermal resistance of the system, R_{th}, is independent of the heater temperature, T_{heater}. We find that R_{th} is typically on the order of $10^5\,\text{K}\,\text{W}^{-1}$ under ambient conditions. When checking all of these assumptions by measuring the temperature of the heater by independent means, we found that the resulting systematic error is far below the measurement errors, fabrication tolerances and scatter. Using this calibration method, we estimate an absolute error of about 30% for the temperature difference $\Delta T = T_{\text{heater}} - \text{RT}$. Relative measurements and temperature changes, however, can be conducted with a temperature resolution of ~0.1 K.

The time scale that these heaters are able to probe is related to the thermal equilibration time of the cantilevers. Although the dynamics is not a simple RC-type exponential [14, 16], it can be approximated by a simple exponential and a single time constant. For the cantilevers used in most of the examples discussed below, the time constant is on the order of 7 to 10 µs; this is then the time constant that limits fast thermal sensing. For a rapid application of heat, however, the heater can be operated in nonequilibrium on timescales down to 1 µs and lower [14]. The transient temperature can be sensed reliably at any time scale by means of the electrical resistance. By design, the time constant can be decreased to values below 1 µs, but there are trade-offs between power consumption, sensitivity, time constant, ease of fabrication and the mechanical stability of the cantilever [16, 17]. Therefore, the optimum design depends critically on the application envisaged. For a detailed study of cantilever designs and time constants, the reader is referred to Ref. [14].

4.3
Scanning Thermal Microscopy (SThM)

The thermal conductivity/diffusivity of a sample can be measured by scanning a heated tip on the sample surface of interest and monitoring the heat flow between probe and surface. The heater/tip is usually mounted on a cantilever so that the surface topography can be measured simultaneously with the thermal signal by applying atomic force microscopy (AFM) techniques. This so-called scanning thermal microscopy (SThM) method has recently been reviewed [10, 18–22].

In general, the heater is also used as a temperature sensor so that changes in the heat flow can be measured during scanning. In the Wollaston wire approach, this is achieved by measuring a thermo-voltage, whereas in the IBM approach the temperature-dependent heater resistance is measured; when the tip is in contact with the surface, the contact forms a heat-loss path. The corresponding thermal conductivity can be visualized by measuring changes in the tip temperature that result from changes in the thermal conductivity as the tip is scanned over the surface. However, to be able to sense this local heat loss, the conductivity must be large enough to produce a measurable signal if compared with the electrical noise in the transducer. An example of a local thermal conductivity map of a polymer blend by Reading *et al.* [18] is shown in Figure 4.5. In such a SThM image the color contrast depends on the local thermal conductivity.

Using this technique, it is relatively straightforward to produce a qualitative image of relative differences in thermal conductivity. Quantitative measurements, however, are significantly more challenging and require knowledge of both the tip–sample contact geometry and all of the various thermal resistances and parasitic heat-loss paths in the system (see Section 4.4). In contrast to the electronic case, heat paths

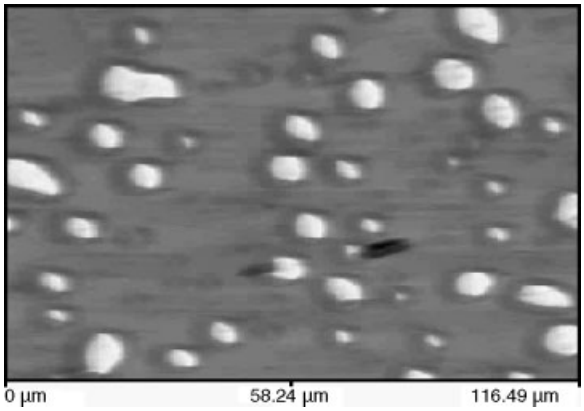


Figure 4.5 Scanning thermal microscopy image of a polymer blend. The two phases can be clearly distinguished in the thermal signal. (Reprinted from Ref. [18]; © 1998; International Scientific Communications, Inc.)

cannot easily be insulated, and therefore a heater typically has several heat-loss paths. The main heat-loss paths, which do not vary with the local sample conductivity, are conduction through the cantilever legs, nonlocal air conduction between heater and sample surface, and radiation cooling. Another potential heat-loss path is through a water meniscus that can form at the point of contact between tip and surface. This will effectively increase the heat-conduction cross-section, leading to a reduced lateral resolution. After performing experiments under ambient conditions, Shi *et al.* [10] have concluded that heat transfer between heater and sample is dominated by conduction through a water meniscus. The heat transfer paths are discussed in more detail in Section 4.4.

The method described above can be refined by modulating the heater drive voltage, which results in a modulation of the heat flow to the sample. The resulting ac component of the heater temperature can be measured using a lock-in amplifier and used to produce an ac thermal image. This ac heat-loss signal changes depending on how the modulation period compares with the diffusion time of heat in the tip-surface contact region – that is, lower-frequency signals diffuse further than do high-frequency signals. Thus, by varying the modulation frequency of the heater, the probing depth can be controlled. This can be seen in Figure 4.6, which shows two ac thermal images taken at 1 and 30 kHz. The sample consists of islands of high-thermal-conductivity material surrounded by low-thermal-conductivity material, both covered by a polymer layer. In the image taken at 1 kHz, the ac signal probes below the polymer layer and strong material contrast is observed, whereas at 10 kHz both probing depth and contrast are significantly reduced.

The limits of the SThM technology are not easily defined. There is a trade-off between time and temperature resolution on the one hand, and spatial resolution on the other hand. For example, increasing the contact area between the tip and sample increases the thermal conductivity, resulting in larger signals and therefore improved sensitivity – but at the expense of reduced lateral resolution. As will be shown below (Section 4.4), the high thermal impedance of the tip and the tip–surface interface make working on samples with high thermal conductivity challenging. Ideally, the

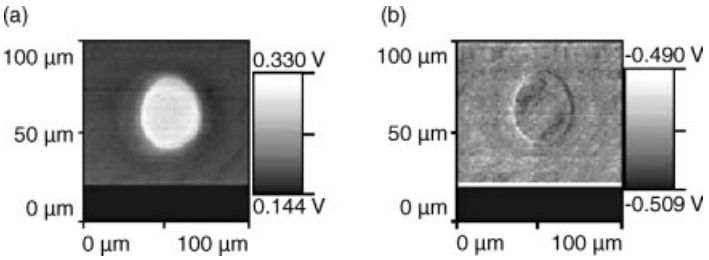


Figure 4.6 Two ac thermal images of a sample with an island of high-thermal-conductivity material within a matrix of low-thermal-conductivity material, both covered by a polymer coating. Image (a) was taken at 1 kHz and image (b) at 30 kHz. (Reprinted from Ref. [18]; © 1998, International Scientific Communications, Inc.)

thermal impedance of the tip and tip–surface interface should be comparable to, or smaller than, those of the sample. The tip and tip–sample interface impedances increase as the tip is made sharper and the contact area reduced, making high-spatial-resolution experiments challenging. Nevertheless, a spatial resolution in the range of some tens of nanometers is feasible on some samples. For example, Shi *et al.* have demonstrated a resolution better than 100 nm on metallic wires [10] and better than 50 nm when imaging a carbon nanotube [12]. A challenge for the quantitative analysis of SThM images is the unknown interaction volume under the probing tip [19]. Nevertheless, the method is very successful in the study of polymers and biological samples. For a recent review, see Ref. [19].

An example of high-lateral-resolution SThM is given in Figure 4.7. Here, the very small contrast between two materials of similar thermal conductivity (silicon oxide and hafnium oxide) is observed. The sample consisted of 2 nm-thick islands of SiO_2 surrounded by a 3 nm-thick film of HfO_2 on a single-crystal silicon substrate. Note that both materials have a considerably higher thermal conductivity than polymers. The measurements were performed in a high-vacuum environment using silicon probes with integrated silicon heaters (as described in Section 4.2). A lateral resolution of ~25 nm was achieved, and the previously unknown thermal conductivity of the 3 nm-thick HfO_2 film was determined. This example nicely demonstrates the potential of using SThM for quantitative measurements, even at high spatial resolution.

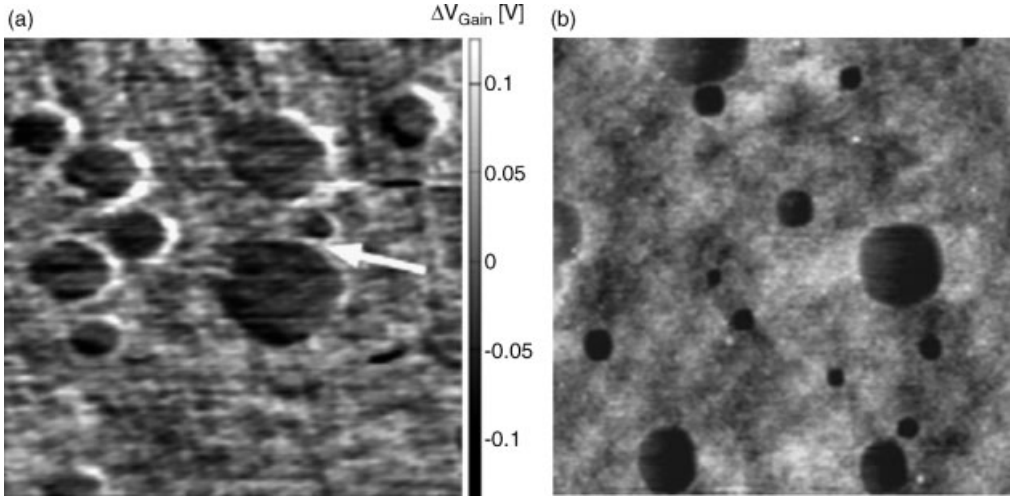


Figure 4.7 (a) Scanning thermal microscopy image (1.6 µm × 1.8 µm) and (b) topography image (2 µm × 2 µm × 5 nm) of a HfO_2 film on a Si substrate. The round holes are filled with 2 nm-thick SiO_2. From the image contrast between the HfO_2 and the SiO_2 regions, the thermal conductivity of HfO_2 can be determined quantitatively. (Images reproduced from Ref. [23].)

4.4
Heat-Transfer Mechanisms

Most of the applications described in this chapter use a sharp, heated probe to deliver heat to a surface on a highly localized scale. However, this process is often rather inefficient because of the high thermal resistance of nanometer-sharp tips and nanometer-sized contact areas, in combination with the other parasitic heat-loss paths present in the system.

In this section, we analyze the various heat-loss paths and mechanisms that can play a role in heated-probe experiments. The analysis in this section is taken from some unpublished results of U. Dürig and B. Gotsmann. The majority of heat-loss paths are undesirable in the sense that they do not contribute to the image contrast in SThM and reduce the efficiency of delivering heat to the sample in other applications. The various heat-loss mechanisms that can contribute to heat loss from a heated tip are illustrated in Figure 4.8 and described in more detail in the corresponding Sections 8.4.1–8.4.6. Which of these mechanisms contribute in a given experiment and the relative magnitudes of their contributions depend on the details of the cantilever design and the experimental conditions. The potentially undesirable heat-loss mechanisms include conductive heat loss through air and the cantilever (Section 4.4.1) and also thermal radiation (Section 4.4.2). Heat conduction through the tip is desirable, but the thermal resistance of the tip (Section 4.4.4) and conduction through a water meniscus (Section 4.4.3) that can form between the tip and sample limit the sensitivity and resolution. The thermal resistance of the tip–surface interface (Section 4.4.6) and the thermal spreading resistance in the sample (Section 4.4.5) are material-specific and determine the image contrast in SThM. The relative magnitudes of these resistances also play an important role in determining the efficiency of heat delivery to the sample. As a quantitative example, we calculate the thermal impedances of the various heat-loss paths for the microfabricated silicon cantilever with integrated silicon heater described at the end of Section 4.2. Finally, in Section 4.4.7, we describe a set of experiments designed to quantify the various thermal resistances.

4.4.1
Heat Transport Through the Cantilever Legs and Air

Microfabricated silicon heater structures integrated into a cantilever structure typically have a minimum size on the order of micrometers. Usually, they are integrated into 'u'-shaped cantilever structures that provide both mechanical support and electrical connections to the heaters. These cantilever structures result in additional heat-loss paths that, however, do not go through the tip but rather into the support structure and from there into the surrounding air. If the cantilever is close to a sample surface, then a fraction of this heat will be conducted through the air into sample, but will not lead to a significant temperature increase in the sample owing to the distributed nature of the heat transfer. The thermal coupling between cantilever and sample is relatively strong if the cantilever–surface distance is comparable to the

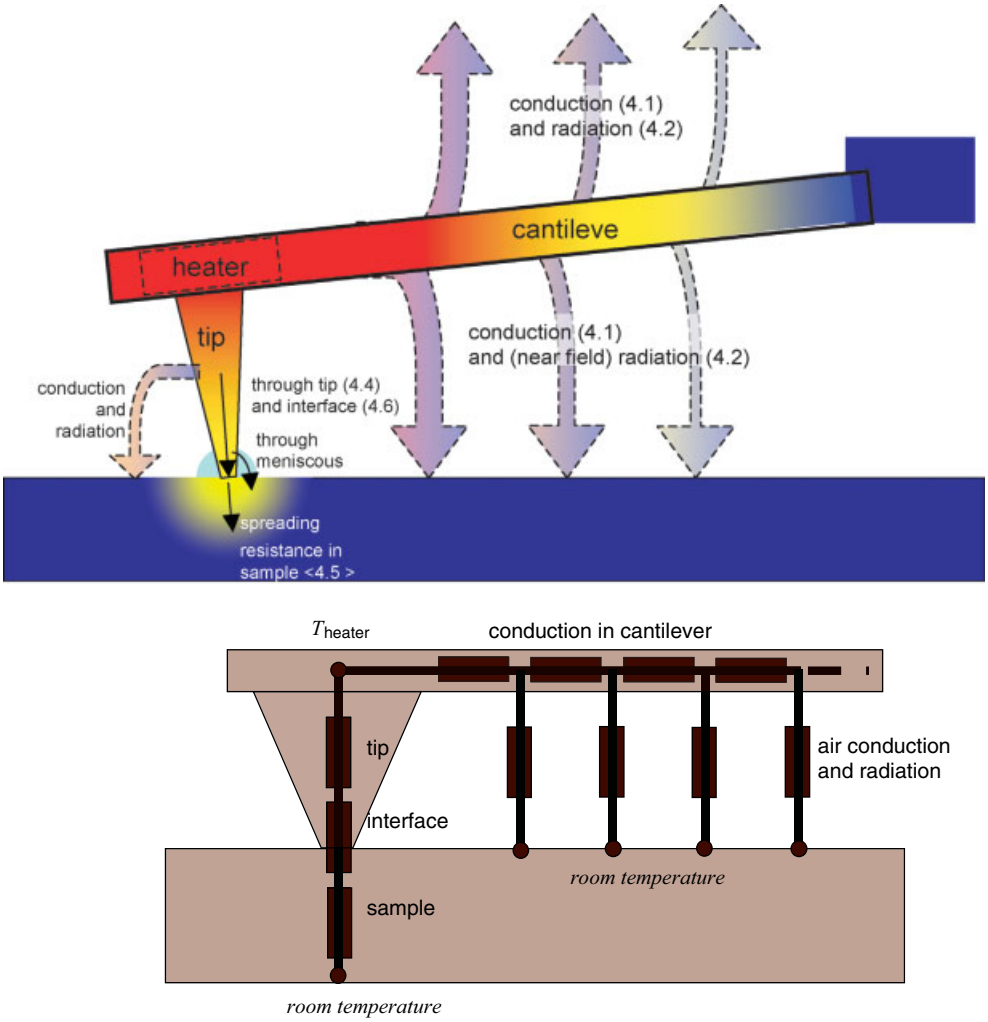


Figure 4.8 (a) Heat paths relevant for experiment using heated probes (numbers refer to text sections in which they are described); (b) Schematic representation as thermal resistances. Note that the distinction between tip, interface and spreading resistances is not possible in every case.

mean free paths of air molecules (~60 nm). For the cantilever design shown in Figure 4.3, the tip height is ~500 nm, resulting in a strong coupling to the substrate. For cantilevers that are long relative to the dimensions of the heater and to the cantilever–surface distance, most of the heat is conducted through the air and into the substrate. For the cantilever in Figure 4.3, the conductivity through cantilever allows the heat to spread along the cantilever by a distance on the order of a few tens of micrometers. Heat conduction along the cantilever and into the air is analogous to a

lossy transmission line. Thus, it can be modeled as a series of thermal resistances, describing conduction through the cantilever with a set of parallel resistances at each node that give the conduction through the air to the sample, as illustrated in Figure 4.8b. Earlier studies of the heat transfer through the cantilever–air gap and within the silicon cantilever/heater [14, 16] predicted that for typical dimensions (see Figure 4.3) – that is, a cantilever thickness of ~200 nm – a cantilever/heater–surface distance of 500 nm, a heater size ~5 × 5 μm and a cantilever width of ~5 μm, the thermal resistance of the combined air/cantilever heat loss path is on the order of $10^5\,\mathrm{K\,W^{-1}}$ and the thermal response times are approximately 10 μs. If the cantilever is operated in a vacuum environment, heat loss through the air is eliminated and heat flows directly through the cantilever to the thick silicon cantilever support structure. For the cantilever design shown in Figure 4.3, the thermal resistance of this heat loss path is on the order of $5\text{–}10 \times 10^5\,\mathrm{K\,W^{-1}}$.

4.4.2
Heat Transfer Through Radiation

Heat loss due to thermal black-body radiation, which involves the propagation of electromagnetic waves from a hot object, is described by the Stefan–Boltzmann equation

$$S = \frac{\pi^2 k_B^4}{60\hbar^3 c^2}\left(T_1^4 - T_2^4\right).$$

Here, the cooling power per area, S, is expressed in terms of the Boltzmann constant, k_B, the Planck constant, $\hbar$, the speed of light, c, the temperature of the heated body, T_1, and the temperature of the environment, T_2. For a heater temperature 100 K above the environment temperature, this corresponds to a thermal resistance of $\sim 6 \times 10^8\,\mathrm{K\,W^{-1}}$ for effective heater dimensions of $(9\,\mu\mathrm{m})^2$. Compared to the thermal resistance of the cantilever and air heat-loss paths of about $1\text{–}10 \times 10^5\,\mathrm{K\,W^{-1}}$, the contribution due to black-body radiation is negligible. In ambient conditions, the overall thermal resistance is dominated by air conduction and in vacuum by conduction through the legs to the support structure.

In heated-probe SPM experiments, the distance between the heater and sample is often less than 1 μm. In this case, the Stefan–Boltzmann equation is only an approximation, and near-field effects must be taken into account. Such effects have a long history of theoretical analysis (see for example Refs [24–28]). Experimentally, however, the effect appears difficult to pin down, and very few reports have been made [29–32]. It has been predicted on a theoretical basis that, compared with Stefan–Boltzmann's law, heat transport by evanescent thermal radiation will depend heavily on the materials involved, with a strong distance dependence ($1/d^2$ for most cases) and a weakened temperature dependence (T^2 for most cases) [25]. The effect is also heavily dependent on the dielectric constants of the heater and the sample material. According to theory, we can expect that the near-field effect for a polymer surface is very small. However, for a silicon surface it can be significantly higher, depending on the doping [26–28], but even in this case the effect is expected to be negligible when compared to the total thermal resistance of the cantilever.

Under ambient conditions, the distant-dependent cooling of the heater/cantilever is dominated by air cooling, and therefore it is not possible to observe near-field cooling effects. Under vacuum conditions, the contribution to cooling due to thermal radiation should become measurable – not so much because of the increased overall thermal resistance without air conduction but rather because we can use the distance dependence to demonstrate the existence of near-field cooling. In air, the distance dependence is dominated by air conduction, whereas in a vacuum the air conduction is of course eliminated and conduction through the cantilever legs does not depend on the heater–sample distance. Thus, on approaching a surface in vacuum, any variation in the thermal resistance that is observed prior to tip–surface contact can likely be attributed to near-field radiation effects.

4.4.3
Thermal Resistance of a Water Meniscus

Under ambient conditions, humidity in the air usually results in the formation of a water meniscus around the tip–sample contact. The size and thermal conductance of the meniscus are a function of both humidity and sample material, and therefore are difficult to control. Thermal conduction through such a water meniscus effectively increases the tip–sample contact area and thereby reduces lateral resolution. On the other hand, the meniscus improves thermal contact, especially on rough surfaces, and may even be necessary to make nanoscale measurements possible in the first place. In a groundbreaking report, Shi and Majumdar concluded that in experiments using a $\sim$100 nm-diameter metal tip on a metal surface, the influence of the water meniscus is of the same order of magnitude as the conduction through the solid–solid tip–sample contact [10]. Moreover, they also concluded that for relatively blunt tips under ambient conditions, thermal conduction through the air gap between the tip sidewalls [33] predominates, whereas for sharper tips, solid–solid and water–meniscus conduction dominate [10]. The effects of conduction through a water meniscus can of course be avoided by operating the heated tip in a low-humidity or vacuum environment.

4.4.4
Heat Transfer Through a Silicon Tip

The thermal resistance of the tip stems from the conductance of phonons in the silicon tip and from the layer of native oxide covering it. In the tip, the resistivity is larger than that of bulk silicon because of enhanced phonon scattering at boundary surfaces [3]. The thermal resistance of the silicon tip can be estimated using predictions for the thermal resistivity of silicon nanowires as a function of diameter [34]. Integrating the expression for the varying diameter of a cone-shaped tip with a typical opening angle of 50° down to the apex with a radius of 5–10 nm yields a thermal resistance on the order of 10^6 to 10^7 K W^{-1} because of phonon scattering. This is in agreement with finite-element calculations [35].

To develop a 'hands-on' feeling for this rather complex subject, we first derive a simple model for heat conduction in conical structures, in particular with regard to ballistic phonon transport. Let us initially consider a cylindrical rod with a cross-section $A = d^2\pi/4$ (d = rod diameter). Let us further assume that a constant current I_{th} of thermal energy flows through the rod, which is driven by a temperature difference along the rod axis (x-axis). For a temperature gradient $\Delta T/\Delta x$, the heat flow is)

$$I_{th} = \frac{\kappa A}{\Delta x}\Delta T, \tag{4.1}$$

where κ denotes the thermal conductance of the rod and ΔT is the temperature difference across a cylindrical slice of thickness Δx. In using Equation 4.1, we assume that the phonon mean free path is large compared with the diameter of the rod. For very thin rods, say $d < 100$ nm for crystalline Si, this assumption is no longer valid and one must account for phonon scattering at the boundaries by renormalizing the thermal conductance according to the so-called Mathiessen's rule [36, 37]:

$$\kappa' = \kappa\frac{1}{1+\frac{\lambda}{d}} \approx \kappa\frac{d}{\lambda}, \quad d \perp \lambda, \tag{4.2}$$

where λ denotes the phonon mean free path.

Let us now consider a rotational symmetric, conical conductor with opening angle Θ (see Figure 4.9). The diameter of a circular slice at distance x from the cone tip is thus $d(x) = 2x|\tan(\Theta/2)$. If $d \gg \lambda$, we can calculate the temperature profile along the cone axis by approximating the cone by a stack of short cylindrical rods of length $\Delta x \gg \lambda$ and diameter $d(x)$, yielding ($x_1 - x_2\ \pi\ \Delta x$):

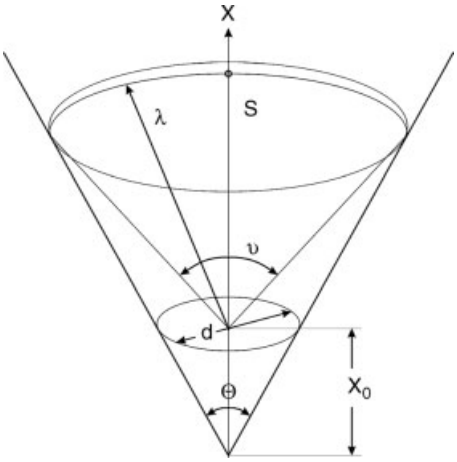


Figure 4.9 Schematic of the conical tip: Θ is cone angle, d the cone diameter at x_0, λ the phonon mean free path, and υ the opening angle for the intersection of the phonon mean free path with the surface of the cone.

$$T_1 - T_2 = I_{\text{th}} \frac{1}{\pi \tan\frac{\Theta}{2}} \int_{x_1}^{x_2} \kappa \frac{dx}{x^2}. \tag{4.3}$$

The cylindrical rod approximation has also been applied for d (λ by simply replacing the constant thermal conductivity κ by the renormalized value κ' [36]. However, it is not obvious that this approach is applicable because the cone cross-section changes significantly over a distance λ along the tip axis. Therefore, let us examine the problem in more detail.

Consider a point on the cone axis at a distance x_0 from the apex (see Figure 4.9). We assume that $d < \lambda$. Let S be the spherical cap defined by the intersection of the conical tip with a sphere of radius λ centered at x_0. A fraction of the phonons emanating from S arrive at x_0 in a direct path without interference from the tip surface. These unperturbed phonons impinge from a solid angle

$$\tan\frac{\upsilon}{2} = \frac{d(x_0+\lambda)}{2\lambda} = \frac{d(x_0)}{2\lambda} + \tan\frac{\Theta}{2}. \tag{4.4}$$

The fraction of the total heat transport seen at x_0 due to these direct phonons is

$$\begin{aligned} \eta^d &= \frac{I_{\text{th}}^d(\upsilon)}{I_{\text{th}}^d(\pi)} \\ &= \frac{2\pi \int_0^{\upsilon} \frac{T(x_0) - (x_0+\lambda)}{dT/dx \cdot \lambda} \cos\upsilon/2 \sin\upsilon/2 \, d\upsilon/2}{2\pi \int_0^{\pi} \cos\upsilon/2 \sin\upsilon 2 \, d\upsilon/2} \\ &= \frac{T(x_0) - T(x_0+\lambda)}{dT/dx \cdot \lambda} \sin^2\upsilon/2 \approx \frac{d(x_0)}{\lambda} \sin^2\upsilon/2. \end{aligned} \tag{4.5}$$

The factor $(T(x_0) - T(x_0 + \lambda))/(\lambda dT/dx)$ accounts for the reduced thermal energy carried by the impinging phonons with respect to the value calculated from the local thermal gradient at x_0. As the temperature difference $T(x_0) - T(x_0 + \lambda) \cong T(x_0) \propto 1/d^2$ (see below), the factor is equal to d/λ. Similarly, for the fraction of heat transported by the phonons scattered off the wall, one can write

$$\begin{aligned} \eta^d &= \frac{I_{\text{th}}^w(\upsilon)}{I_{\text{th}}(\pi)} \\ &= \frac{2\pi \int_0^{\pi} \frac{\Delta T^w(\upsilon')}{\Delta T(\upsilon')} \cos\upsilon'/2 / 2 \sin\upsilon'/2 \, d\upsilon'/2}{2\pi \int_0^{\pi} \cos\upsilon'/2 \sin\upsilon'/2 \, d\upsilon'/2} \\ &\approx \frac{d(x_0)}{\lambda} \cos^2\upsilon/2. \end{aligned} \tag{4.6}$$

As for the direct phonons, the factor $\Delta T^w/\Delta T$ denotes the fraction of heat carried by a phonon scattered from the wall with respect to a thermal equilibrium phonon. A calculation (A. Dürig, unpublished results) yields

$$\Delta T^w(\upsilon') \approx \Delta T(\upsilon') \begin{cases} \frac{d(x_0)}{\lambda} \; \frac{d(x_0)}{\lambda} < 1 \\ 1 \; \frac{d(x_0)}{\lambda} \geq 1 \end{cases}, \tag{4.7}$$

where the equality holds for $\upsilon = 0$ and deviations for $\upsilon > 0$ have been neglected. Hence, we obtain as final result

$$\kappa'(x_0) = \kappa(\eta^d + \eta^w) \approx \kappa \frac{d(x_0)}{\lambda}, \tag{4.8}$$

in exact agreement with Mathiessen's rule.

According to Equation 4.3, the thermal resistance of a conical heat conductor with an apex diameter $d_0 \perp \lambda$ can be written as

$$\begin{aligned} R &= \frac{1}{\kappa}\frac{4\lambda}{\pi} \int\limits_{d_0 \ll \lambda}^{\lambda} \frac{1}{d^3}\, dx \\ &\approx \frac{1}{\kappa}\frac{2\lambda}{\pi \tan\Theta/2}\frac{1}{d_0^2} \\ &= \frac{3}{8}\frac{1}{\tan\Theta/2} R_s, \end{aligned} \tag{4.9}$$

where

$$R_s = \frac{1}{\kappa}\frac{4\lambda}{3\pi}\frac{1}{(d_0/2)^2} \tag{4.10}$$

is the so-called Sharvin resistance for ballistic transport through a circular aperture (see e.g. Ref. [38]). It is interesting that, despite the short effective mean free path due to boundary scattering, the resistance of a conical conductor still retains the characteristics of ballistic transport expressed by the inverse d_0^2 dependence. Also, the length of the cone does not enter because more than 90% of the temperature change occurs over a distance on the order of three times the apex diameter. Substituting the corresponding values for the thermal conductivity $\kappa = 165\ \mathrm{W\,K^{-1}\,m^{-1}}$ and the mean free path $\lambda = 100\ \mathrm{nm}$, one obtains the following for the thermal resistance of a Si tip:

$$R \approx 3.86 \times 10^8\ \mathrm{KW^{-1}\,nm^2}\, \frac{1}{\tan\Theta/2 d_0^2}. \tag{4.11}$$

The cone angle dependence and explicit values of R for a representative set of Θ and d_0 values are tabulated in Table 4.1. Note that the tip resistance increases markedly if the cone angle is less than $\sim 45^\circ$.

Next, we investigate the influence of interface scattering at the apex for a tip in contact with a substrate surface. Specifically, let us assume that the substrate is a

Table 4.1 Normalized thermal resistance of a conical tip as a function of the cone angle and thermal resistance for $\Theta = 90^\circ$ as a function of apex diameter.

$R(\Theta)/R(90^\circ)$	1	1.73	2.41	3.73	7.60
Θ	90	60	45	30	15
R (90°) (KW^{-1})	3.86×10^{8}	9.65×10^{7}	1.54×10^{7}	3.86×10^{6}	9.65×10^{5}
d_0 (nm)	1	2	5	10	20

poorly conducting material, such as a polymer film, which has a correspondingly short phonon mean free path $\perp d_0$. The heat transmitted through the apex is carried away radially, where the temperature drop in the substrate is given by the spreading resistance (see e.g. Ref. [38] and Section 4.4.5):

$$R_{\mathrm{sp}} = \frac{1}{2\kappa_s d_0}. \tag{4.12}$$

Here, κ_s denotes the thermal conductance of the substrate.

Consider a heat current impinging on the interface. A fraction η is reflected back into the tip owing to scattering at the apex (see Figure 4.10a). Whether the scattering is elastic or inelastic is not important for the subsequent discussion. Because of the ballistic nature of the propagation, thermalization of the reflected phonon will occur far away from the interface, and therefore will not influence the local thermal equilibrium at the interface. Let T and I_{th} be the interface temperature and the thermal current in the absence of scattering, respectively. The heat transmitted into the substrate is thus $I^t_{\mathrm{th}} = (1-\eta)I_{\mathrm{th}}$. Correspondingly, the substrate temperature at the interface is $T_s = T - \eta(T - T_0)$, where T_0 is the substrate temperature far from the interface.

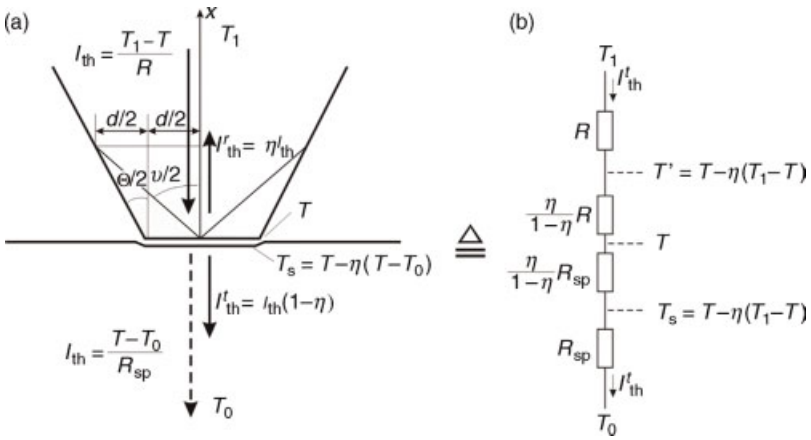


Figure 4.10 Schematic of boundary scattering at the tip apex.

The net heat current flow in the tip is $I^t_{\rm th} = I_{\rm th} - I^r_{\rm th}$, where $I^r_{\rm th} = \eta I_{\rm th}$ denotes the reflected current in the tip. As argued above, the tip temperature at the apex is not altered. Nevertheless, we introduce an effective temperature at a virtual tip interface $T' = T + \eta(T_1 - T)$, where T_1 is the tip temperature far from the interface (see Figure 4.10b). With this definition, the heat balance can be satisfied with regard to the net current, $T' = T_1 - I^t_{\rm th} R$, where the heat conduction in the tip is represented by the single resistive element, R. The virtual tip interface is connected to the substrate surface by means of a resistive element that must satisfy the equation $T' - T_s = I^t_{\rm th} R_b$. Hence, one obtains

$$R_b = \frac{\eta}{1-\eta}(R + R_{\rm sp}). \tag{4.13}$$

The scattering physics is captured in the reflection coefficient

$$\eta = a(1 - \cos\upsilon/2) \approx a\left(1 - \frac{1}{\sqrt{4\tan^2\Theta/2 + 1}}\right), \tag{4.14}$$

where $0 \leq a \leq 1$ is a type of accommodation factor for the back-scattering of the phonons into the tip and $(1 - \cos\upsilon/2$ accounts for the fraction of the solid angle covered by the back-scattered phonons that escape form the tip apex and thermalize in the heat bath. The particular choice for υ is heuristically motivated by the observation that the temperature gradient is roughly one order of magnitude lower at a distance above the apex that corresponds to a cone diameter of twice the aperture diameter.

The substrate temperature is one of the key parameters for studying thermomechanical material properties. It can be written as

$$T_s = \frac{R_{\rm sp}}{R + R_b + R_{\rm sp}}(T_1 - T_0) + T_0 = \frac{1-\eta}{1 + R/R_{\rm sp}}(T_1 - T_0) + T_0, \tag{4.15}$$

Note that already for rather weak back scattering the interface resistance accounts for a significant fraction of the total resistance, that is, $a = 0.3$ yields R_b λ0.5 $(R + R_{\rm sp})$ (see Equations 4.13 and 4.14). Figure 4.11 shows the substrate temperature as a function of tip cone angle for $a = 0.5$ and 0.75 and for various ratios of $R + R_{\rm sp} = (\kappa_s\lambda)/(\kappa d_0) < 1$ corresponding to cases in which the spreading resistance is dominating the tip resistance. Under such conditions, the substrate temperature is rather insensitive to the values substituted for the accommodation coefficient a and the cut-off angle υ.

As observed experimentally [8, 14, 35, 39, 40] and also described in Section 4.4.8, the simple model predicts that the temperature rise at the substrate interface is on the order of 0.4 to 0.7 times the total temperature difference between tip and substrate for parameters that correspond to typical experimental conditions. It is clear that the model cannot capture the complexities of phonon scattering in a predictive manner. Instead, a phenomenological parameter a must be introduced to match experimental observations with model predictions. However, the model provides a means for assessing the scaling properties and provides qualitative

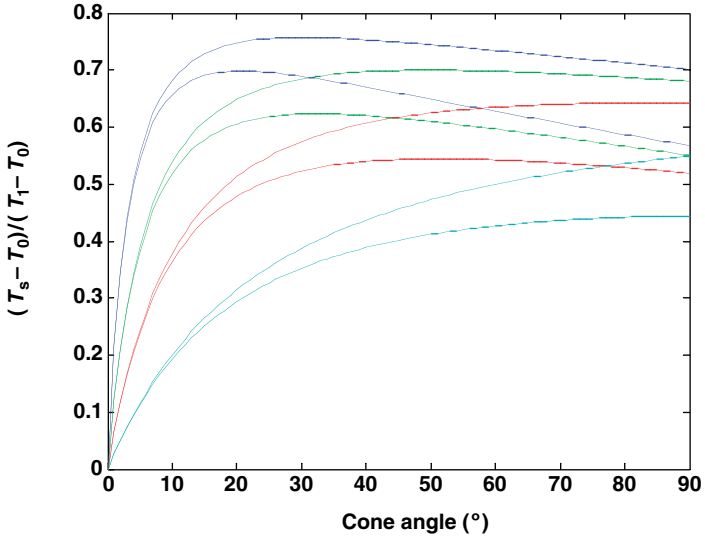


Figure 4.11 Substrate temperature at the tip apex: $a = 0.75$ (solid lines) and 0.5 (dashed lines) and $R/R_{sp} = (\kappa_s\, \lambda)/(\kappa\, d_0) =$ 0.025 (blue), 0.05 (green), 0.1 (red) and 0.25 (cyan).

insight based on intuitive physical arguments. A deeper discussion of physical mechanisms governing thermal transport of nanometer-scale tip–surface contacts is presented in Sections 4.4.5–4.4.8.

4.4.5
Thermal Spreading Resistance

The spreading resistance in the sample is probably the best understood of all the thermal resistances involved. Commonly, it is well approximated by Equation 4.12, which says that the resistance scales inversely with the contact diameter d_0. The scaling is borne out by the fundamental heat conduction Equation 4.1 by observing that the mean gradient $\Delta T/\Delta x$ scales as $1/d_0$ for diffusive transport in a half-space.

For a thin film on a substrate, one can account for the effect of the substrate by using an approximate solution proposed by Yovanovich *et al.* [41]:

$$R_{sp} = \frac{1}{2\kappa_s d_0} - \frac{1}{2\pi\kappa_s t}\log\left(\frac{2}{1+\kappa_s/\kappa_{sub}}\right). \tag{4.16}$$

Here, κ_s and t are the thermal conductance and the thickness of the film, respectively, and κ_{sub} denotes the thermal conductivity of the substrate on which the film has been deposited. In all experiments discussed below, the film thickness is at least one order of magnitude larger than the contact diameter, and we can disregard the finite-size correction term.

For polymer films, a value of $\kappa_s \sim 0.2$–$0.3\,\mathrm{W\,K{\cdot}m^{-1}}$ is typical. The thermal conductance can increase by up to a factor of 2 under a pressure of 1 GPa. As the

stress under the tip varies during an experiment and is transient within the tip–surface interaction volume, we must resort to estimating an effective pressure-increased thermal conductivity [42]. For the experiment described below, we use a value of k_{pol} of 0.3–0.6 $W K \cdot m^{-1}$. For a contact diameter of $d_0 = 10$ nm, we obtain an estimated R_{sp} of approximately $(0.8 - 1.6) \times 10^8 K W^{-1}$ for polymers, $\sim 10^6$–$10^7 K W^{-1}$ for oxides, $\sim 3 \times 10^5 K W^{-1}$ for silicon, and down to $10^4 K W^{-1}$ for metals.

4.4.6 Interface Thermal Resistance

As discussed in Section 4.4.4, the thermal resistance of the interface R_{int} defies accurate prediction. The situation is further complicated because the interface resistance usually depends heavily on the quality of the interface and the contact pressure. For most of the cases described in this chapter, we can assume a single-asperity contact characterized by a contact diameter d_0. Contact mechanics models can be invoked to estimate d_0. For a tip with a radius of 10 nm and applied loads of a few tens of nanoNewtons, d_0 is on the order of a few nanometers.

On the other hand, the values of the interface resistance reported in the literature were measured on macroscopically large areas (rather than a tip–surface contact). Typical values for silicon–polymer interfaces are in the range of 10^{-8} to $10^{-7} Km^2 W^{-1}$ [43, 44]. For a silicon–silicon interface, the corresponding value obtained for phonon scattering is $2.1 \times 10^{-9} Km^2 W^{-1}$. The subject of interface scattering has been extensively reviewed by Swartz and Pohl, who discuss the thermal interface (or boundary) resistance between various materials as well as related models [45].

Returning to the nanoscale tip contact, it is not immediately clear how to relate the macroscopic data to the nanoscale interface resistance. One simple approach adopted by King [35] is to treat the interface as a scattering site for phonons in silicon and to assume that the total interface resistance is inversely proportional to the contact area, which yields $R_{int} = 2.1 \times 10^{-9} Km^2 W^{-1} \times 4/(\pi \ d_0^2) \sim 2.7 \times 10^7 K W^{-1}$ for $d_0 = 10$ nm. Alternatively, if we substitute the measured value for the polymer–silicon interface resistance, we obtain $R_{int} = 10^{-8}$ to $10^{-7} Km^2 W^{-1} \times 4/(\pi \ d_0^2) \sim 1.3 \times 10^8$ to $10^9 K W^{-1}$ for $d_0 = 10$ nm.

Alternatively, it is shown in Section 4.4.4 that the resistance due to boundary scattering at the tip apex is proportional to the sum of the thermal resistances associated with the conduction paths through substrate and tip. Therefore, this has two components: one scaling as $1/d_0$ and corresponding to the spreading resistance in the substrate; and the other scaling as $1/d_0^2$ and corresponding to the tip resistance. The question then arises how this is to be reconciled with the $1/d_0^2$ scaling suggested by extrapolating from the macroscopic scale.

The interface resistance is a somewhat artificial construct, which bridges the gap in the conduction path where the temperature of the phonon gas cannot be defined unequivocally. The temperature is well defined only if the phonons thermalize by means of mutual scattering. Therefore, the gap typically extends over a distance on the order of the phonon mean free path on either side of the interface. For consistency with the concept of a thermal resistance, the interface resistance is defined as the tempera-

ture difference divided by the net heat flux across the gap, as measured by an imaginary observer with an apparatus that is in thermal equilibrium with the phonon gas. Note, however, that unlike a regular thermal resistor, the interface resistance cannot be broken up into a string of series resistors to calculate the temperature at any point along the gap. In fact, such temperatures have no meaning and merely serve as a mathematical concept. The interface temperature of the tip, T' (which was introduced in Figure 4.10b), is an example of such a fictitious temperature. Moreover, the ballistic tip resistance, R (see Equation 4.9 in Section 4.4.4) constitutes part of the overall interface resistance. Therefore, we must write the following for the interface resistance:

$$R_{\text{int}} = R_b + R, \tag{4.17}$$

which spans the entire ballistic propagation path of the phonons through the conical tip, including boundary scattering at the tip–substrate interface up to their thermalization in the substrate. It is also clear from the above discussion that we cannot simply extrapolate from macrosopic results to nanoscale thermal contacts without accurately accounting for the conduction path. What one can do, however, is to extract a mean backscattering probability from macroscopic experiments. Using the same type of reasoning as in Section 4.4.4, one obtains the following for the interface resistance for a unit area of a planar contact:

$$r_{\text{int}} = \frac{\eta}{1-\eta}\frac{\lambda}{\kappa}. \tag{4.18}$$

With $\eta \sim a$, and assuming $\lambda \sim 1$ nm and $\kappa \sim 0.3\ \text{W Km}^{-1}$ for the mean free path and the thermal conductance of polymers, respectively, one must substitute $a \sim 0.75$ to 0.97 in order to obtain the experimentally observed values of $r_{\text{int}} \sim 10^{-8}$ to $10^{-7}\ \text{Km}^2\ \text{W}^{-1}$ [43, 44]. The upper bound for the measured interface resistance yields a somewhat unrealistically high value of 0.97 for the backscattering probability. However, it must be borne in mind that it is difficult to obtain good contact uniformity in a large-scale experiment, and therefore the experimental values must be seen as upper bounds.

4.4.7
Combined Heat Transport Through Tip, Interface and Sample

We define the *heating efficiency c* (similar to – but a simplification of Equation 4.15 and Figure 4.11) as the increase in the sample surface temperature divided by the total temperature difference between heater and substrate:

$$c = \frac{R_{\text{sp}}}{R_{\text{tip}} + R_{\text{int}} + R_{\text{sp}}}. \tag{4.19}$$

Here, R_{tip} denotes the nonballistic, diffusive component of the tip resistance (the ballistic part is captured in R_{int}, as explained in Section 4.4.6). This definition is useful for understanding sensitivity issues when using heated probes. The heating efficiency c is a strong function of tip size – that is, of the lateral resolution. Small values of c imply that also the measured signal will be small, indicating that achieving high lateral resolution becomes increasingly difficult.

As outlined in Section 4.4.4 and inferred experimentally [8, 14, 35, 39, 40], typical values for c range from 0.3 to 0.7 for polymer samples. In the case of better thermal conductors, c can be much lower; for example, on metals we estimate $c \sim 10^{-3}$–10^{-4}, for semiconductors $c \sim 10^{-2}$–10^{-3}, and for oxides $c \sim 10^{-1}$. This points to the challenges expected when extending the SThM method to both the nanoscale (e.g. $a = 5$ nm in the above calculations) and to sample materials having a higher thermal conductivity than the commonly used polymers or oxides.

We note that, although the heating efficiency reflects the temperatures, it does not reflect how efficient a probe is in terms of *heating power*. For the cantilever type shown in Figure 4.3 operating in air, the ratio of power going through the tip to that lost to other heat paths is approximately $10^5/10^8$ (K/W) $= 0.001$. Thus, from a power consumption point of view, the delivery of heat to the sample through the tip is very inefficient. Improving the efficiency requires either a reduction in the interface and the tip thermal resistances or an increase in the air/cantilever thermal resistance. The tip and interface resistances can of course be decreased by using a blunter tip, but at the expense of lateral resolution. Increasing the cantilever thermal resistance requires a decrease in the heater and lead cross-sections and/or an increase in the cantilever length. Among the design issues that restrict the freedom to reduce the heater/lead cross-sections are the mechanical stability, cantilever stiffness, mechanical response time, power consumption of the heater, electrical resistance and noise, thermal response time and fabrication tolerances.

4.4.8
Heat-Transport Experiments Through a Tip–Surface Point Contact

The sheer number of heat-transport paths described above renders an understanding of heat transport in heated probes challenging, let alone the direct measurement of individual components of heat transport. In order to distinguish between different heat paths, we have chosen a threefold approach:

- Bringing the tip into and out of contact with the sample opens and closes heat channels.
- Operation in a vacuum removes the distance-dependent cooling path through the air, which tends to dominate the small change in cooling that occurs when the tip is brought into contact with the sample in ambient conditions and completely eliminates conduction through the water meniscus.
- By varying the contact area, a, only some of the contributions will be affected (interface and spreading part).

The total thermal resistance of the heater, R_{th}, is given by

$$R_{\mathrm{th}} = (T_{\mathrm{heater}} - \mathrm{RT})/P, \tag{4.20}$$

where T_{heater} is the temperature of the heater, RT is the room temperature, and P the heating power. The thermal resistance due to conduction through the cantilever legs

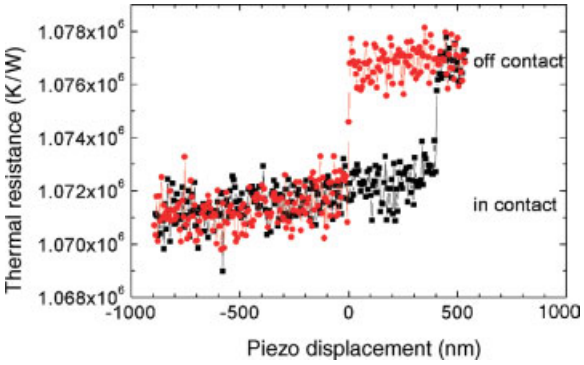


Figure 4.12 Thermal resistance of heated cantilever/tip as a function of displacement and contact with an 80 nm-thick SU8 film on a silicon substrate. (Reproduced from Ref. [14]; © Springer-Verlag.)

and radiation can be determined from the data obtained before the tip contacts the sample. The tip–surface thermal conductance can then be determined by subtracting this value from the data measured with the tip in contact with the sample. Figure 4.12 shows an example of such an experiment performed using a thermal probe similar to that shown in Figure 4.3 and a sample consisting of 80 nm of SU8 (an epoxy-based photoresist) on a silicon substrate. Out of contact, the displacement translates into a distance change between heater and sample. In contact, the displacement translates into a load force as the tip is pressed against the polymer. In this experiment, the average T_{heater} was approximately 315 °C, and the change in T_{heater} resulting from contact was about 1.5 K. From the difference between the thermal resistance out-of-contact ($\sim$1.077 MK W^{-1}) and the thermal resistance in-contact (1.071–1.073 MK W^{-1}), the thermal resistance due to heat transport through the tip–surface contact was calculated to be $2 - 3 \times 10^8$ MK W^{-1} (see Figure 4.14).

The tip–sample resistance is given by

$$R_{\text{ts}} = R_{\text{tip}} + R_{\text{int}} + R_{\text{sp}}, \tag{4.21}$$

as illustrated in Figure 4.13, where R_{tip} is the diffusive thermal resistance of tip, R_{int} is the tip–sample interface resistance, and R_{sp} is the spreading resistance in the polymer.

To experimentally quantify the different contributions to R_{ts}, we can vary the individual contributions by varying the sample material and the applied force. As the contributions depend on the contact radius, 0.5 d_0, it is useful to vary this parameter. To study the contact area dependence of the overall thermal resistance of the tip–polymer contact, we vary the force during an approach experiment. The contact area is calculated using the JKR model [46]. For this purpose, the applied force, the pull-off force and the tip radius need to be known. The applied force and pull-off force are determined from the cantilever spring constant and the known motion of the tip holder relative to the sample surface. The tip radius is measured *ex situ* by means of

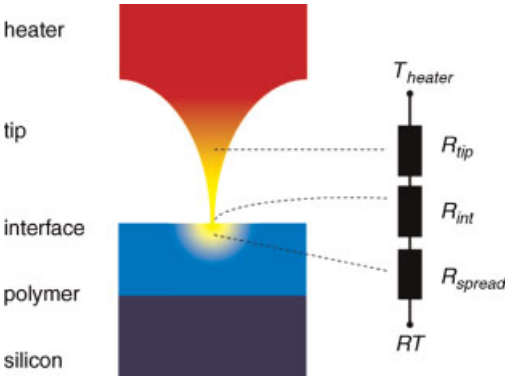


Figure 4.13 Schematic representation of the thermal resistances involved in heat transfer between the heater and a polymer sample. (Reproduced from Ref. [14]; © Springer-Verlag.)

scanning electron microscopy. The data shown in Figure 4.14 were obtained for a tip with $R_{tip} \cong 13.5$ nm corresponding to a variation of the contact diameter from 7 nm just before the contact breaks at a pull-off force of −15 nN to 13 nm at the maximum load of 35 nN.

Clearly, the thermal resistance depends on the tip force and hence on the contact diameter. Motivated by the above discussion, we propose the following ansatz for the thermal resistance:

$$R_{ts} = A_0 + A_1/d_0 + A_2/d_0^2. \tag{4.22}$$

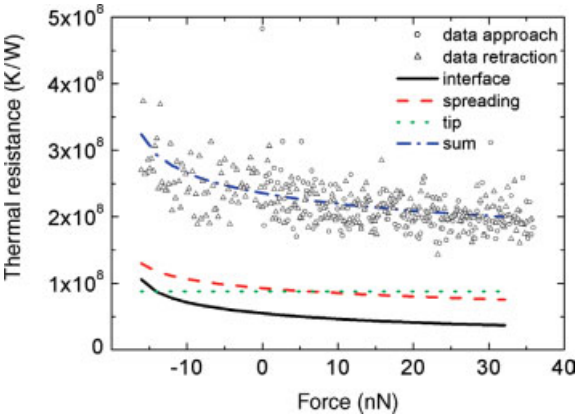


Figure 4.14 Experimental thermal resistance R_{th} of a typical tip–sample contact on a polymer (SU8) film and fit (blue, dash-dotted line) containing the three components R_{tip} or A_0 (green, dotted), R_{int} or A_2 (black, solid) and R_{sp} or A_1 (red, dashed). (Reproduced from Ref. [14]; © Springer-Verlag.)

Indeed, a good fit to the data is obtained with this second-order ansatz. In Figure 4.14, the individual contributions are shown as a green dotted line for the A_0 term, a red dashed line for the A_1/d_0 term, and a solid black line for the A_2/d_0^2 term.

Recalling the results from Sections NaN.2.4–NaN.2.6, it is surprising that there is a significant A_0 term ($\sim 9 \times 10^7$ MK W^{-1}). All components – tip, interface and spreading – should have an explicit dependence on d_0. We attribute the contribution A_0 to the thermal resistance of the tip R_{tip}. This may appear as a rather strong assumption, because we argued in Section 4.4.4 that the thermal resistance of the tip is predominated by ballistic conduction, and therefore the diffusive (nonballistic) thermal resistance of the tip should be vanishingly small in comparison. However, in practical applications the inner structure of the tip may play a role. In particular, the oxide cap covering silicon tips can contribute decisively to the thermal resistance of the tip. The value of this contribution can be estimated on its own as an independent thermal resistance by using approximate dimensions or as a mesoscopic link enabling 'phonon tunneling' between the silicon cone and the sample [47]. The uncertainty remains large, and we estimate 10^7–10^8 MK W^{-1} for the oxide cap. Accordingly, the total value of R_{tip} might be dominated by the value for the oxide cap, and we therefore also expect 10^7–10^8 MK W^{-1} for R_{tip}. This interpretation is supported by control experiments using the same tip on a silicon sample.

The A_2/d_0^2 term is an unequivocal sign for ballistic transport, and therefore it must be assigned to the interface resistance R_{int} (see Section 4.4.6). The A_1/d_0 term stems from the spreading resistance in the polymer sample. As discussed in Sections 4.4.4 and 4.4.6, it is composed of a real diffusive component and an interface contribution. It is difficult to assess the magnitude of the latter without detailed knowledge of the tip structure at the interface, which would allow one to make realistic assumptions on the scattering efficiency at the interface to the polymer. We note, however, that the magnitude is consistent with the diffusive spreading resistance for polymers (see Section 4.4.5). Hence, most likely the interface scattering contribution is rather small. It is interesting to note that all three terms contributing to the overall thermal resistance are of similar magnitude, namely, on the order of 10^8 MK W^{-1}, which results in a heating efficiency of $c \sim 40\%$ in this example (assuming that the A_1/d_0 term corresponds to a purely diffusive spreading resistance).

4.5 Thermomechanical Nanoindentation

Thermomechanical nanoindentation can be viewed as a powerful nanoscale extension to the existing methods of indentation (hardness testing) and dynamic thermomechanical analysis (modulus testing). The process involves pressing a heated tip into a sample using a defined tip temperature, load/heat duration, and load force. The indentation dynamics and the yield of the sample can be used to understand its material properties. Apart from the metrology discussed in this section, the technique has applications in data storage and in nanoscale patterning and lithography.

The indentation process yields considerable insight into the thermomechanical properties of materials – in particular of polymers – on the nanometer scale [48–52]. Traditionally, indentation processes are used to determine the hardness of a material [53], in which experiments an indenter produces a permanent deformation at the surface of the material under investigation. The hardness is determined by the size of the indentation with respect to the loading force. As the geometry of the indenter plays a fundamental role, common hardness definitions are based on individual, defined geometries, such as a ball (Brinell) or a pyramid (Vickers) [53]. In the experiments, typically a specific load is applied for specific time durations to yield comparable results.

Although it has not been widely used in traditional hardness testing, temperature is an additional important parameter in indentation experiments [14, 52]. As the mechanical properties of materials typically depend largely on temperature, the control of this parameter opens up interesting new areas of investigation. Heated probes provide an easy means to vary the temperatures on any given surface. In addition, controlling the probe temperature is relatively straightforward, and the low heat capacity of the probes allows the temperature of the probe to be switched at relatively high rates.

Polymers are one class of materials in which the mechanical and viscoelastic properties change dramatically with temperature, for example, at the polymer's glass transition temperature, T_g [1, 54]. At this temperature, the internal configurational changes within the polymer chain (which are linked to the translational motion of the chain) become slow compared to the typical experimental observation time scale of 1–100 s. As a result, the material drops out of equilibrium into a so-called 'glassy' state, and its materials properties change dramatically. The elastic modulus, for example, increases by orders of magnitude.

Heated probes are ideally suited to investigate the mechanical properties of polymers on a nanometer scale over a wide range of temperatures, from room temperature to several hundred degrees Celsius, and time scales varying over orders of magnitude down to the microsecond regime.

As an example, Figure 4.15 shows the result of an indentation experiment using heated probes. The indentations were written as a function of the load force, F, and the tip temperature, T, for indentation times of 10 µs using a tip having a radius of about 10 nm. The image shows the topography measured in contact mode after writing the indentations using the same tip. Whereas, in the lower left part of the image, no permanent indentations were formed, they appeared very clearly in the upper right part at high temperatures and forces. Clearly, in order to produce permanent indentations a certain minimum force and/or temperature must be applied. The corresponding characteristic line at the onset of indentation formation is called the *writing threshold* (see shaded region in Figure 4.15). For practical purposes, this dividing line can be defined as the load/temperature combination that leads to indentations of 1 nm depth: $T(F)_{d\,=\,1\,\mathrm{nm}}$.

Figure 4.16 exemplifies the threshold behavior of writing indentations with increasing temperature at a given load of 50 nN and a pulse duration of 10 µs. Whereas, below the threshold temperature T_h, no permanent indentations can be formed, above T_h their depth increases linearly with temperature.

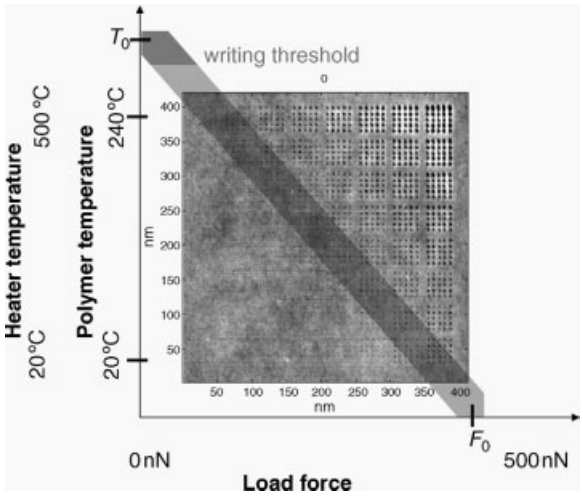


Figure 4.15 Atomic force microscopy image of indentations written into a polymer film at various combinations of load forces and tip temperatures using an indentation time of 10 µs. Blocks of 5 × 5 indentations spaced 36.6 nm apart are written using the same parameters. The blocks are written with increasing force and temperature along the *x*- and *y*-axis, respectively. The heater temperature is determined as described in the text. The polymer temperature under the tip is estimated using the results of the heat-transfer experiments described in Section 4.4.

Before turning to the physical interpretation of such temperature–load plots, let us briefly see how they can be used in practice. Of particular interest are the well-defined intersections of the writing threshold curves with the axes – that is, the writing temperature in the limit of no load force applied, T_0, and the load force in the limit of

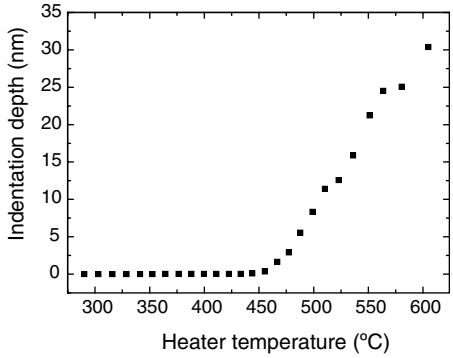


Figure 4.16 Depth of indentations in a polymer film (polymethylmethacrylate, PMMA) as a function of heater temperature for a load of 50 nN and heat- and force-pulse durations of 10 µs. Above a threshold heater temperature T_h of ~450 °C, the indentation depth increases linearly with the heater temperature. For the parameters applied, T_h is therefore closely related to the glass-transition temperature, T_g, of the polymer.

no heat applied, F_0. The quantities T_0 and F_0 are a function of both indentation time and tip geometry.

T_0 is the writing temperature needed in the limit $F \rightarrow 0$, which is also the limit of zero stress. In this limit the T_g is defined and therefore, for a given tip radius and indentation time, T_0 is a measure of the T_g.

For a given heater temperature, the temperature reached in the polymer underneath the tip depends on the geometry of the tip (see Section 4.4). Whereas, the opening angle of the tip cone has relatively little effect, the contact area – and therefore the heat transport from the tip to the polymer – differ for blunt and for sharp tips.

To reach more quantitative statements about the T_g of a polymer, we must normalize the effects of the tip geometry. There are two possible solutions to this:

- First, to obtain comparable results, one can use the same tip for all samples being studied. Assuming that the tip shape stays constant over the range of experiments, this procedure yields comparable results that can be correlated to traditionally measured T_g values of polymers.
- The second approach is to pick one of the polymers as a reference sample and to normalize the results on the other polymers with respect to this reference polymer.

Clearly, the first approach is prone to difficulties relating to the necessary constant geometry of the tip, is restricted to the use of a single tip, and therefore cannot be applied as a general method. In the second method, the T_g values measured are rescaled with respect to a reference tip on the reference sample, which cancels to first order the relative difference in heat-transfer properties resulting from the use of different tips.

By using these two approaches, a correlation to T_g measured by conventional means can be made [6, 14], as shown in Figure 4.17. All experimental data were either

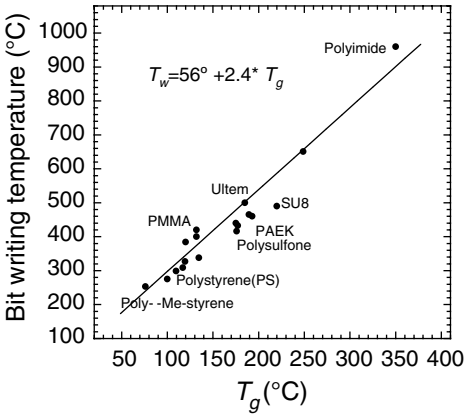


Figure 4.17 Indentation-writing temperature T_0 as a function of the conventionally determined T_g for various polymers. The indentation-writing temperature has been normalized to a reference tip, as described in the text. All data points are within 10% of the linear fit to the data.

obtained with a tip of approximately 10 nm radius or rescaled using reference samples [poly(methylmethacrylate) (PMMA) and polystyrene] to the case of a particular tip with specific opening angle and tip radius ($R_{tip} = 10$ nm), which yielded T_h at 400 °C for 10 µs-long heat pulses. The writing temperatures measured correlate very well with the respective T_g values of the polymers. This holds for uncrosslinked polymers, such as PMMA or Poly-a-Me-styrene, as well as for highly crosslinked polymers, such as the epoxy-resist SU8. In fact, the largest deviation of the data from the linear fit is less than 10%.

The excellent correlation of the indentation-writing and T_g values demonstrates the applicability of the method for determining T_g for unknown samples. This is insofar surprising as the T_g values for the given samples were determined using macroscopic bulk methods of dynamic mechanical analysis (DMA) or differential scanning calorimetry (DSC) and, in particular, also because these methods work on much longer time scales (1–100 s).

The second parameter obtained from the threshold curve, F_0, can be used to determine the hardness of the materials. As a demonstration, measured indentation-writing threshold curves are shown in Figure 4.18 for a class of similar polymers. The samples are thin films (120 nm thick) of polystyrene crosslinked using benzocyclobutene (BCB) as crosslinking groups [55]. This system is an ideal system to study the effects of the crosslink density on the thermomechanical properties of polymers. Increasing the crosslink density by incorporating a larger amount of BCB monomers increases both the hardness and T_g [56].

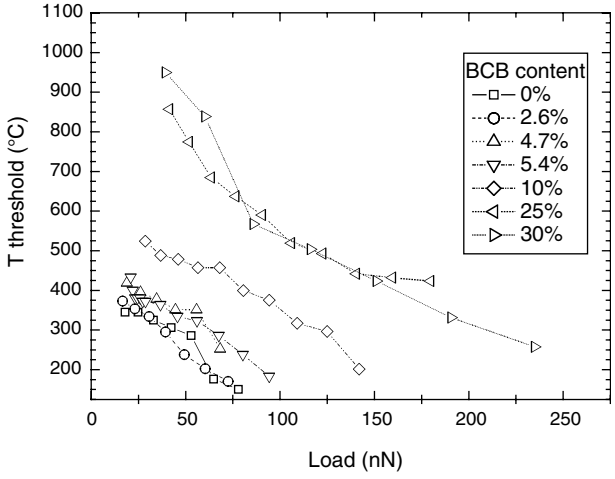


Figure 4.18 Indentation-writing threshold plots determined by writing indentation arrays, as shown in Figure 4.15. Each datum point refers to the temperature needed to write an indentation of 1 nm depth at a given load. Here, thins films of polystyrene that have different crosslink densities are compared. The percentage values in the inset refer to the relative amount of crosslinking benzocyclobutene (BCB) monomers with respect to styrene monomers in the polymer.

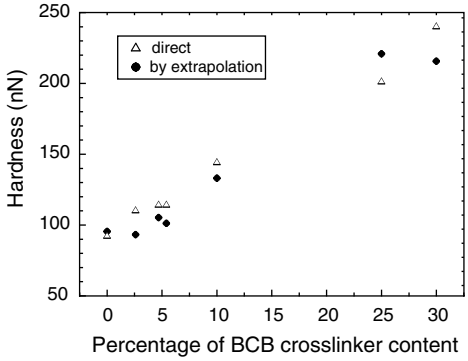


Figure 4.19 Comparison of two methods to determine sample hardness by nanoindentation. In the first method, the value (solid circles) is obtained by extrapolating the writing-threshold data from Figure 4.18 to room temperature. The second method (open triangles) measures the hardness in a more conventional manner: the minimum force required to obtain an indentation of 1 nm is determined, with tip and sample at room temperature. Within experimental uncertainty, the two methods yield identical results.

To a good approximation, the temperature needed to write an indentation decreases linearly with the writing load applied. This linearity can be used to extrapolate T_0 and F_0 from data taken in a limited force/temperature range. For example, in Figure 4.18 the writing-threshold curves reach neither T_0 nor F_0. Nevertheless, both values may be determined by extrapolation to zero force and room temperature, respectively. To verify the validity of this extrapolation, direct measurements of F_0 were performed at room temperature using the same tip. A comparison of the extrapolated and the directly measured hardness values for the polymers is shown in Figure 4.19.

A well-known issue in hardness testing is the wear of the indenter. This poses a problem in particular in the case of nanoindentation because, in order to draw valid conclusions, the indenter geometry must be known. Typically, this issue is circumvented by choosing indenters having relatively blunt apexes and/or using diamond materials. For true nanoscale applications using indentations that are confined in both the normal and the lateral direction, the wear of the probe is an unsolved issue.

The writing-threshold experiment as depicted in Figure 4.15 can be a workable solution, because the forces acting on the tip can be minimized. A grid with limited writing forces and subsequent extrapolation to room temperature can be applied to minimize tip wear. Thus, measuring the temperature dependence of writing permanent indentations is an elegant way to measure the real hardness data by extrapolation.

The indentation experiments shown above demonstrate how sensitively thermomechanical nanoindentation depends on the load and temperature. Underlying, of course, are the material properties of the polymer, such as hardness and the T_g, as well as the tip geometry and heat-transfer properties of the cantilever/tip. All of these govern the indentation formation. As discussed above, the indentation-writing

threshold experiment results in a relationship of $T_{\mathrm{thresh}}(F)$ (or $F(T_{\mathrm{thresh}})$) that is linear within the uncertainties given.

One explanation of the existence of a defined threshold for the force needed to write a permanent indentation is to argue that the stress build-up in the polymer has to overcome a critical stress, the yield stress of the polymer. It has been found macroscopically that the yield stress σ_y [57, 58] of polymers is a function of temperature. More precisely, around T_g, σ_y varies linearly with temperature, with $\sigma_y \sim 0$ for $T = T_g$, which is consistent with our observation of the linear shape of the writing-threshold curves. Hence, we can write $F(T_{\mathrm{thresh}}) \propto \sigma_y(T)$. This model of yielding is also supported by the analysis of the indentation shapes as a function of indentation parameters [59] (T. Altebaeumer, unpublished results).

A model that explains the observed indentation behavior simply as a yielding phenomenon, however, is not fully satisfactorily. Yield implies a permanent deformation of the material. In polymers, such a permanent change is linked to a change in the topology of the material, which proceeds via chains sliding with respect to each other. In the case of yielding, this sliding is forced by the external stress, which has to overcome the inherent monomer-sliding friction in the polymer [60].

In macroscopic yield experiments, two types of yield behavior are generally observed, namely *shear yielding* and *crazing*. Shear yielding occurs in partially crystalline and tough polymers (such as polycarbonates) which can extend to a multiple of their initial lengths. Just above a critical yield stress, the polymers often form shear bands on a macroscopic scale. Crazing is observed in brittle polymers such as polystyrenes; these polymers can elongate by only a small percentage before they rupture, and therefore the stress–strain curve is only slightly bent just before fracture. At the same time, one often observes elongated voids in the material called 'crazes'.

Clearly, both macroscopic phenomena encounter difficulties at the nanometer scale. Both typically have a length scale much larger than the length scale of the nanoindentation experiments. Moreover, in many of the materials studied here such mechanisms should be largely constrained by the high crosslink density. Therefore, we note that the macroscopic definition of yielding must be applied with care on the nanometer scale.

In an alternative model, the material is not assumed to undergo yielding but rather a viscoelastic deformation (like rubber) in the heated state at a temperature above T_g. Elastic deformation in this regime still proceeds via the deformation of polymer chains, which implies a relative movement of polymer chains. In contrast to yielding, the chains in rubbery deformation are almost free to slide and are only held in place by entanglement or crosslink sites. The monomer relaxations in the polymer backbone, which couple to the translational motion of the polymer chains, are fast, and the friction between the monomers is reduced to very low values. Polymer motion is mainly limited by the chain-like nature and the network constraints of the material. In viscoelastic deformation, the external force is mainly needed to deform the polymer network because monomer friction is low.

After cooling to temperatures below T_g, monomer relaxations in the backbone become orders of magnitude slower and limit the translational motion of the polymer

chains: the indentation is 'frozen in' and the 'loaded rubber springs' are kinetically hindered from relaxing.

This picture of rubbery indentation is more compliant with our nanoscopic length scales because no macroscopic changes in the material are involved. In addition, as has been argued [14, 61], this picture of 'rubbery indentation' also captures some apparent physics better than the yield picture. For example, it works much better at ultrafast indentation times, which are possible experimentally. Moreover, even polymers with extremely high degrees of crosslinking can undergo a rubbery deformation if the deformation is small. On the other hand, rubbery indentation implies temperatures above T_g, in clear contradiction to findings of a linear threshold curve all the way down to F_0. In the threshold curve no apparent transition through the glass transition exists.

In nanoindentation experiments, T_g is not easy to quantify. It can be expected, however, that T_g increases for the high indentation rates typical of these experiments and decreases for the high stresses. Even for the lowest stresses that can be applied in the experiments, significant shear stresses of ~100 MPa will have to be considered. Although at compressive stresses, T_g always increases in macroscopic experiments, it is expected that under shear or tensile stress the underlying alpha-transition is eased. We note that a theory on yielding by Robertson [62] predicts WLF (Williams–Landel–Ferry) kinetics below T_g under shear stress, but this theory was found to be useful only near T_g [63]. All in all, T_g is difficult to predict for such experiments.

As mentioned above, an important aspect of both models is the difference in the indentation dynamics because in the yield picture monomer friction is predominant, whereas in the rubbery picture the chain/network topology of the polymer is the limiting factor.

In the rubbery picture, polymer backbone dynamics above T_g generally follow the so-called 'time–temperature superposition' [1, 64] and their kinetics are well described by the WLF equation:

$$T = \frac{\log(\tau_{\mathrm{ref}}/\tau)T_\infty - c_1 T_{\mathrm{ref}}}{\log(\tau_{\mathrm{ref}}/\tau) - c_1}.$$

Here, τ, $T$and k_B are the indentation time, indentation temperature of the polymer, and the Boltzmann constant, respectively. The WLF parameters τ_{ref}, T_{inf}, T_{ref} and c_1 are the fit parameters characteristic for individual polymers. Note that usually these parameters are found to be independent of the actual quantity measured, be it shear modulus, viscosity or heat capacity. We therefore expect rubbery indentation to be essentially controlled by backbone kinetics following WLF.

In the yielding model, again the indentation kinetics is controlled by the dynamics of the backbone and is essentially of Arrhenius-type with a single activation energy E_{a} [65]:

$$\frac{1}{\tau} = \frac{1}{\tau_0}\exp\frac{E_a}{k_B T}.$$

Thus, to distinguish between 'rubbery indentation' and yielding, the indentation kinetics should be investigated.

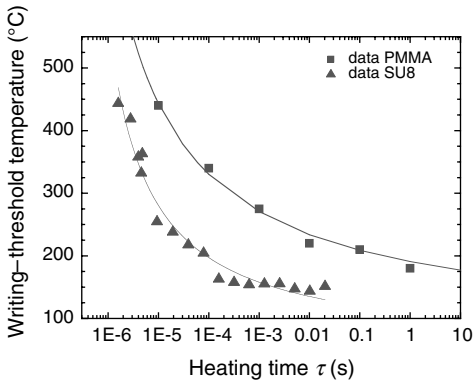


Figure 4.20 Writing-threshold heater temperature (i.e. the temperature required to write an indentation of 1 nm depth at a constant load force) as a function of heating time at a fixed load for a linear polymer PMMA (■) and a highly crosslinked epoxy SU8 (▲). The solid lines are fits using WLF kinetics (for which it was taken into account that the actual polymer temperature is significantly lower than the heater temperature).

First experiments [66] revealed that the indentation kinetics measured using PMMA and SU8 samples between 1 μs and 1 s cannot be fitted by a single activation energy (i.e. Arrhenius kinetics). An overall fit using WLF was satisfactory. An example of such an experiment is shown in Figure 4.20. Writing-threshold curves, defined as the minimum heater temperature needed to achieve an indentation depth of 1 nm at a given indentation time and at constant load force, were measured. Two prototype polymers are used: one is a thin film of PMMA; and the other a highly crosslinked thermoset, the epoxy SU8. A good fit with WLF can be obtained for PMMA. The temperatures needed are clearly above the T_g of about 120 °C, in agreement with our rubbery-indentation picture.

In a highly crosslinked system such as SU8, viscous flow can no longer account for the indentation. In this case, the load force was relatively high (200 nN), so that indentation times down to 1 μs were feasible with limited heater temperatures. Here also, a reasonable WLF fit was obtained despite the fact that, for long indentation times, the writing-threshold temperatures were considerably lower than the T_g of about 200 °C. Note, however, that for these data points the quality of the data does not allow the exclusion of a transition to Arrhenius-type behavior at long indentation times.

More detailed experiments using a crosslinked polystyrene sample were performed to investigate the topic further. These data are shown in Figure 4.21, in the form of an Arrhenius plot. Although the curve can be fitted linearly at long times, there is a clear deviation from the linear fit above a temperature of 180 °C that coincides with the T_g of the material measured using DSC. Above T_g, the data is well-fitted using WLF and a Vogel temperature T_∞ of $T_g - 50$ °C. Below T_g, however, the simple linear Arrhenius model is the best fit. Despite the many uncertainties, the

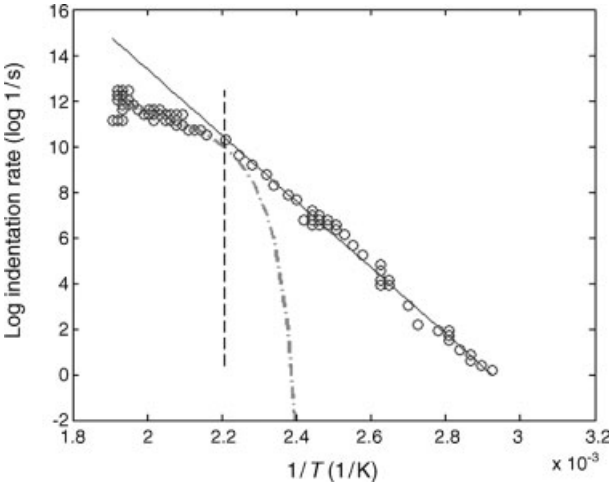


Figure 4.21 Arrhenius plot of thermomechanical indentation kinetics experiment. Below the glass transition temperature (T_g) of 180 °C (dashed line) an Arrhenius fit (solid line) is used. Above T_g, WLF yields a good fit (dash-dotted line). The sample used was a crosslinked polystyrene film; the indentation depth was 4 nm.

activation energy can be quantified and is found to be comparable with the activation energy of macroscopic polystyrene samples (1–2 eV).

It is concluded that WLF kinetics predominates at higher temperatures, and that a smooth transition to Arrhenius kinetics can be observed at lower temperatures/longer times. As expected, the transition occurs close to the T_g of the polymer.

These two physical pictures only manifest themselves in the different indentation dynamics. Only in the rubbery model above T_g does the chain-like nature of the polymers become apparent. On the other hand, it becomes clear that both physical pictures are useful to understand the experiments, and a distinction between them may appear artificial. The reason for this difference to the macroscopic polymer world is twofold: (i) because of the nanometer scale of the experiment and the crosslinked nature of the materials, macroscopic phenomena such as shear bands or crazes are absent; and (ii) the variations of shear stress and indentation rates are rather extreme. This forces a transition between the two conventionally fully separate regimes, which usually are switched only by the temperature and T_g.

In summary, a more unified picture of the indentation process emerges. For better clarity, we would like to propose a schematic, qualitative picture (see Figure 4.22). In this schematic, we capture the mechanics of the material using springs and dashpots. Springs k_s and k_p are connected in series and, correspondingly, in parallel to the dashpot. The elastic part of the medium is symbolized by k_s, whereas k_p is the elastic part linked to conformational changes of the polymer network. The dashpot, γ, is linked to the glass-to-rubber transition in the polymer. Below T_g, the dashpot is locked and can only be deformed by high stress; above T_g, it is open, representing the low-friction sliding of the polymer chains.

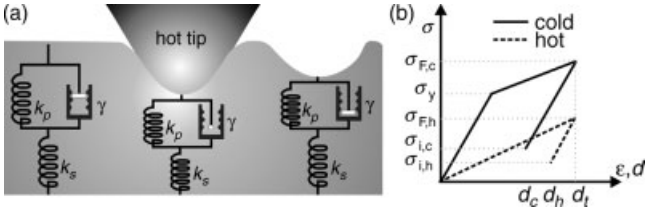


Figure 4.22 (a) Model representation of the polymer and the indentation process. From left to right: Undeformed polymer, polymer heated and deformed by the tip, and relaxed cold indentation; (b) Schematic of the stress–strain curves during an indentation experiment with hot and cold tips. See text for details.

Figure 4.22a describes (from left to right) the events during an indentation using a hot tip. If the hot tip is in contact with the polymer sample, the polymer is above T_g, which means that the dashpot is open and essentially free to move. Upon application of external stress, both springs will be deformed according to their strength, as shown by the dashed line in the stress–strain diagram in Figure 4.22b. Let us assume that the total deformation is d_t. Upon cooling, we lock the dashpot in the deformed state, and by releasing the external stress, $\sigma_{F,h}$, spring k_s relaxes to its uncompressed length. This state is show in the center of Figure 4.22a. We observe a partial loss of the indentation depth (to d_c for the cold case) as we retract the tip which results from elastic recovery in the material.

However, the $k_p - \gamma$ system is still deformed, and a residual stress given by the deformation of the polymer network spring k_p is locked in the deformation. In fact, this residual stress can be used to erase the indention (as will be shown in Section 4.6), and we also found stress-dependent relaxation in retention studies of the indentations. This deformation above T_g implies that the dynamics follows WLF kinetics because the backbone relaxation dynamics also follows this law. And indeed, we did observe this behavior for hot tips, as shown above.

If we deform the polymer below T_g, the situation is rather different, however. Under a cold tip, the dashpot is initially in the locked state. If we apply stress, the entire deformation at low stress values will first be absorbed only by the elastic spring k_s. The dashpot only opens once a critical stress, the yield stress σ_y, has been attained, as indicated by the solid line in Figure 4.22b. At the yield stress, the backbone motion is forced by the external stress and we have reached the writing threshold. As the dashpot opens, k_p is deformed accordingly, producing internal stress and, similarly to the hot case, we obtain elastic stress relaxation upon removal of the external force.

The state of the polymer after indentation is therefore remarkably similar in the two physical pictures discussed. Stress is stored in the deformed polymer network and is frozen in by the glassy state of the cold polymer. Only the amounts of elastic recovery (to d_c and d_h) and of the internally stored stress ($\sigma_{i,c}$ and $\sigma_{i,h}$) differ slightly. Experimentally, there is evidence of a higher remaining stress in cold indentations than in hot written ones because, at elevated temperature, the former relax faster (A. Knoll, unpublished results). The other significant manifestation of the two mechan-

isms is in the indentation dynamics, where we see a transient crossover from yielding to rubbery deformation with increasing temperature.

It is concluded that nanoindentation is a universal technique to study the deformation physics of polymers at the nanometer scale. By varying load, force, heat and temperature, important material properties – such as glass temperature, hardness, shift factors and yield-activation energies – can be extracted.

4.6 Application in Data Storage: The 'Millipede' Project

The capability of scanning probe techniques to modify and image a surface on the nanometer scale makes these techniques obvious candidates for data-storage applications. In fact, since the invention of the STM, many demonstrations of bit formation and imaging have been reported in the literature using almost every SPM technique and many different storage media and write mechanisms. Perhaps the most impressive of these demonstrations – at least from a density point of view – is the manipulation of individual atoms on a surface [67]. Although the storage densities that could be achieved with such techniques are very impressive, the construction of an actual storage system based on one of these ideas requires that numerous issues be addressed, including automated bit detection, system data rate, error correction, bit retention, power consumption, eraseability/cyclability, servo/tracking, reliability and cost. Many of these requirements are actually in competition with each other. For example, the highest storage densities demonstrated so far – that is, atomic scale – were achieved with very slow read-back speeds and had rather complex system requirements, such as ultra-high-vacuum conditions and low temperatures.

One scanning probe storage technology that achieves a balance between the many competing system requirements is the thermomechanical approach developed by IBM and referred to internally as the 'millipede' project [6, 7, 68]. In order to achieve a data rate comparable to those of conventional storage technologies, IBM has used microelectromechanical systems (MEMS) technology to fabricate large arrays of cantilevers that can be operated in parallel, with each cantilever writing and reading data in its own small storage field. The internal name of the project – 'millipede' – refers to the approximate 1000 cantilevers that were used in one of the first prototype systems.

4.6.1 Writing

The *write mechanism* used is the thermomechanical nanoindentation of polymers, as described in Section 4.5. The basic write process is illustrated in Figure 4.23. Data are written by pulsing the voltages applied to the cantilever to obtain suitable heat and force pulses while the tip is being scanned over the surface. Indentations placed at predefined positions along the data track can be used to encode data, with for

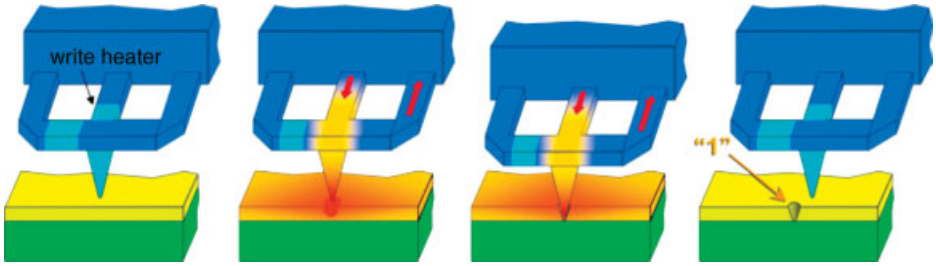


Figure 4.23 The principle of thermomechanical writing. The tip is heated by applying a current pulse to a resistive heater integrated in the silicon cantilever, directly behind the tip.

example, an indentation representing a logical '1' and the absence of an indentation a logical '0'. Storage densities greater than 1 Tb in^{-2} have been demonstrated using this scheme in combination with appropriate coding [69].

4.6.2 Reading

The data are read back by measuring the topography of the polymer surface using same tip that wrote the data. In the IBM approach, this is done using a read-back mechanism based on heat-transport sensing. For this purpose, a second heater has been integrated into the cantilever structure. This second heater is remote from the tip, and can be heated without causing much of an impact on the tip temperature. When operated in ambient air conditions, the thermal resistance of the read heater exhibits a strong dependence on the distance between heater and medium surface, as discussed in Section 4.4.1. This thermal resistance dependence results in turn in a heater temperature dependence and hence also an electrical resistance dependence on the distance between heater and medium surface. (The electrical resistance change with temperature is an intrinsic property of silicon, as discussed in Section 4.2.) This situation can be exploited to sense the topography by applying a constant voltage to the heater and monitoring the changes in the electrical resistance that result as the tip is being scanned over the surface. For example, when the tip moves into an indentation (a '1'), the distance between cantilever and surface is reduced and the heat-transfer rate increased. This leads to a resistance change of the $\sim$1 kΩ heater of $\Delta R/R \sim 10^{-4}$ per nanometer (Figure 4.24).

4.6.3 Erasing

Erasing [6] is achieved by exploiting the mechanical stress that is stored in an indentation. The thermomechanical writing process described above results in indentations that are a metastable, deformed state of the polymer with a significant amount of stored elastic energy. If a sufficiently hot tip is pressed against the surface

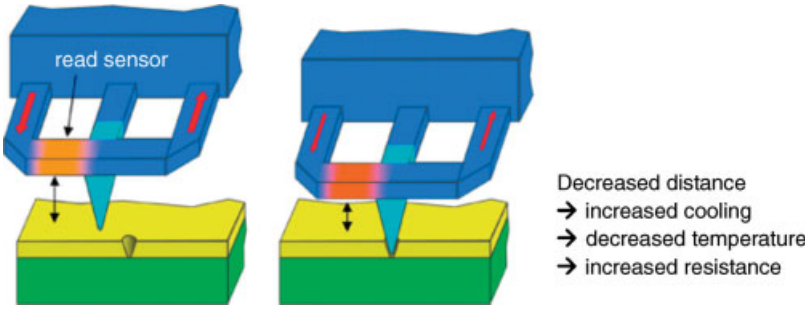


Figure 4.24 The principle of thermomechanical reading. Heat is generated by applying a current to a resistive heater integrated into the silicon cantilever. The heat transfer between heater and medium surface varies as a function of the distance between the cantilever and surface. Decreasing the distance between tip and medium leads to an increase in the cooling, which in turn decreases the temperature and increases the resistance, producing a detectable signal.

in the close vicinity of an indentation – that is, in the region of the rim around the indentation – then the increase in temperature in the indentation will result in a decrease in the viscosity of the polymer, which allows the elastic stress in the indentation to relax, effectively erasing the indentation. This process usually results in the creation of a new indentation, which can then be erased by repeating the procedure. Thus, a previously written data track can be erased by overwriting the data track with a series of closely spaced indentations. With this procedure, each new indentation erases the preceding one such that, at the end of the data track, all indentations will have been erased except for the last one. A demonstration of this principle is shown in Figure 4.25.

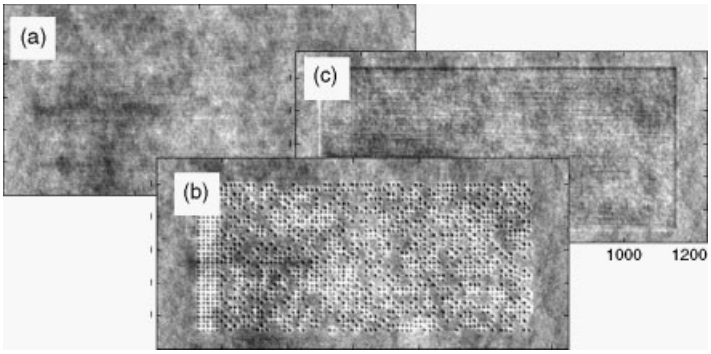


Figure 4.25 Atomic force microscopy topographical images illustrating the principle of thermomechanical writing and erasing. The images show (a) an empty area; (b) an area with indentations written at 1 Tb in^{-2}; and (c) an erased area. The grayscale covers 5 nm in all three images.

4.6.4
Medium Endurance

Medium endurance is a critical issue. The first challenge for polymer media is to be robust against repeated scanning with a sharp tip. In general, polymers tend to quickly roughen (and form ripples) when they are scanned repeatedly with a sharp tip, even at low load forces (see Section 4.6.5). Of the many solutions proposed to overcome medium wear, only a few can be readily applied to the nanoscale. On the nanometer scale, the homogeneity of the medium is crucial for nanoscale data-storage applications, and thus phase separation, filler particles or similar ideas cannot be used.

One elegant way to solve the issue of medium wear is to introduce a high degree of crosslinking into the polymer. This not only solves the roughening issue during sliding (reading) [56] but also facilitates erasing, because it provides the medium with a means of storing elastic energy and results in a type of 'shape memory'. To date, more than 10^4 write/erase cycles have been demonstrated using highly crosslinked polymer media (H. Podzidis *et al.*, unpublished results).

The dramatic improvement in wear endurance that occurs with increasing crosslinking is demonstrated in Figure 4.26, where the wear rate is plotted as a function of crosslink density for a set of polystyrene samples that were repeatedly scanned with a sharp tip. At a critical value of crosslinking, the mobility of the polymer is significantly reduced. This occurs when there is a sufficient number of crosslinks so that each region of cooperative polymer motion (typically ~1–3 nm in size) is affected.

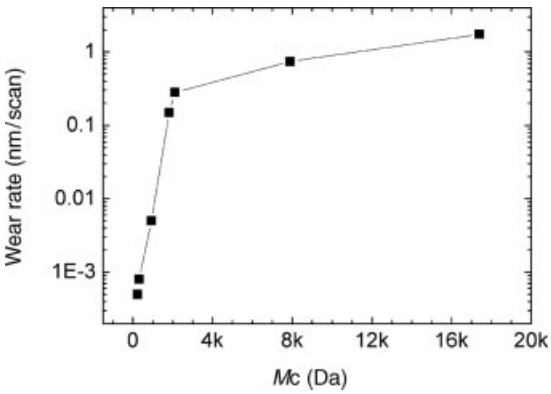


Figure 4.26 The peak amplitude of roughening induced by the wear of crosslinked polystyrene using a sharp tip scanning the surface, as in the reading operation. Here, the wear rate is defined as the cumulative amplitude of surface-topography changes normalized by the number of repeated scans of the same area. The wear rate is a strong function of crosslinkage. The crosslink density is given in units of molecular weight between crosslinks in the polymer chain. (Reprinted from Ref. [56]; © 2006, American Chemical Society.)

4.6.5
Bit Retention

Bit retention and the long-term stability of written data are also governed by the polymer mobility below the *Tg* value. This mobility is fundamentally provided by the activation energy of a backbone motion – that is, the so-called *alpha-relaxation*. Depending on the polymer, this can be as be as much as several electron volts, and can thus be sufficiently high for typical lifetime requirements. Lifetimes of 10 or more years at operating temperatures of up to $\sim$80 °C have been extrapolated from experimental data.

4.6.6
Tip Endurance

Tip endurance may limit the feasibility of several SPM-based data-storage schemes that involve mechanical contact between probe and surface. The endurance requirements of a tip will, of course vary, depending on the application and the system architecture. In general, however, a single tip will have to scan distances ranging from 10^4 to 10^8 m during the lifetime of the device, without losing its ability to read and write data. In thermomechanical data storage, the polymer medium is relatively soft compared to the hard silicon tips used for reading and writing. However, even for this combination, tip wear still is an important issue, and other – even harder – tip materials are also currently being investigated. Lubrication has proved to be key in improving the endurance of hard-disk drives, and may also prove to be useful for probe-based storage.

The density limits of thermomechanical data storage are predicted to be well above the 1 Tb in^{-2} mark. Ultimately, the density limits will be determined by the mobility of the polymer that corresponds to finite regions in which cooperative motions of polymer chains or chain segments occur. These regions range in size from 1 to 3 nm. As a small number of such regions must occur in each indented zone, a limit will appear somewhere at or below an indentation spacing of 10 nm.

4.6.7
Data Rate

The data rate is commonly one of the weaker aspects of SPM-related data-storage schemes. In the thermomechanical approach, two factors contribute to data-rate limitations:

- The cantilevered tip must be able to follow the topography mechanically; this translates into the requirement of a high mechanical resonant frequency.
- Temperature-based displacement sensor must be able to respond to these topography-induced height changes, ideally with a low power consumption.

The situation is further complicated by the requirement for low applied forces during the read operation in order to minimize tip and polymer wear. This low-force

requirement in turn entails the need for a small spring constant, which tends to reduce the resonance frequency. Finally, during the write operation, the cantilever must also be able to apply and withstand forces on the order of hundreds of nanoNewtons. Thus, in order to achieve a competitive thermomechanical storage technology, all of these competing requirements must be carefully balanced and the cantilever design highly optimized. However, even with optimization, a data rate per cantilever/tip well above 1 MHz appears speculative. Consequently, a high degree of *parallelization* of 10^2 to 10^4 tips operating in parallel is required to achieve a sufficient user data rate, and this is feasible only if the fabrication employs VLSI silicon technology. To date, the fabrication of prototype cantilever arrays with thousands of tips has been demonstrated, as illustrated in Figure 4.27. Moreover, parallel read/write operation at high densities using a small subset of cantilevers has been achieved [70]. Currently, three electrical connections to the array chip are required for each cantilever that is to be operated, and thus the number of cantilevers that can be operated in parallel is limited by the area available for bonding wires. The demonstration of higher degrees of parallelization will require the integration of some of the system electronics behind the cantilevers, and this is an area of current research. The other basic components required to make an actual prototype storage system based on this technology, including a MEMS scanner, a position-sensing and servo-control scheme, a bit-detection scheme and error-correction codes as well as a

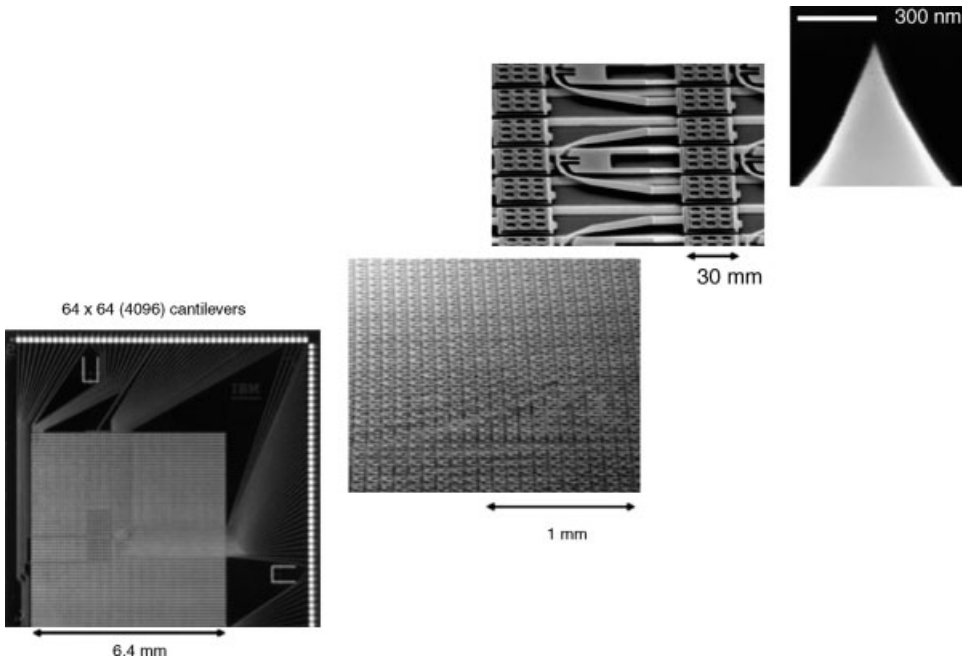


Figure 4.27 Microfabricated 64 × 64 cantilever/tip array for thermomechanical data storage.

system controller, have also been developed. All of this makes the route to highly parallel SPM-based storage appear feasible.

4.7 Nanotribology and Nanolithography Applications

Applications of heated probes going beyond thermal imaging and thermomechanical indentation can more generally be described as exposing surfaces to heated tips. In this section, examples are presented that lead to the modification and patterning of surfaces. First, experiments are described that involve scanning with a hot tip on polymer samples, with the aim of understanding nanoscale wear. Second, the *controlled* removal of material with the application to scanning-probe lithography (SPL) is developed and analyzed. Finally, dip-pen nanolithography using heated tips will be discussed.

4.7.1 Nanowear Testing

Nanowear testing using AFM is commonly performed to understand the nanoscale wear of various materials. Wear in general is a complex phenomenon, and often very different physical mechanisms come simultaneously into play, such as thermally activated bond rupture, adhesion, frictional shear stress, third-body lubrication, and so on [71]. Thus, in wear experiments, certain parameters are varied to elucidate the wear mechanisms. For example the repetition of wear cycles, the load force on the scanning tip, the scanning speed, or environmental conditions, such as humidity, are examined. Temperature is less suitable as a variable in such experiments, as mentioned above in the context of thermal imaging. However, temperature is a decisive parameter controlling wear, in particular of polymers. Let us, for example, consider the wear of polymer surfaces around T_g. The glass-transition region is often very sharp and covers only few degrees; for thin films T_g may vary by several degrees because of finite-size effects. A temperature variation of a sample around T_g therefore involves ramping the temperature in relatively small steps, with each step including settling for minutes, or even hours, to equilibrate the sample. In contrast, the *tip temperature* is easily varied and settles within microseconds. Therefore, heating the tip rather than the sample makes it possible to perform wide temperature variations in a single experiment [8, 14, 72].

As an example, let us consider experiments in which a variably heated tip is raster-scanned over polymer surfaces. While the tip temperature is continuously increased along the slow-scan axis, low loading forces are maintained between the tip and polymer surface. A real-space image of thermal degradation (a 'wear track') is generated, in which each fast scan line corresponds to a certain tip temperature. A wear experiment of a polymer film of polystyrene (PS), a standard linear polymer, is shown in Figure 4.28a. Here, three regimes can easily be identified:

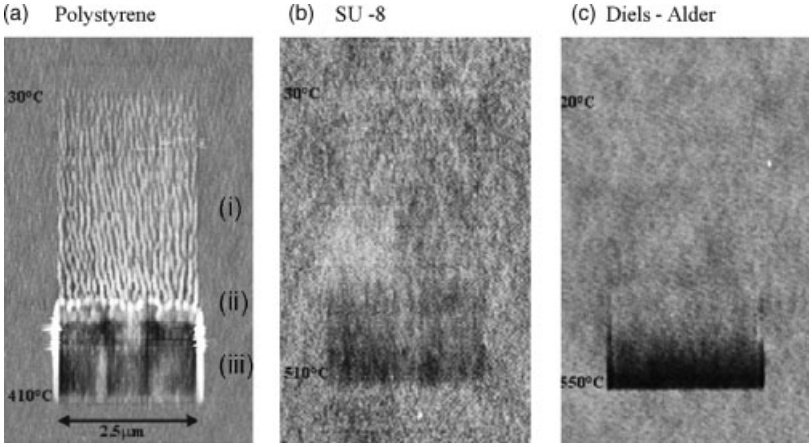


Figure 4.28 Wear tracks on different polymer samples obtained by raster-scanning a heated tip over a surface. The tip temperature is increased along the slow-scan direction (vertical). The wear track consists of 512 lines of 2.5 μm length scanned at 10 Hz. After the wear process, the images were obtained with an unheated tip in AFM imaging mode. (a) A polystyrene surface scanned four times; the grayscale covers an image corrugation of 17 nm. (Reprinted from Ref. [39]; (c) 2004, American Chemical Society.); (b) SU8 thermoset scanned 10 times; the grayscale covers a corrugation of 1.7 nm; (c) Diels–Alder polymer scanned 10 times; the grayscale covers a corrugation of 5 nm.

(i) At low temperatures, a *ripple pattern* is generated. The activated kinetics of ripple formation of PS has been studied using both variable sample temperatures [73] and heated tips [39]. Activation energies of the ripple process that exhibit a similarity with the yield process discussed above have been established.

(ii) A second regime is the glass-transition region – that is, at tip temperatures that lead to a heating of the polymer in the interaction region under the tip up to T_g. There rather drastic effects become apparent in a dramatic increase of the ripple amplitude.

(iii) In the third regime, above T_g, the material becomes so ductile that it is swept to the sides of the heated scan.

A very different trend is observed when the same experiment is performed with a thermoset material in which material transfer is more suppressed, for example in a highly crosslinked epoxy, such as a 100 nm-thick film of SU8 (see Figure 4.28b). Because of the high crosslink density of SU8, no ripple pattern can be formed. Overall, the surface remains unchanged by the wear test, and only a marginal depression is observed at the largest temperatures. If any debris is formed it is apparently so volatile that it cannot be traced on the surface with the probing tip after the wear procedure.

The third example (see Figure 4.28c) demonstrates yet another characteristic degradation mode, where a chemical reaction is induced by the heated tip. The

material is a highly crosslinked material in which the crosslinks are thermally reversible by virtue of a retro-Diels–Alder reaction [8]. The crosslinks are opened when a temperature of 130 °C is reached, and the material constituents are small molecular fragments that are rather volatile and may either diffuse onto the tip or evaporate. As observed in Figure 4.28c, no change in the surface is seen in areas scanned with heater temperatures of up to 320 °C. On further increasing the temperature, the depth of the wear track increases linearly up to 550 °C, although there is a clear lack of debris. In addressing the mechanism for track formation, compression is ruled out based on the observation that measured wear is cumulative. Repeating the experiment in a given area yields a proportional increase of the wear depth. Hence, it can be concluded that the material is lost by evaporation or diffusion onto the tip. The debris-less removal of material renders this method and the Diels–Alder polymer candidates for lithography applications. However, before we test this idea further let us briefly consider maskless lithography (ML2), in particular SPL.

4.7.2 Nanolithography Applications

In *microelectronics*, the time and expense required to produce a mask set is an issue in the prototyping of integrated circuitry [74]. For this reason, electron-beam lithography (EBL), which is an ML2 technique, is often used to prototype individual devices. The main drawback of ML2 efforts with respect to conventional lithography is the comparatively low exposure throughput. To remedy this, efforts are underway to develop EBL systems to operate a multitude of beams in parallel so as to reduce the overall exposure time [74]. However, further development and research is needed to control the crosstalk due to the high-voltage control signals and source brightness. The need for UHV conditions is also seen as an obstacle.

Currently, numerous SPL-based systems are under development – or have at least been proposed – and, in comparison to EBL tools, these will be more compact and simpler systems. The fabrication of large arrays of probes for massively parallel operation has been realized. One of the most powerful demonstrations of research-scale SPL used electrons extracted from a conducting tip to expose a resist [75]. Structures with resolutions of 30 nm have been transferred, and parallel operation has been demonstrated. The main challenges to this particular technique are the need for a conducting substrate, the high voltage necessary to extract electrons, the reliability of the electron-extraction process, and the lack of a simple overlay strategy.

Heated probes can also be used for SPL. For example, local heating has been used to induce the crosslink reaction of a conventional photoresist locally [76], after which the pattern is developed in the same way as in conventional lithography. Based on the example given in Figure 4.28c, an alternative SPL method can be applied that directly removes material from the exposed region. This approach offers specific advantages:

- It combines exposure and development in a single step, which not only makes the method simpler but also enables direct inspection of the exposure and direct repair

in a separate repair step. A prerequisite here is that the same probe can be used for imaging, which is possible because the exposure creates a surface topography that can easily be measured. Thereby, the difficult issue of reliability of SPL is directly tackled.

- It also facilitates the use of a simple overlay strategy based on exposure/development of alignment marks. These marks can be imaged and used to align the new layer to be exposed. Note that the same strategy could be used for stitching error correction or self-calibration of probe positioning.
- In contrast to conventional nanoindentation or nanoscratch-based methods [77], this variant helps to achieve high resolution because the material is not simply displaced but is actually removed from the exposure site [8]. This permits achievements similar to those with other SPL techniques.

A demonstration of a high-resolution exposure of such a material is given in Figure 4.29 [78]. The imaged area was inscribed using pixels spaced 3.0 nm apart in a line, with a spacing of 107 nm between the lines. Each pixel was written using a heat pulse of 20 μs duration and a loading force of 40 nN. The Diels–Alder polymers were shown to permit the production of repeatable lithographic features of 12–15 nm in size (full width at half maximum) by sequential inscription. This makes it possible to overlay line patterns and faithfully reproduce line-crossings, both features which are desirable in lithography.

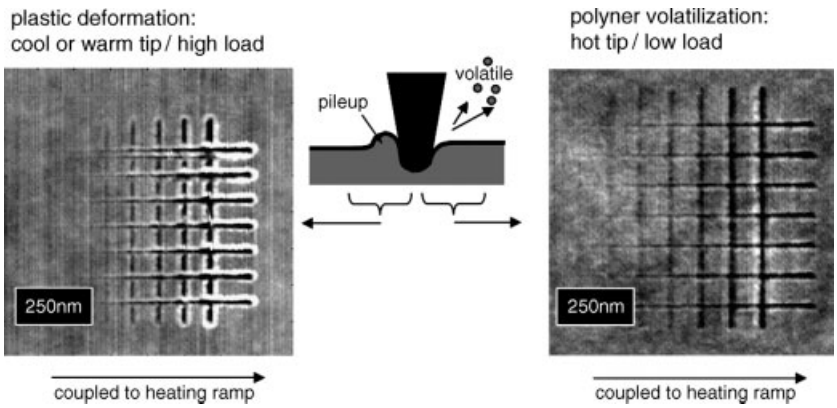


Figure 4.29 Comparison of lithography results using tips to (re)move material in a direct quasi-single exposure and development step. The material used is a reversibly crosslinked polymer described in Ref. [8]. Left panel: Conventional 'scratch'-type lithography using high forces (300 nN). The tip-heater temperature is varied from left to right, from room temperature to 440 °C. Independently of temperature, it is shown that the plastic deformation leads to a *displacement* of material to the sides of the individual lines written, making line-crossing impossible. Right panel: Similar experiment using a low load force (30 nN) and a heating ramp between 300 and 500 °C. In this case, material is *removed* and not just translated, and line-crossings are possible. A sketch of the difference between the two lithography processes is shown in the center panel.

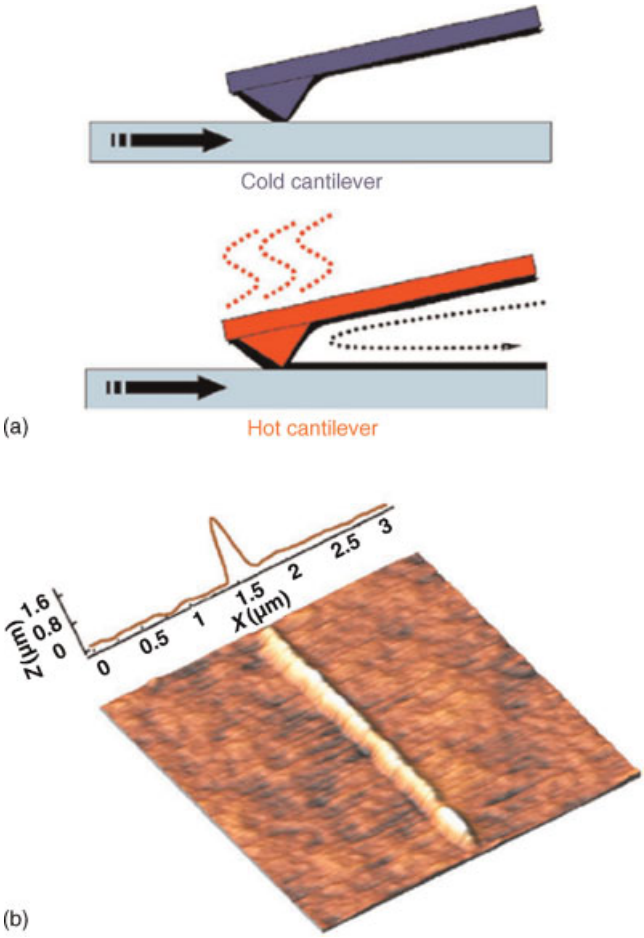


Figure 4.30 (a) Schematic of the operation of thermal dip-pen lithography (tDPN), which uses a heated AFM cantilever with a tip coated with a solid 'ink'. When the tip is hot enough to melt the ink, it flows onto the substrate. No deposition occurs when the tip is cold, thus enabling imaging without unintended deposition; (b) A topographic AFM image of a continuous nanostructure deposited from an indium-coated tip onto a borosilicate glass substrate. (Reprinted from Ref. [80]; (c) 2004, American Institute of Physics.)

The concept of removing material can be amplified by using materials that undergo an exothermic reaction when being volatized; in this regard, explosives have been used by King *et al.* [72]. The volatilization rate of a thin film of pentaerythritol tetranitrate exhibits a strong dependence on the tip temperature and exposure time. Although the study of King and coworkers emphasized the analytical application of heated probes, the nanostructuring context is evident.

For nanostructuring in a broader sense, it is interesting not only to *transfer* material along a surface (as in 'nanowear' experiments) or to *remove* material (as in the SPL examples above), but also to *deposit* materials. The deposition of material is particu-

larly appealing for biological applications. Although various methods exist to deposit material from liquid or gas phases using a scanned probe tip (such as local oxidation of semiconductor surfaces [79]), the direct deposition of liquid droplets or lines using dip-pen lithography (DPN) has recently attracted considerable attention. In this method, an AFM tip that has been covered with a liquid to be deposited is brought into contact with a surface at locations where the liquid is to be deposited. In numerous experiments performed over the past few years, a range of materials have been successfully deposited, notably those used for biopatterning applications. A major challenge of DPN is the controlled switching on and off of the deposition process. Since in many applications, arrays of probes must be run in parallel in order to achieve sufficient throughput, individual probes cannot easily be brought into and out of contact independently; otherwise, all probes would write the same pattern. One strategy to circumvent this problem relies on heated probes, where the temperature of the tip – the sidewalls of which are the 'ink' reservoir – can be used to turn the deposition on and off. This has been demonstrated by Sheehan *et al.* [80] (see Figure 4.30), where lines of ink (octadecylphosphonic acid) were written at linewidths down to 100 nm in a controlled manner.

Acknowledgments

The authors would like to thank C. Bolliger for carefully proof reading the manuscript of the chapter. They are also grateful to the 'Millipede' teams at the IBM Research Laboratories in Zurich and Almaden for their continued collaboration and helpful scientific discussions. Previously unpublished results were obtained in collaboration with J. Frommer, C. J. Hawker, J. Hedrick, M. Hinz and R. Pratt.

References

1 Ferry, J.D. (1980) *Viscoelastic Properties of Polymers*, 3rd edn, John Wiley & Sons, New York.

2 Cahill, D.G., Ford, W.K., Goodson, K.E., Mahan, G.D., Majumdar, A., Maris, H.J., Merlin, R. and Phillpot, S.R. (2003) *Journal of Applied Physics*, **93**, 793.

3 Chen, G. (2000) *International Journal of Thermal Sciences*, **39**, 471.

4 Chen, G., Borca-Tasciuc, D. and Yang, R.G. (2004) in *Encyclopedia of Nanoscience and Nanotechnology*, Vol. 7 (ed. H.S. Nalwa), American Scientific Publishers, p. 429.

5 Balandin, A.A. (2005) *Journal of Nanoscience and Nanotechnology*, **5**, 1015.

6 Vettiger, P., Cross, G., Despont, M. *et al.* (2002) *IEEE Transactions on Nanotechnology*, **1**, 39.

7 Eleftheriou, E., Antonakopoulos, T., Binnig, G. K. *et al.* (2003) *IEEE Transactionson Magnetics*, **39**, 938.

8 Gotsmann, B., Dürig, U., Frommer, J. and Hawker, C.J. (2006) *Advanced Functional Materials*, **16**, 1499.

9 Reading, M., Price, D.M., Grandy, D.B. *et al.* (2001) *Macromolecular Symposia*, **167**, 54–62.

10 Shi, L. and Majumdar, A. (2002) *Journal of Heat Transfer*, **124**, 329.

11 Majumdar, A., Lai, J., Chandrachood, M., Nakabeoppu, O., Wu, Y. and Shi, Z.

(1995) *Review of Scientific Instruments*, **66**, 3584.

12 Shi, L., Plyasunov, S., Bachthold, A., McEuen, P.L. and Majumdar, A. (2000) *Applied Physics Letters*, **77**, 4295.

13 Despont, M., Brugger, J., Drechsler, U., Dürig, U., Häberle, W., Lutwyche, M., Rothuizen, H., Stutz, R., Widmer, R., Rohrer, H., Binnig, G. and Vettiger, P. (2000) *Sensors and Actuators A*, **80**, 100.

14 Gotsmann, B. and Dürig, U. (2006) in *Applied Scanning Probe Methods IV: Industrial Applications* (eds B. Bhushan and H. Fuchs), Springer, Berlin, Heidelberg, p. 215.

15 Sze, S.M. (1981) *Physics of Semiconductor Devices*, 2nd edn, John Wiley & Sons, New York.

16 Dürig, U. (2005) *Journal of Applied Physics*, **98**, 044906.

17 Wiesmann, D. and Sebastian, A. in (2006) *Proceedings 19th IEEE International Conference on Micro Electro Mechanical Systems, 2006, Istanbul, Turkey, IEEE*, p. 182.

18 Reading, M., Houston, D.J., Song, M., Pollock, H.M. and Hammiche, A. (1998) *American Laboratory*, **30**, 13.

19 Price, D.M., Reading, M., Hammiche, A. and Pollock, H.M. (1999) *International Journal of Pharmaceutics*, **192**, 85.

20 Majumdar, A. (1999) *Annual Review of Materials Science*, **29**, 505.

21 Shi, L. and Majumdar, A. (2001) *Microscale Thermophysical Engineering*, **5**, 251.

22 (a) Pollock, H.M. and Hammiche, A. (2001) *Journal of Physics D - Applied Physics*, **34**, R23; (b) Shi, L. and Majumbdar, A. (2004) in *Applied Scanning Probe Methods I*, (eds B. Bushan, H. Fuchs and S. Hosaka), Springer, Berlin, Heidelberg, New York, p. 327.

23 Hinz, M., Marti, O., Gotsmann, B., Lantz, M.A. and Dürig, U. (2008) *Applied Physics. Letters*, **92**, 043122.

24 Polder, D. and Van Hove, M. (1971) *Physical Review B - Condensed Matter*, **4**, 3303.

25 Loomis, J.J. and Maris, H.J. (1994) *Physical Review B - Condensed Matter*, **50**, 18517.

26 Volokitin, A.I. and Persson, B.N.J. (2004) *Physical Review B - Condensed Matter*, **69**, 045417.

27 Pendry, J.B. (1999) *Journal of Physics: Condensed Matter*, **11**, 6621.

28 Mulet, J.-P., Joulain, K., Carminati, R. and Greffet, J.-J. (2002) *Microscale Thermophysical Engineering*, **6**, 209.

29 Hargreaves, C.M. (1973) *Philips Research Reports Supplements*, **5**, 1.

30 Xu, J.-B., Laeuger, K., Dransfeld, K. and Wilson, I. H. (1994) *Journal of Applied Physics*, **76**, 7209.

31 Mueller-Hirsch, W., Kraft, A., Hirsch, M.T., Parisi, J. and Kittel, A. (1999) *Journal of Vacuum Science & Technology A - Vacuum Surfaces and Films*, **17**, 1205.

32 DiMatteo, R.S., Greiff, P., Finberg, S.L., Young-Waithe, K.A., Choy, H.K.H., Masaki, M.M. and Fonstad, C.G. (2001) *Applied Physics Letters*, **79**, 26.

33 Chapuis, P.-O., Greffet, J.-J., Joulain, K. and Volz, S. (2006) *Nanotechnology*, **17**, 2978.

34 Volz, S. and Chen, G. (1999) *Applied Physics Letters*, **75**, 2056.

35 King, W.P. (2002) Thermomechanical Formation of Polymer Nanostructures, PhD Dissertation, Stanford University, CA, USA.

36 King, W.P., Santiago, J.G., Kenny, Th.W. and Goodson, K.E. (1999) *Proceedings, American Society of Mechanical Engineering, MEMS*, **1**, 583.

37 Dames, C., Dresselhaus, M.S. and Chen, G. (2004) Phonon thermal conductivity of superlattice nanowires for thermoelectric applications. Materials Research Society Proceedings, **793**, S1.2.1.

38 Jansen, A.G.M., van Gelder, A.P. and Wyder, P. (1980) *Journal of Physics C - Solid State Physics*, **13**, 6073.

39 Gotsmann, B. and Dürig, U. (2004) *Langmuir*, **20**, 1495.

40 Lantz, M., Gotsmann, B., Dürig, U.T., Vettiger, P., Nakayama, Y., Yoshikazu, S.,

Tetsuo, T. and Tokumoto, H. (2003) *Applied Physics Letters*, **83**, 1266.
41 Yovanovich, M.M., Culham, J.R. and Teerstra, P. (1998) *IEEE Transactions on Components, Packaging, and Manufacturing Technology, Part A*, **21**, 168.
42 Ross, R.G., Andersson, P., Sundqvist, B. and Backstrom, G. (1984) *Reports on Progress in Physics*, **47**, 1347.
43 Hu, C., Kiene, M. and Ho, P.S. (2001) *Applied Physics Letters*, **79**, 4121.
44 Govorkov, S., Ruderman, W., Horn, M.W., Goodman, R.B. and Rothschild, M. (1997) *Review of Scientific Instruments*, **68**, 3828.
45 Swartz, E.T. and Pohl, R.O. (1989) *Reviews of Modern Physics*, **61**, 605.
46 Cappella, B. and Dietler, G. (1999) *Surface Science Reports*, **34**, 1.
47 Patton, K.R. and Geller, M.R. (2001) *Physical Review B - Condensed Matter*, **64**, 155320.
48 Briscoe, B.J. (1998) *Journal of Physics D - Applied Physics*, **31**, 2395.
49 VanLandingham, M.R., Villarrubia, J.S., Guthrie, W.F. and Meyers, G.F. (2001) *Macromolecular Symposia*, **167**, 15.
50 Klapperich, C., Komvopoulos, K. and Pruitt, L. (2001) *Transactions of the American Society of Mechanical Engineering*, **123**, 624.
51 Fischer-Cripps, A.C. (2002) *Nanoindentation*, Springer, New York.
52 Hinz, M., Kleiner, A., Hild, S., Marti, O., Dürig, U., Gotsmann, B., Drechsler, U., Albrecht, T.R. and Vettiger, P. (2004) *European Polymer Journal*, **40**, 957.
53 Balta Calleja, F.J. and Fakirov, S. (2000) *Microhardness of Polymers*, Cambridge University Press, Cambridge.
54 van Krevelen, D.W. (1997) *Properties of Polymers*, 3rd edn, Elsevier, Amsterdam.
55 Harth, E., Van Horn, B., Germack, D.S., Gonzales, C.P., Miller, R.D. and Hawker, C.J. (2002) *Journal of the American Chemical Society*, **124**, 8653.
56 Gotsmann, B., Duerig, U.T., Frommer, J. and Hawker, C.J. (2006) *Nano Letters*, **6**, 296.
57 Kody, R.S. and Lesser, A.J. (1997) *Journal of Materials Science*, **32**, 5637.
58 Brooks, N.W.J., Duckett, R.A. and Ward, I.M. (1998) *Journal of Polymer Science Part B - Polymer Physics*, **36**, 2177.
59 Sills, S., Overney, R.M., Gotsmann, B. and Frommer, J. (2005) *Tribology Letters*, **19**, 9.
60 Strobl, G. (1996) *The Physics of Polymers*, 2nd edn, Springer, Berlin, Heidelberg, New York.
61 Binnig, G.K., Cherubini, G., Despont, M., Dürig, U., Eleftheriou, E., Pozidis, H. and Vettiger, P. (2007) *Springer Handbook of Nanotechnology, Part F Industrial Applications* 2nd edn, (ed. B. Bhushan), Springer-Verlag, Berlin, Heidelberg, New York, p. 1457.
62 Robertson, R.E. (1966) *Journal of Chemical Physics*, **44**, 3950.
63 Argon, A.S. and Bessonov, M.I. (1977) *Polymer Engineering and Science*, **17**, 174.
64 Williams, M.L., Landel, R.F. and Ferry, J.D. (1955) *Journal of the American Chemical Society*, 3701.
65 Ree, T. and Eyring, H. (1955) *Journal of Applied Physics*, **26**, 793.
66 Gotsmann, B., Dürig, U., Frommer, J. and Hawker, C.J. (2006) *Advanced Functional Materials*, **16**, 1499.
67 Eigler, D.M. and Schweizer, E.K. (1990) *Nature*, **344**, 524.
68 Pozidis, H., Häberle, W., Wiesmann, D., Drechsler, U., Despont, M., Albrecht, T.R. and Eleftheriou, E. (2004) *IEEE Transactions on Magnetics*, **40**, 2531.
69 Wiesmann, D., Dürig, U., Gotsmann, B., Knoll, A., Pozidis, H., Porro, F. and Vecchione, R. (2007) Innovative Mass Storage Technologies "IMST 2007", Enschede, The Netherlands.
70 Sebastian, A., Pantazi, A., Cherubini, G., Eleftheriou, E., Lantz, M. and Pozidis, H. (2005) Proceedings of the 2005 American Control Conf. "ACC 2005", Portland, OR, June 2005, IEEE, Vol. 6, p. 4181.
71 Bhushan, B.(ed.) (1995) *Handbook of Micro/Nano Tribology* CRC Press, London.

72 King, W.P., Saxena, S., Nelson, B.A., Weeks, B.L. and Pitchimani, R. (2006) *Nano Letters*, **6**, 2145.

73 (a)Schmidt, H.R., Haugstad, G. and Gladfelter, W.L. (2003) *Langmuir*, **19**, 10390; (b) Wang, X.P., Loy, M.M.T. and Xiao, X. (2002) *Nanotechnology*, **13**, 478.

74 Groves, T.R., Pickard, D., Rafferty, B., Crosland, N., Adam, D. and Schubert, G. (2002) *Microelectronic Engineering*, **61–62**, 285.

75 Wilder, K., Quate, C.F., Singh, B. and Kyser, D.F. (1998) *Journal of Vacuum Science and Technology B*, **16**, 3864.

76 Hung, M.-T., Kim, J. and Sungtaek Ju, Y. (2006) *Applied Physics Letters*, **88**, 123110.

77 Kunze, U. and Klehn, B. (1999) *Advanced Materials*, **11**, 1473.

78 Gotsmann, B., Dürig, U., Frommer, J. and Hawker, C.J. (2006) *Advanced Functional Materials*, **16**, 1499.

79 Tello, M., Garcia, F. and Garcia, R. (2006) in *Applied Scanning Probe Methods IV: Industrial Applications* (eds B. Bhushan and H. Fuchs), Springer, Berlin, Heidelberg, p. 215.

80 Sheehana, P.E., Whitman, L.J., King, W.P. and Nelson, B.A. (2004) *Applied Physics Letters*, **85**, 1589.

5
Materials Integration by Dip-Pen Nanolithography

Steven Lenhert, Harald Fuchs, and Chad A. Mirkin

5.1
Introduction

The concept of using a tip coated with an ink – that is, a pen – to write on a surface has been used throughout history and is widely used today for recording or communicating information by hand. Although the most ancient written texts appear to have been carved in surfaces using sharp tools such as a knife or chisel, there are several reasons why the pen has eventually become the hand-writing tool of choice. First, the constructive nature of the pen typically enables a higher contrast than carving, making it possible to distinguish the writing from the surface background without further processing steps. Second, pen writing is relatively independent of the contact force in comparison with carving. And finally, if desired, a variety of different inks can be readily integrated on the same surface.

The same conceptual advantages that make the pen a useful tool on the macroscale also translate to the nanoscale when the tip of an atomic force microscope is used as an ultra-sharp pen to transfer material to a surface with nanometer scale resolution, a method known as dip-pen nanolithography (DPN) [1]. By using this technique, high-resolution chemical patterns can be constructed on surfaces in a single deposition step. Because the ink-transfer is independent of the contact force between the atomic force microscope tip and the substrate in almost all known cases, it is possible to carry out DPN reproducibly and in parallel, without the requirement for feedback from individual tips. By coating different tips with different inks, it then becomes possible to integrate a wide variety of molecules on a surface. As with other scanning probe lithography (SPL) methods (e.g. mechanical modifications, oxidation, local thermal treatments), DPN offers ultra-high lateral resolution, well below 20 nm. As a direct write lithographic method, DPN enables arbitrary patterns to be drawn without the need for a mask, with capabilities comparable to those of electron-beam lithography (EBL). Additionally, it is a tool that is ideally situated to rapidly produce laboratory prototypes and structures that are incompatible with the harsh conditions associated with conventional microfabrication techniques (soft biological structures in particular).

Nanotechnology. Volume 6: Nanoprobes. Edited by Harald Fuchs

ISBN: 978-3-527-31733-2

Importantly, DPN makes it possible to integrate different materials on scales (both in size and complexity) that appear impossible to reach by any other direct-write fabrication method. Such a method is clearly desirable for the fabrication of biomolecular arrays, and opens entirely new possibilities in the study and development of nanotechnology. This chapter will introduce the fundamental concepts in DPN technology, with a focus on aspects which enable nanoscale materials integration. In order to gain an understanding of what to expect, theoretical models will be introduced, followed by experimental approaches to controlling ink transport of various ink–substrate combinations, tip-coating methods, driving forces and characterization methods. Examples of unique applications of materials integration by DPN will then be described that cannot be achieved by any other method. Excellent reviews have been produced by Mirkin and others that summarize the DPN literature, and the reader is referred to those works for a more complete description of the vast amount of work already published on DPN to date [2–4].

5.2 Ink Transport

DPN is made possible by the transport of a material (ink) from the tip of an atomic force microscope to a surface at point where the tip contacts the surface (Figure 5.1a). As in the case of a macroscopic pen, the ink must flow from the tip of the pen to the paper, and this transport process is typically driven by an interaction between the ink and the substrate. However, quantitative differences appear when this technique is carried out at the nanoscale tip of an atomic force microscope. Most striking is that DPN is able to produce patterns consisting of a single molecular layer. The most

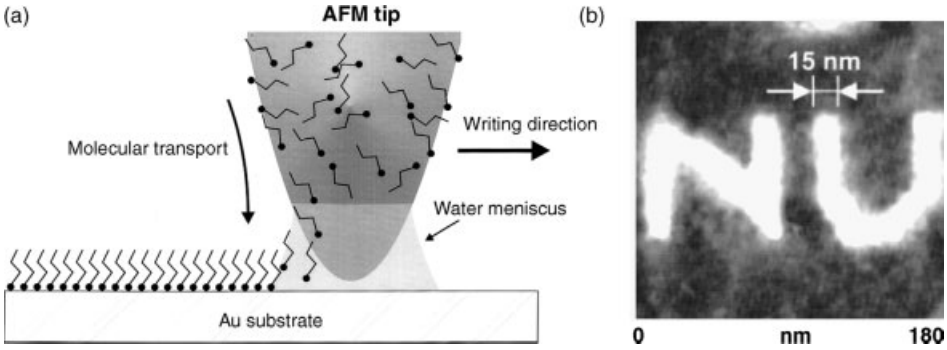


Figure 5.1 (a) Schematic illustration of the concept of dip-pen lithography DPN. (Reprinted from Ref. [1], with permission from the American Association for the Advancement of Science.); (b) Atomic force microscopy (AFM) friction image of patterns of the thiol mercaptohexadecanoic acid patterned on gold ⟨111⟩ with 15 nm line widths. (Reprinted from Ref. [5], with permission from the American Association for the Advancement of Science.)

thoroughly studied (and widely reproduced) ink–substrate combination for DPN is the patterning of alkanethiols on gold surfaces. Alkanethiols spontaneously self-assemble on gold surfaces under the appropriate conditions to form tightly packed self-assembled monolayers (SAMs). Typical thiol inks include octadecanethiol (ODT), which forms a methyl-terminated SAM, and mercaptohexadecanoic acid (MHA), which forms a carboxylic acid-terminated SAM. Figure 5.1b shows an example of a high-resolution DPN pattern of MHA on a single crystalline gold surface, with line widths of 15 nm. As the radius of curvature of the tip used to make that pattern was approximately 10 nm, it has been hypothesized that a sharper tip may enable the fabrication of even smaller features [5]. The ultimate resolution limit of DPN has yet to be determined. It has even been proposed that DPN can be used to generate features consisting of single molecules, and that the practical limit lies in detecting such small features by atomic force microscopy (AFM) [6].

In a typical experiment aimed at characterizing the transport rate of an alkanethiol such as MHA or ODT from the atomic force microscope tip to a gold surface, the coated tip is placed in contact with different areas of the surface for different amounts of time. These contact areas can then be imaged *in situ* by rapidly scanning the patterned area with the same tip in lateral force mode to obtain a friction contrast image such as that shown in Figure 5.2a. Upon plotting the area (or radius, r) of the spots as a function of contact time, it is possible to quantify ink transport rate. It is reproducibly observed that the surface area covered by a single dot is roughly proportional to the contact time, albeit with some exceptions – for example, in the case of very long contact times [7]. Based on this simple assumption, a single parameter can be used to describe the transport rate – namely a spreading constant C expressed in units of $\mu m^2\,s^{-1}$ (dashed line in Figure 5.2b). Once this transport rate has been determined, and is considered in a calibration, it then becomes possible to control dot dimensions and fabricate arbitrary patterns in a lithography process [8].

5.2.1
Theoretical Models for Ink Transport

The ability to obtain quantitative data on transport rates of inks, as well as the morphological information obtained by *in situ* AFM imaging, opens the possibility of testing theoretical models for the nanoscale ink transport in DPN. In addition to the perfectly round and sharp spots reproducibly achieved when patterning thiols on gold under optimal conditions, occasionally it can be observed that some spots appear more diffuse, are surrounded by a halo of lower lateral force microscopy (LFM) contrast or consist only of a ring. Furthermore, in some cases an 'anomalous diffusion' is observed where, instead of circles, fractal-like branches appear. Figure 5.3 shows schematics of four models that have been developed to explain and understand the different spreading phenomena observed in DPN experiments. The first three models (see Figure 5.3a–c) focus on the deposition of thiol SAMs on gold, where a strong chemical binding of the ink molecule to the substrate is expected and anomalous diffusion is not observed. The fourth model explains anomalous diffusion in terms of strong intermolecular interactions within the ink.

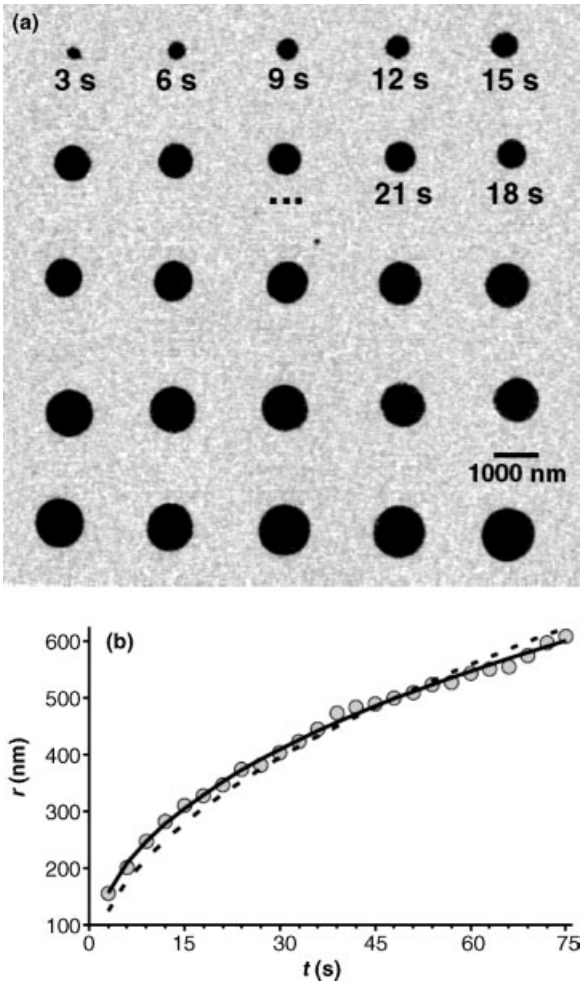


Figure 5.2 (a) Lateral force image of octadecanethiol dots deposited on a gold surface at different contact times; (b) A typical plot of the radius (r) of the spots versus contact time (t), including fits to two theoretical models. (Reprinted with permission from Ref. [9]; © 2002, American Physical Society.)

The first two models (Figure 5.3a and b) are based on diffusion theory, and are similar in that they both assume the tip to be an infinite point source, and use diffusion theory to describe the spreading of a thiol monolayer on a gold surface. The first model (Figure 5.3a) assumes a constant flux of molecules flowing from the tip, with a concentration of zero outside a spread island. That idea is consistent with the experimental observations that the area tends to increase linearly with contact time, and that the monolayer islands have sharp edges in the majority of AFM images. The second model (Figure 5.3b) assumes a constant concentration at the tip and an area of

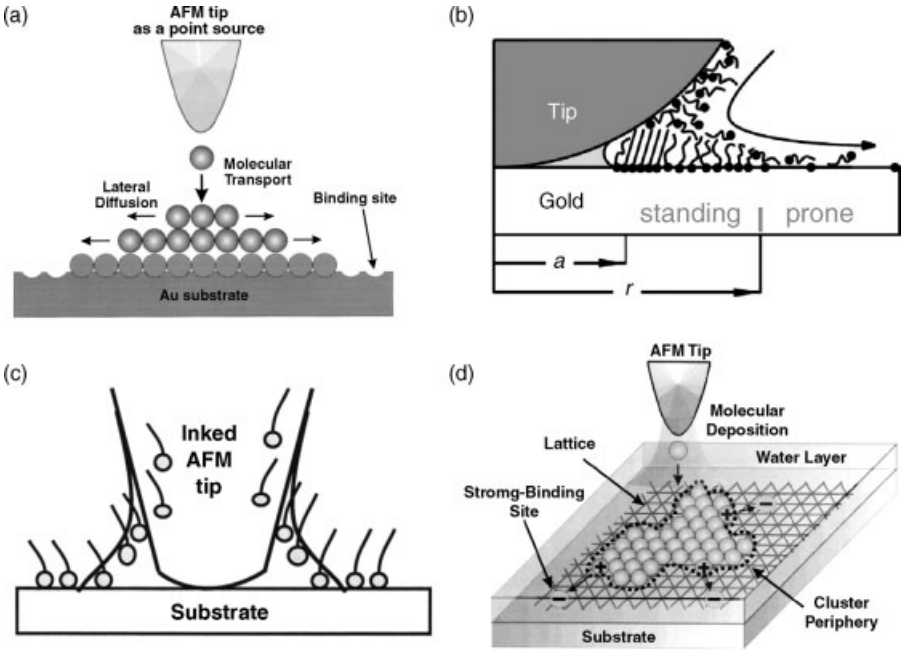


Figure 5.3 Drawings of different models of ink transport. (a) Diffusion model with constant flux from the tip. (Reproduced with permission from Ref. [14]; © 2001, American Institute of Physics.); (b) Diffusion model with a constant concentration at the tip. (Reprinted with permission from Ref. [9]; © 2002, American Physical Society.); (c) Meniscus interface transport model. (Reproduced with permission from Ref. [15]; © 2006, American Institute of Physics.); (d) Model for anomalous diffusion based on collective behavior. Reproduced with permission from Ref. [13]; © 2006, American Institute of Physics.)

a lower density of thiol molecules on the surface diffusing from the tightly packed monolayer. This second model allows for variation in the flux from the source, explains the occasional presence of halos and provides a slightly better fit to the experimental data (solid line in Figure 5.2b). The constant-concentration model also allows the derivation of an absolute diffusion coefficient using three physically relevant fit parameters (tip contact area, ratio of the concentration on the tip to that of a tightly packed monolayer and the diffusion coefficient).

In addition to the general idea of modeling the tip as a point source from which ink molecules diffuse, the effect that a microscopic condensed water meniscus forming at the tip–substrate contact point in the presence of humidity has been considered in order to further unravel the mechanisms by which the ink molecules make their way to the surface [10]. In particular, amphiphilic molecules such as MHA, which are only very slightly soluble in water, might be expected to show an affinity for the air–water interface. The meniscus interface transport model (Figure 5.3c) was therefore developed and can be used to explain the formation of hollow ring patterns. This model can also predict the transport behavior of a variety of amphiphilic molecules, including some that physisorb on the substrate.

Molecular dynamics simulations have suggested that SAM growth on the nanoscale involves a molecular basis that cannot be adequately described by analytical diffusion models [11]. For instance, it was shown that, even in the case of strong binding (e.g. alkanethiols on gold), the monolayer will grow as molecules from the tip displace molecules already bound to the surface, in a mechanism more akin to spreading than to diffusion. Computer simulations have also been able to recreate anomalous diffusion [12] *in silico* by considering the collective behavior of the molecules in a SAM (Figure 5.3d) [13]. There is evidence supporting aspects of each of these models, and there is no consensus as to which is the best. Further innovations from theoreticians, as well as carefully planned experiments designed to test the different models, are necessary to determine the true situation at the DPN tip.

5.2.2 Experimental Parameters Affecting Ink Transport

Several experimental parameters have been observed to influence the ink transport in DPN, including: driving forces (chemical interactions and external fields), ink composition, surface properties (chemistry and roughness), humidity, temperature and tip geometry. It is necessary to understand and control these parameters in order to optimize DPN processes for a particular application. This is especially important if different materials and nanostructures of the desired materials are to be integrated on the same substrate.

5.2.2.1 Driving Forces

In order for the ink to move from the tip to the substrate, a driving force is required, otherwise the ink will simply remain on the tip. Internal driving forces (i.e. in the absence of external fields) are typically based either on a chemical reaction of the ink with the substrate (chemisorption) resulting in a SAM, or on physical adhesion of the ink to the substrate (physisorption). Another approach is to apply an external field, for instance by heating the tip (thermal DPN) or applying a voltage (electrochemical DPN). It is worth noting that all DPN fabrication process require some sort of interaction between the ink and the substrate, even when the driving force is provided externally.

5.2.2.2 Covalent Reaction with the Substrate

The covalent reaction of the ink molecules with the substrate to form a SAM is the most straightforward and widely used driving force in DPN. The most reproduced and well-studied system is the formation of thiol SAMs, as described in Section 5.2.1. However, covalent self-assembly has also been used as a driving force for other ink molecules. As an example, considerable effort has been made to develop reproducible methods for DPN on semiconducting or insulating surfaces such as silicon and glass, as thiols are limited to self-assembly on metallic surfaces.

Functional silane molecules are widely used for the fabrication of SAMs on semiconductor surfaces, and are the therefore the first choice. However, a drawback of patterning silanes by DPN in air is that they polymerize in the presence of water,

and tend to be liquid in their monomeric state (thiol inks in contrast are typically solid at room temperature). Despite these challenges, it has been shown that through careful optimization of the experimental conditions – for example, selection of the molecule, control of humidity and functionalizing the AFM tip – it is possible to pattern SAMs of functional silanes [16–18]. Another approach is to choose a functional group for self-assembly that is not as sensitive to water; for example, silazanes have been shown to be suitable inks on semiconductor surfaces by covalent reaction with OH groups on the surface [19].

Perhaps the most generally applicable strategy for depositing and integrating arbitrary molecules on an arbitrary surface by DPN is to prefunctionalize the desired surface with a bulk SAM, which can then be covalently linked to the ink molecule of choice. This approach has been particularly useful for the patterning of biofunctional molecules (this will be discussed in more detail in Section 5.6). Briefly, it has been used successfully for direct DPN of synthetic macromolecules [20], peptides [21] and DNA [22].

5.2.2.3 Noncovalent Driving Forces

While chemisorption of the DPN ink results in highly stable patterns, covalent reactions tend to be highly specific – and therefore in some cases it is desirable to be able to deposit an ink noncovalently. Examples include patterning on inert substrates, the integration of different materials on the same substrate, and/or the fabrication of multilayer structures. Such noncovalent deposition is, for example, the method used for patterning with macroscopic pens.

Numerous examples of noncovalent patterning have been reported in the literature. For example, the first instance of controlled deposition of organic materials from an AFM tip was the deposition of thiols on mica [23]. Electrostatic interactions have been used as a driving force to pattern charged conducting polymers [24], as well as polyelectrolytes which could be used as templates for layer-by-layer assembly [25] on silicon substrates. Inorganic nanostructures were fabricated by depositing inorganic precursors dispersed in a copolymer surfactant or dissolved in an ethylene glycol solvent which wets the substrate [26, 27]. Luminescent polymer nanowires were patterned on glass using only adhesion as a driving force [6, 28]. Nanoparticles (Fe_3O_2 and gold) have been picked up by an AFM tip and deposited noncovalently in controlled fashion onto mica surfaces in air by DPN [29, 30]. Semiconductor precursors, which are expected to react with each other and precipitate CdS only in the water meniscus, were used as inks to fabricate semiconductor nanostructures on mica [31]. Another approach is to mix a functional molecule with a well-characterized ink, as has been demonstrated by the DPN patterning of binary ink mixtures [32]. Finally, surfactants can be added to the ink in order to tune the 'wettability' of the ink on the substrate, providing another parameter that can be used to control ink transport [33].

5.2.2.4 Tip Geometry and Substrate Roughness

In addition to chemical interactions between the ink and substrate, it is also apparent that the topography of the substrate and geometry of the tip play a role in ink

transport. The effect that these parameters have on the minimum feature size was systematically investigated in the case of alklythiol patterning on gold surfaces, where the smallest line widths (14 nm) could be achieved with the sharpest tips, and on the smoothest gold available [8]. In another study on the effect of tip-geometry, the AFM tips were deliberately made blunt using laser ablation [34]. It was then found that not only did the minimum feature size depend on the tip radius, but also the rate of ink transport – an idea consistent with several of the ink transport models described above.

5.2.2.5 **Humidity and Meniscus Formation**

A significant amount of evidence is available which suggests that the transport of ink molecules with polar groups (such as the thiol MHA) are heavily dependent on humidity, with higher humidity showing higher diffusion constants. Although there seems to be only a slight (if any) humidity dependence for the nonpolar molecule ODT [9, 10], the effect cannot be ignored in the patterning of just about all other molecules. *Humidity* is therefore an important parameter that must be controlled in order to optimize DPN conditions. Ideally, this is achieved by encasing the DPN apparatus in an environmental chamber, or locally by placing a water-containing capillary tube near the atomic force microscope tip. Although the possibility that the humidity might affect the ink properties or substrate reactivity has not been excluded, it is generally thought that the mechanism for humidity dependence depends on the presence of a meniscus that condenses at the tip of an AFM when it contacts a surface in the presence of humidity [35].

Theoretical studies of meniscus formation at an atomic force microscope tip based on Monte Carlo simulations have predicted that the meniscus should depend not only on humidity, but also on the tip geometry and surface chemistry [36]. Striking images confirming the presence of such meniscus formation (and indeed showing that the meniscus can grow larger than expected) have been made possible by using environmental scanning electron microscopy (ESEM), as shown in Figure 5.4 [37]. Interestingly, studies of the kinetics of meniscus formation between an atomic force microscope tip and gold surfaces showed similar trends as the early patterning rates of thiols on gold. It has therefore been hypothesized that growth of the water meniscus may in some cases be the rate-limiting step in DPN ink transport [15, 38].

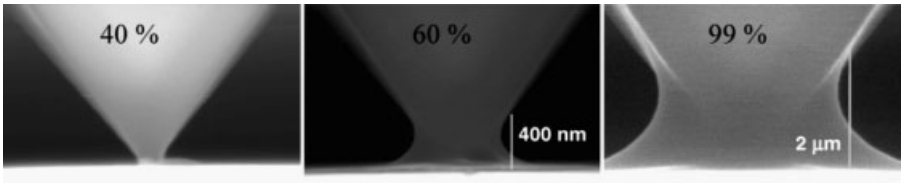


Figure 5.4 Environmental scanning electron microscopy (ESEM) series of meniscus formation on a cantilever tip at three different relative humidities (left, 40%; center, 60%; right, 99%). (Reproduced with permission from Ref. [37].)

It should be noted that, although the humidity clearly influences ink transport in DPN, it does not appear to be a prerequisite, as patterning has been achieved at 0% humidity and even in ultra-high vacuum [9, 10].

5.2.2.6 External Driving Forces

Another approach to controlling the transport of materials from an atomic force microscope tip to the surface is to apply an external driving force between the tip and the substrate. Although this poses an engineering challenge in fabricating externally addressable tips, the ability to switch a particular pen on and off greatly increases the versatility, especially when one considers parallel arrays of tips where each may be addressable for large scale integration. Two examples are the use of heatable cantilevers (thermal DPN) and the application of a voltage between the tip and sample (electrochemical DPN).

5.2.2.7 Thermal DPN

The idea of thermal DPN is similar to that of a soldering iron; that is, the material on the tip of the microscope should be heated above its melting temperature in order to facilitate transport to the surface. The concept is shown schematically in Figure 5.5. Although heating the ink can be a disadvantage for biological inks, which may be sensitive to high temperature and dehydration, it provides a useful means of patterning other materials. For example, octadecylphosphonic acid (OPA) was the first compound to be patterned on silicon by using thermal DPN [39]. It was observed that OPA only began to write when the tip was heated above a critical temperature.

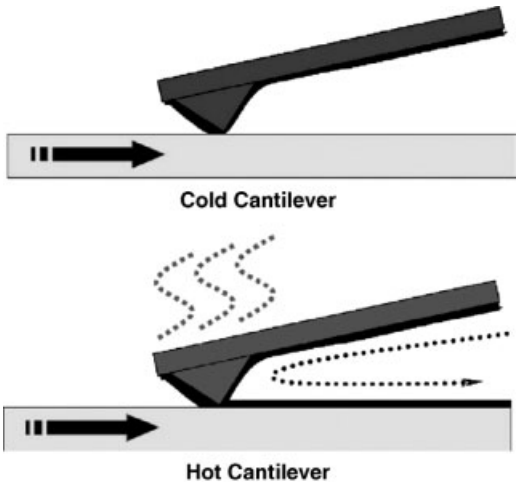


Figure 5.5 Schematic of the concept of thermal DPN. At low temperature, the ink does not transfer from the tip (top), while upon heating the tip the ink flow can be controlled (bottom). (Reproduced with permission from Ref. [39]; © 2004, American Institute of Physics.)

The method has since been applied to the deposition of conducting polymers [40]. Although it was initially suggested that thermal DPN is necessary for organic compounds with high melting points, it has since been shown that such compounds can also be patterned at room temperature (well below their melting temperatures) by humidity-controlled DPN. For instance, OPA as well as other compounds with melting points up to 230 °C have been patterned at room temperature under the appropriate humidity [41]. The mechanism of transport in thermal DPN of organic inks therefore remains unclear, although it appears to be possible to control the ink transport by controlling the tip temperature. Most striking is that indium metal nanostructures could be directly written by thermal DPN [42]. Furthermore, by filling a single carbon nanotube with molten copper and then dispensing it under observation with a transmission electron microscope, it has been suggested that using such a carbon nanotube-based spotwelder in thermal DPN might enable the ultra-high resolution of molten metals by thermal DPN [43]. Such direct, nanoscale writing of water-insoluble metals has not been shown to be possible below the melting point of the metal.

5.2.2.8 **Electrochemical DPN**

Another approach to controlling the transport of ink from the atomic force microscope tip is to use the water meniscus as a nanoscale electrochemical cell, where metal salts can be dissolved, reduced and precipitated to form metal nanostructures on the surface; this method, which is referred to as electrochemical DPN (E-DPN), is illustrated schematically in Figure 5.6 [44]. This method was further applied to the controllable transport of his-tagged proteins, which have an affinity to certain metal ions such as Ni^{2+}. By carrying out E-DPN on nickel-coated surfaces, the surface could be locally ionized, thereby allowing proteins on the microscope tip to transport the surface and bind. The same approach of combining local surface oxidation with material transport from the atomic force microscope tip was used to locally oxidize a pre-existing SAM and to simultaneously deposit organic inks to those same areas [45].

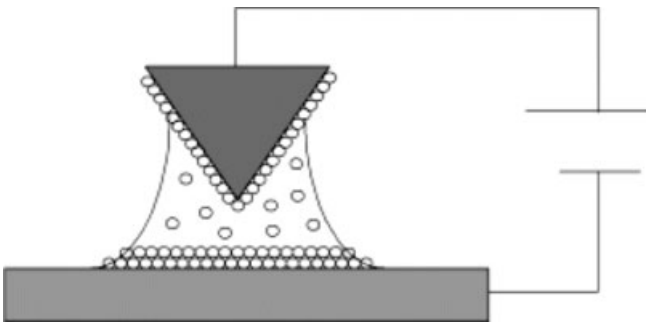


Figure 5.6 Schematic of electrochemical DPN (E-DPN). The water meniscus that condenses in air between the AFM tip and sample is used as a nanometer-sized electrochemical cell. (Reproduced with permission from Ref. [44].)

5.3
Parallel DPN

5.3.1
Passive Arrays

The constructive and chemically driven nature of DPN makes it uniquely amenable to being carried out in parallel, using arrays of tips, and without the need for accessing each tip electronically for force feedback. A rough alignment of the tip-array with the surface is sufficient, since if the tips are touching the surface then the ink will be transported at a constant rate which is determined primarily by the ink–substrate combination. The first demonstration that DPN could be readily carried out in parallel, employed micromachining processes to fabricate one-dimensional arrays with 32 silicon nitride tips or eight boron-doped silicon tips, the latter having sharper tips at the expense of pen densities [46]. The number of tips in a single linear array was then scaled up to the centimeter scale, using arrays of up to 250 tips, all of which wrote simultaneously [47]. Parallel DPN was then scaled up again to a two-dimensional arrays of tips that covered a square centimeter and consisted of 55 000 probes writing simultaneously; an example is shown in Figure 5.7 [48]. The parallel and constructive capabilities of DPN are what give it the potential to integrate materials on unprecedented scales.

5.3.2
Active Arrays

One factor which limits the complexity of patterns that can be generated by parallel DPN as described above, is that every tip necessarily writes the same

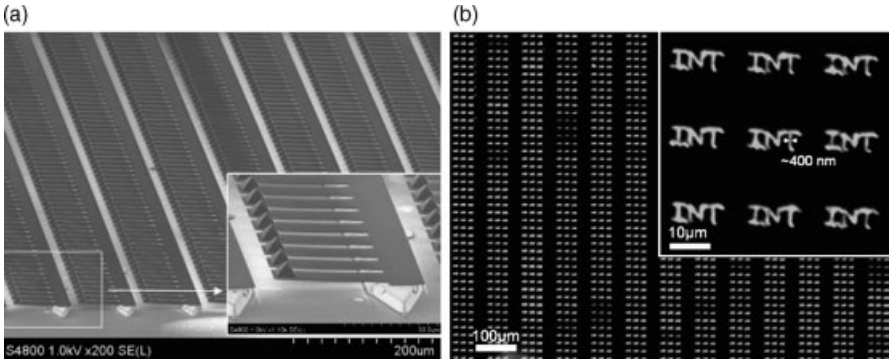


Figure 5.7 Massively parallel dip-pen nanolithography (DPN) with two-dimensional (2-D) tip arrays. (a) Scanning electron microscopy image of a small section of a 55 000 tip array covering an area of 1 cm^2. (Image courtesy of NanoInk.); (b) Fluorescence image of phospholipid patterns generated with the 2-D arrays at a throughput of 5 $cm^2\ min^{-1}$ [49].

pattern, provided that it is coated with ink. That is, it would be impossible to get each tip in an array to write a different pattern using passive tips. In addition to the possibility of controlling the driving force by external fields (e.g. by thermal DPN or E-DPN, as described earlier), another innovative step in the development of DPN probes for parallel materials integration was taken in which the cantilevers could be externally actuated. The first approach of this type used thermal bimorph cantilevers, where heating one side of the cantilever caused it to bend down so that the tip contacted the substrate and the ink could flow [50]. Again, by optimizing the tip fabrication the resolution of patterns generated by active pen arrays could be brought down below 100 nm [51]. Another approach would be to use electrostatically actuated cantilevers, where the cantilever bends towards the surface based on an applied electric field in order to avoid the possibility of unwanted heating of the tip or thermal crosstalk between neighboring cantilevers [51]. Actuated probes not only increase the available complexity of patterns that can be generated by a single ink, but also open the door to the integration of different materials from different tips in an array on an area smaller than the dimensions of the tip array itself. This can be done by writing with one tip, and then moving the array such that a neighboring tip writes on or near the same area that has already been patterned.

5.4 Tip Coating

5.4.1 Methods for Inking Multiple Tips with the Same Ink

An essential part of any DPN process is to bring the ink onto the atomic force microscope tips. This is typically achieved either by thermally evaporating the ink onto the tips, or by immersing the tips in the ink in a type of dip-coating process. Thermal evaporation is rather straightforward and typically results in a homogeneous ink coating, for instance of ODT. However, the majority of ink molecules – and especially the more polar ones – do not seem suited to evaporation, and tend to function better when the tip is coated from solution. For solution coating, the entire cantilever chip can be dipped by hand into an ink solution and, upon removing the tip and allowing the solvent to dry (e.g. under a stream of inert gas), a typically homogeneous coating results. However, the distribution of the ink on the tip when coated in this way will inevitably depend on exactly how the ink wets (or de-wets) the tip, and also on how the ink solutes concentrate on the tip as it is dried. Functionalization of the atomic force microscope tip before coating is therefore sometimes beneficial. For instance, in order to reproducibly pattern proteins it was found useful to precoat the tips with a SAM of a thiolated polyethylene glycol [11-mercapto-undecylpenta(ethylene glycol)disulfide(PEG)], which makes the tip hydrophilic but prevents the denaturing of adsorbed proteins [52].

5.4.2 Ink Wells

Inking the tips in the ways described above is well suited for single tips, or in the case that the requirement is to coat all tips with the same ink. However, in order to take advantage of the potential for DPN to integrate different materials on the same substrate in parallel, it is necessary to deliver different inks selectively to different tips in an array. For this purpose, microfluidic ink-delivery systems have been developed [53, 54]. An example of tips being dipped in wells that coat only every second tip in an array is shown in Figure 5.8a. An array where only every second tip is coated is useful for many experiments that require uncoated tips as negative controls, or in cases where there is a need to have the patterns spaced further apart than the spacing of the tips in the array. Today, ink wells are available commercially (from the company NanoInk) that allow the integration of up to 24 different inks on a one-dimensional array. Figure 5.8b shows an example of two different fluorescently labeled phospholipids integrated on a single cantilever array [49]. In similar fashion, a chip has also been developed that allows local vapor coating onto tip arrays [55].

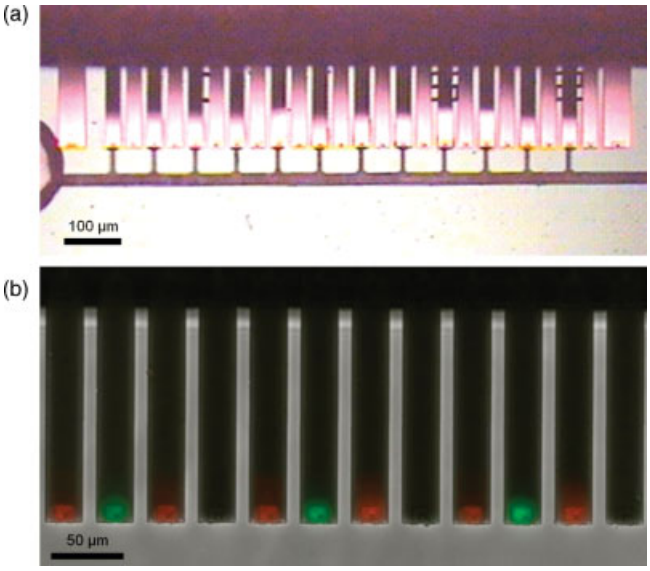


Figure 5.8 One-dimensional tip arrays coated with phospholipids using ink wells. (a) Optical micrograph of tips in contact with the ink wells. Every second tip in the array is being dipped with one ink; (b) Multi-channel fluorescence and bright-field micrograph of phospholipids doped with different inks. Every second tip is coated with a red dye, every fourth tip with a green dye, and the remaining tips function as negative controls.

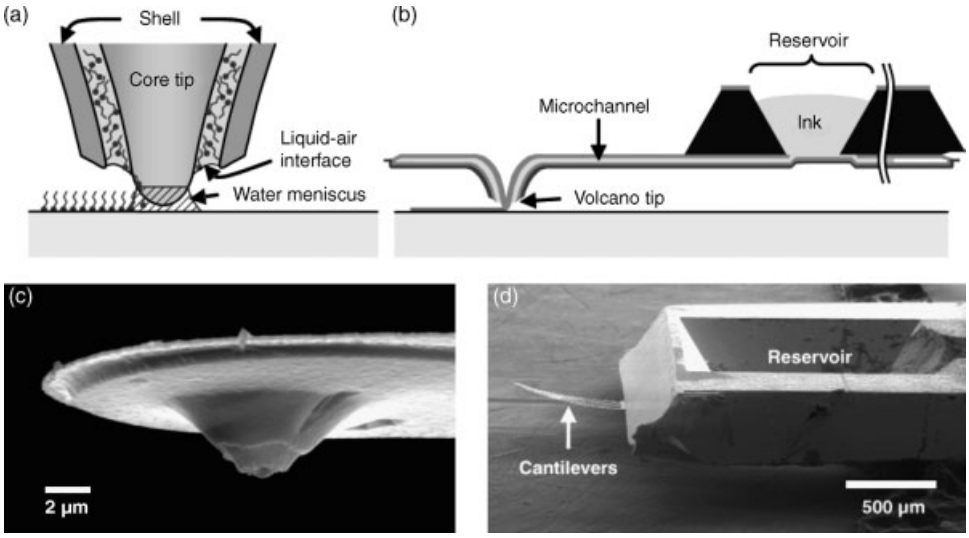


Figure 5.9 Fountain pens. (a) A cross-sectional schematic of the volcano-like tip in the process of writing; (b) A schematic of the entire chip including the reservoir; (c, d) SEM images of the tip and entire chip, respectively. (Reproduced with permission from Ref. [56]; © Wiley-VCH Verlag GmbH & Co. KGaA.)

5.4.3
Fountain Pens

One particularly elegant approach to delivering ink to the atomic force microscope tip is through the integration of microfluidic channels directly onto the tip itself, in order to generate a nanofountain probe (NFP), such as that shown in Figure 5.9 [56]. As standard microfabrication techniques were used, it has been possible to generate parallel arrays of NFPs integrated on a single chip, with different ink reservoirs leading to different tips in the same array, thus enabling the parallel integration of different inks [57]. As in the case of macroscopic pens, NFPs can be expected to be particularly useful for the patterning of inks where the solvent must remain in the ink until after patterning, as tips coated by dipping in solution are subject to drying.

5.4.4
Nanopipettes

Although micropipettes and nanopipettes differ technically from DPN (in that they do not necessarily utilize an atomic force microscope tip), they are conceptually similar to DPN and NFPs in several ways, and are therefore worthy of brief mention at this point. Cantilevered micropipettes similar to those used for scanning near-field optical microscopy (SNOM) have been used for the local delivery of an etchant to a chrome film, with a resolution of 1 μm [58]; indeed, subsequent studies led to the

fabrication of spots of 280 nm diameter [59]. In similar manner, an electrochemical fountain pen was used to fabricate freestanding platinum nanowires with a diameter of 150 nm. By using a voltage-controlled feedback circuit derived from scanning ion conductance microscopy (SICM), submicron and multicomponent features consisting of biomolecules such as DNA and protein have also been fabricated [60, 61]. Although the practical resolution of micropipettes and nanopipettes is much lower than for DPN, and it is difficult to imagine them being used in parallel, they do have a significant advantage for the patterning of biomolecules in that they are able to function under water [62].

5.5 Characterization

In addition to tip inking and writing, a third indispensable part of the DPN process is the characterization and quality control of the resultant patterns. A convenient capability of DPN, which is also shared by most scanning probe-based lithography processes, is that the same tip can be used for both patterning and imaging, in the case of DPN by AFM. In particular, lateral force imaging is typically used for the characterization of chemical contrast in covalently bound inks such as thiols on gold (as described earlier and shown in Figure 5.2). There are two practical issues to be aware of when characterizing DPN patterns generated by the same tip that has been used for imaging:

- An inked tip will typically continue to write during imaging, and therefore high scan speeds must be used to minimize this effect.
- The vast majority of DPN is carried out in contact mode, using cantilevers with a too-low spring constant for use in intermittent contact or tapping mode imaging in air. This is especially the case for parallel DPN, where it is impractical to have a separate tapping feedback mechanism for each tip. As the contact mode typically provides inaccurate heights in air, it is often necessary to change the tip and realign it to find the patterned area in order to obtain quantitative height information.

Although these two issues represent disadvantages in a high-throughput lithography process, they can actually serve as significant advantages when characterizing the ink transport. For instance, in an early study using DPN, monolayer growth could be observed *in situ* by scanning the same area repeatedly at high resolution; this allowed observation of the monolayer growth dynamics as alkanethiols were transferred from the tip to a gold substrate [63]. It is also possible to carry out DPN in tapping mode by using a single tip for the simultaneous deposition and imaging of soft materials, as well as obtaining accurate height information [64]. When this method was applied to the deposition of poly-D,L-lysine hydrobromide onto mica surfaces, it was possible to observe the nucleation and growth dynamics of polymer crystals at a submicron scale that is inaccessible to other methods (Figure 5.10) [65].

In the DPN patterning of a new ink or substrate, it is essential to determine that the patterns generated are indeed composed of the intended ink. If a particular

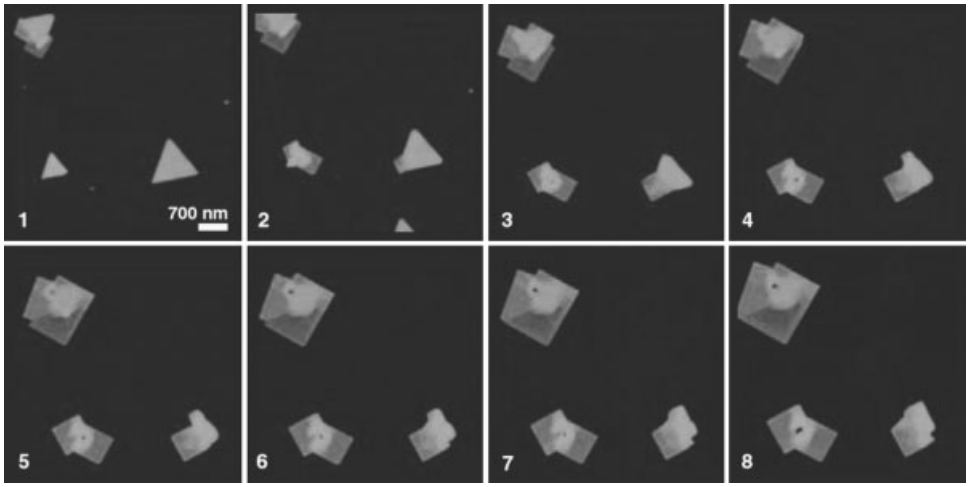


Figure 5.10 Topographical AFM image sequence (1 to 8) of epitaxial crystal nucleation and morphology changes from threefold to fourfold symmetry during growth of poly-D,L-lysine crystals on a mica surface as molecules are transferred from the microscope tip during each scan. (Reproduced from Ref. [65] with permission from the American Association for the Advancement of Science.)

topographical morphology is known, then this information can be obtained from AFM measurements. For instance, the AFM images of DPN-patterned collagen fibrils showed a helical repeat of 65 nm, which was consistent with scanning electron microscopy (SEM) observations of collagen fibrils. Another example is the formation of anisotropic structures formed during the DPN patterning of peptide amphiphiles [66]. Another way of distinguishing DPN patterns from artifacts that might result from mechanical contact of the tip, condensed water or residual solvent, would be to carry out the appropriate negative controls using either uncoated tips or tips dipped only in the solvent, in the absence of the desired ink molecules.

More often than not, it is necessary to confirm the chemical identity of the ink molecules on the surface by other analytical methods. For example, X-ray photoelectron spectroscopy (XPS) is often used to identify elements present in DPN patterns [17, 21, 25, 29, 31, 67–69]. Infrared spectroscopy [68] and mass spectrometry [67] are also powerful characterization tools that have the added advantage of providing structural information. Furthermore, AFM is a rather slow characterization method and, in the case of high-throughput, parallel DPN characterization can easily become the rate-limiting step in the fabrication process. In that case, optical characterization is ideal, and this can be achieved by using DPN-generated thiol monolayers as a resist against chemical etching so that the patterns become visible under optical microscopy [48]. Another approach would be to use a fluorescently labeled ink, which not only enables rapid characterization but also provides some chemical information about the patterns [6, 28, 49, 70–72]. Optical characterization using scanning near-field optical microscopy (SNOM) detection of single molecules deposited by DPN

suggests that the practical resolution limits of DPN may not be in the patterning, but rather in the ability to detect patterns with dimensions smaller than the AFM tip used to fabricate them [6]. Finally, it is often useful to characterize the AFM tips as well as the presence and distribution of the ink on the tips, which can typically be achieved using SEM or optical microscopy.

5.6 Applications Based on Materials Integration by DPN

Functional chemical patterns fabricated by DPN have been used for a wide variety of scientific applications, and it can be expected that industrial applications will follow. Even in the many published applications, including for example etch resists [73] or templates for selective deposition [74], when only a single ink molecule is patterned onto a single surface, DPN has several advantages over conventional direct-write lithographic techniques such as EBL. While the latter is able to provide competitive lateral resolution, it is severely limited in its throughput as well as its cost. Furthermore, in contrast to DPN, EBL involves removing material from the substrate, which requires an extra development step. One advantage of placing resists and template molecules directly onto the surface is that the remainder of the surface is left free of contaminants. That being said, the ability to generate multicomponent nanostructures opens entirely new possibilities that are inaccessible by any other method. Some of the more striking examples will be briefly described in the following sections, in order to provide an overview of the types of unique application made possible by DPN. Whilst the examples are categorized based on selective adsorption, combinatorial chemistry and biological arrays, these categories are by no means complete and a significant amount of overlap is apparent between them.

5.6.1 Selective Deposition

Although the selective deposition of materials onto patterned surfaces is not limited to DPN patterns, the rapid prototyping capabilities and ability for DPN to generate multicomponent nanostructures on a small scale adds a new dimension to the field. A few of the strategies that can be used to immobilize different particles from solution by selective adsorption or templating are shown in Figure 5.11. In addition to the adsorption strategies shown, covalent binding, nonpolar adhesion forces and entropic effects can also be used to direct binding towards desired areas of the substrate. The surface passivation of the background is often a crucial step in fabricating templates for selective adsorption. Most often, successful selective adsorption will involve combinations of more than one of these strategies.

As a first example of electrostatic templating on DPN patterns, positively charged colloidal particles were immobilized electrostatically onto negatively charged MHA patterns. By fabricating a variety of MHA dot array patterns with different dot sizes and spacings, it was possible to screen the pattern dimensions for those capable of

Figure 5.11 Schematic representation of different strategies for selective adsorption or templating. Left to right: charge-based recognition; macromolecular encoding (i.e. DNA); and specific binding of ligands. (Image courtesy of NanoInk.)

organizing the particles such that each dot had exactly one particle bound, in a combinatorial fashion [74]. Such an approach was later applied to the fabrication of arrays of individual bacterial cells [75]. In another approach, the positive and negatively charged polyelectrolytes poly(diallyldimethylammonium) chloride (PDDA) and poly(styrenesulfonate) (PSS) have been directly patterned on silicon surfaces by DPN. Upon the addition of a complementary polyelectrolyte, selective adsorption was observed which suggested a compatibility of the method with layer-by-layer assembly [25]. Furthermore, such electrostatic templates have been used in combination with *molecular combing* to organize *aligned DNA strands*, a biological molecule which also falls into the polyelectrolyte category and can readily be adsorbed electrostatically [76].

DNA-directed self-assembly on DPN patterns has been used to organize two different-sized nanoparticles into nanoarrays, with the spacing between different-sized particles in an array being as low as 500 nm [74]. This method of self-sorting was improved by using noncomplementary DNA as a *passivation layer*, enabling larger particles to be immobilized [87]. Protein nanoarrays selectively bound to MHA patterns [77], or patterned by direct write methods [52], have been used to subsequently immobilize other protein molecules by molecular recognition. Covalent linking of the biofunctional group biotin was carried out selectively on DPN patterns, enabling the binding of streptavidin protein by molecular recognition and subsequent binding of biotinylated materials [72]. Similarly, the covalent coupling of proteins to DPN templates was carried out using succinimide chemistry [78]. Finally, nonpolar carbon nanotubes were selectively adsorbed to DPN templates based on differences in surface energy between the patterned and passivated regions [79].

Although not often emphasized in the literature, *passivation* is a crucial step in selective adsorption, such that the materials to be integrated on the surface do not simply bind everywhere. In a DPN experiment, the background is typically blocked by a low-surface energy (methyl-terminated) compound such as ODT [74] or, in the case

of protein or cells, the adsorption of PEG-terminated passivation layers is preferred due to their nondenaturing character [52]. Nonspecific adsorption is in fact a major problem that limits the applicability of templating-based fabrication methods, although it can be overcome to some degree through chemical modification of the surface and optimization of the solution conditions (e.g. pH and ionic strength). Nonetheless, the fact that nonspecific adsorption cannot be completely eliminated is a qualitative advantage that constructive, parallel DPN has over serial methods based on sequential selective adsorption and patterning steps.

5.6.2 Combinatorial Chemistry

The idea of placing different chemical compounds on the same surface, for exposure to identical solution conditions, lends itself well to combinatorial chemistry. In addition to the possibility of ultra-high-density chemical arrays, the nanoscale resolution of DPN also enables studies of collective molecular interactions, as well as how the properties of nanoscale aggregates might differ from bulk behavior. For instance, the ability for nanopatterned SAMs to function as resists against the chemical etching of metallic films has been investigated combinatorially as a function of pattern dimension in order to minimize feature sizes [73]. Solid-state nanostructures with features as low as 15 nm have been fabricated by the direct deposition of etchant [67], while nanostructures of various metals such as gold, silver and palladium have been generated with 35 nm dot diameters and 53 nm line widths [80, 81].

As a first demonstration of the ability to screen the chemical behavior of different compounds on the same surface, four different thiol molecules were patterned within an area of 5 μm^2 on a single gold surface to form combinatorial libraries, as shown in Figure 5.12. The libraries were then used to study molecular

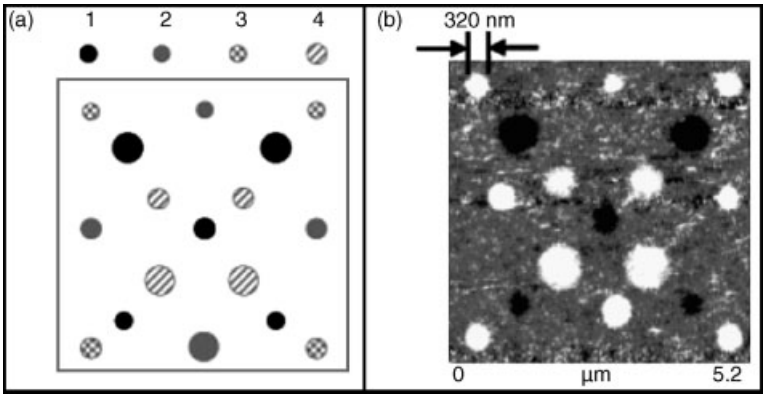


Figure 5.12 (a) Schematic representation of the combinatorial library design consisting of the four different molecular inks; (b) Lateral force microscopy image of the library described in panel (a). (Reproduced with permission from Ref. [82].)

displacement from the surface by repeatedly scanning the same area with an ink-coated tip and observing the order in which the spots disappeared as they were exchanged by new molecules from the tip [82]. In another creative approach to combinatorial chemistry, DPN was used to pattern molecules on cantilevers which could, in principle, be used as force sensors for determining interactions between molecular libraries [83].

Another unique application for DPN is in the study of microscale and nanoscale phase separation. For instance, phase separation and pattern formation in conjugated polymers spin-coated onto combinatorially nanostructured DPN templates was studied as a function of polymer concentration and MHA dot diameter [84]. The information resulting from those screenings of pattern dimensions enables one to control pattern formation in thin organic films, with potential applications in the fabrication of organic electronic and optical devices. Furthermore, by coating the tip with a mixture of two (or more) inks it becomes possible to observe if, and how, the molecules de-mix during the DPN writing process by using *in situ* AFM measurements [85]. The use of such separated phases for selective adsorption provided the ability to reduce line widths down to 10 nm.

5.6.3 Biological Arrays

The complex integration of biological molecules such as DNA, protein and phospholipids on hierarchical scales, ranging from molecular dimensions up to the size of entire organisms, provides the basis for the molecular machinery that makes life possible. The ability to understand, control and even mimic such interactions has long been a dream, not only for biologists but also for nanoscientists in a variety of disciplines. Robotic spotting methods have already proven valuable in the biotechnology industry for the fabrication of biological arrays, even with spot sizes on the order of hundreds of micrometers. Smaller spots not only have evolutionary advantages of higher spot density, sensitivity and lower requirements for sample volumes, but also open entirely new possibilities. For instance, the ability to integrate more than one biomolecule or biological entity (e.g. a virus) under the surface area covered by a single adherent cell in a spatially defined manner is especially exciting for unraveling the roles of intermolecular interactions in biological systems.

The patterning of DNA by DPN was first achieved by the selective deposition of oligonucleotides onto thiol-patterned surfaces [86]. Later, reproducible protocols for direct deposition from the AFM tip were developed [22, 87]. Figure 5.13 shows an example of a binary DNA array that was used to reversibly hybridize both fluorescently labeled oligonucleotides as well as gold nanoparticles of two different diameters [22]. This illustrates the ability for DNA arrays to organize nanomaterials in a highly specific manner. Furthermore, whilst fluorescence is an invaluable tool for characterizing biomolecular arrays, many biological materials are not easily obtainable in functional form with fluorescent labels. Therefore, label-free detection methods such as AFM topography show great promise [71]. The physical–chemical

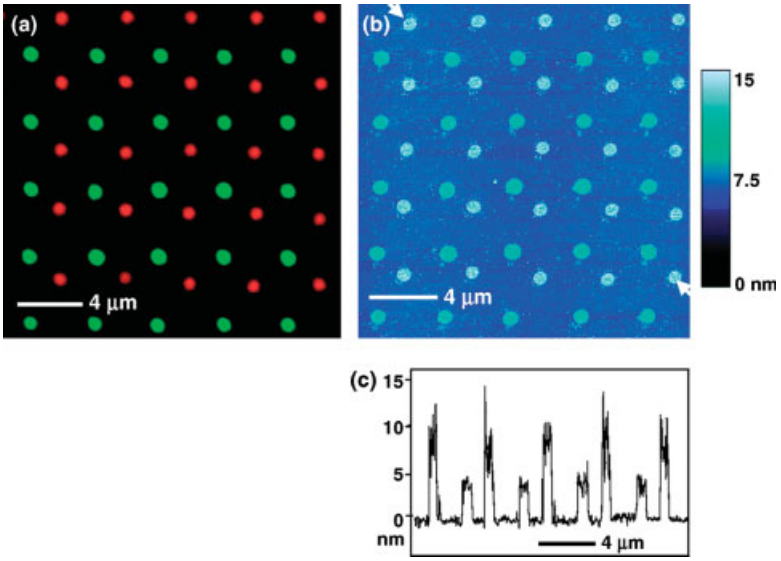


Figure 5.13 A binary DNA array fabricated by direct-write DPN. (a) Fluorescence image showing two different fluorescent probes selectively hybridized to the different spots; (b) AFM topography of the same area after dehybridization of the fluorescent probes and attachment of oligonucleotide-labeled nanoparticles of two different diameters (5 nm and 13 nm). (Reproduced from Ref. [22] with permission from the American Association for the Advancement of Science.)

similarity between different DNA strands of the same sequence renders them promising for parallel arraying, as DNA strands of different sequence can be patterned under the same environmental conditions.

A variety of different proteins have also been patterned by DPN, both by selective adsorption and by direct-write processes. In the majority of cases, selective adsorption using various coupling strategies appears to be most successful in fabricating functional protein nanoarrays to date, as selective adsorption can be carried out without dehydration of fragile proteins [70, 72, 77, 78, 88, 89]. For example, an antibody nanoarray-based detection assay for HIV in patients' plasma exceeded the limit of detection of conventional enzyme-linked immunosorbent assay (ELISA)-based immunoassays (5 pg ml^{-1} plasma) by more than 1000-fold [90]. Direct-write approaches hold the promise of larger-scale integration, and have also proven their ability to successfully generate functional protein nanoarrays. For example, collagen was the first protein to be directly deposited onto gold by first reducing disulfide bridges in the biopolymer, and then allowing the collagen fibrils to reassemble on the gold surfaces. Functional his-tagged proteins have been deposited directly onto nickel oxide surfaces via metal affinity [91]. Unmodified antibodies were directly patterned onto gold surfaces by nonspecific adsorption [52].

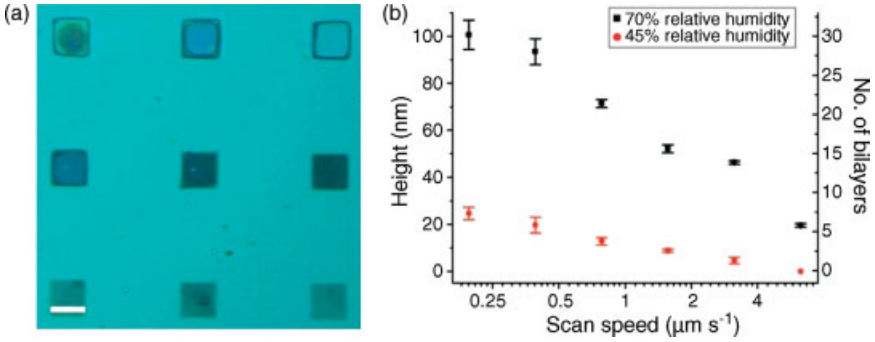


Figure 5.14 Phospholipid DPN patterns with control of multibilayer stacking. (a) Reflection-mode optical micrograph of phospholipid squares patterned on plasma-oxidized silicon at various speeds (scale bar = 5 μm); (b) The height of phospholipid multilayers (and corresponding number of bilayer stacks) measured by AFM is plotted as a function of scan speed (on a logarithmic scale) at two different relative humidities. (Reproduced with permission from Ref. [49]; © Wiley-VCH Verlag GmbH & Co. KGaA.)

In addition to DNA and protein, phospholipids represent another ubiquitous biomolecule that can be patterned noncovalently by DPN on a variety of surfaces, with linewidths as small as 100 nm [49]. In contrast to the transport behavior of most other inks – where the dot and linewidth can be controlled by the tip-contact time and scan speed, as well as humidity – phospholipid inks tend to stack into multilayer structures where the thickness of the film can be controlled by those same parameters (Figure 5.14). Upon immersion into aqueous solution, the multilayers can be spread on the surface to form a single monolayer, a lipid bilayer membrane, or remain as stable multilayers, depending on the substrate. A unique property of phospholipids is that they amphiphilic and are lyotropic liquid crystals that tend to organize into different supramolecular structures, where their fluidity depends on the hydration. Their transport rate from ink wells onto the tips, as well as from the tip to the substrate, can therefore be precisely controlled.

If one views viruses as nanoparticles, the methods described earlier for templating can be applied directly to the immobilization of viruses. For example, arrays of virus particles have been generated using genetically modified cow pea mosaic virus that covalently couple to maleimides patterned by DPN [32, 92]. Another approach based on several steps of selective adsorption to the DPN pattern of MHA has resulted in arrays of influenza virus particles [70] and of single tobacco mosaic virus particles [93]. In this process, Zn^{2+} are first bound to the MHA nanopatterns, after which antibodies specific for the virus in question are selectively adsorbed; finally, the virus particles are adsorbed on top of those two layers. Remarkably, it was shown that this method could be used to position functional viruses capable of infecting cells cultured on the arrays, as illustrated in Figure 5.15. The use of DPN-based methods to organize biological molecules and particles on a subcellular level, and to interface them with living systems, opens numerous possibilities in the emerging field of nanobiology.

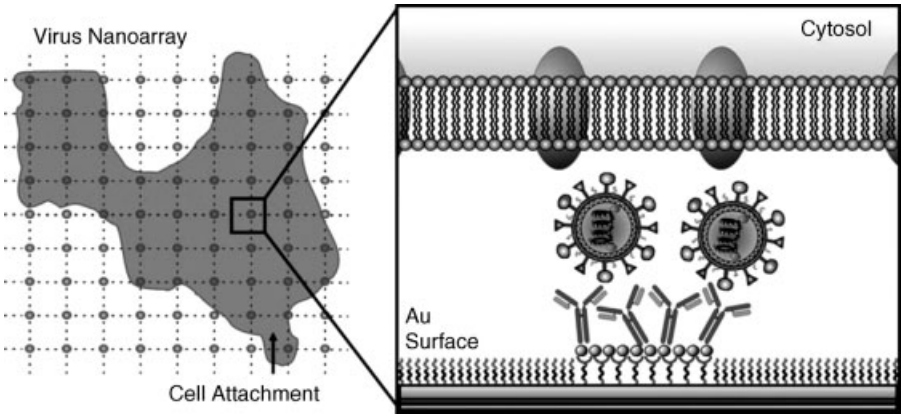


Figure 5.15 Schematic illustration of experiments where cells were cultured on and infected by DPN-patterned virus arrays. (Reproduced with permission from Ref. [94]; © Wiley-VCH Verlag GmbH & Co. KGaA.)

5.7 Conclusions

In conclusion, DPN provides the ability to simultaneously integrate and nanostructure a diverse range of materials on a variety of surfaces, at unprecedented levels of both spatial resolution and complexity. In principle, several thousand different materials could be integrated in thousands of different combinations with nanoscale resolution over square-centimeter (or larger) surface areas. Although much remains to be done before this dream is achieved, the barriers appear to be surmountable. Clearly, a basic understanding of the mechanisms behind the nanoscale transport of ink will be necessary in order to reproducibly carry out and extend the capabilities of DPN. In addition, innovative probe designs and inking strategies will be essential in order to expand the capability of DPN for large-scale materials integration.

References

1 Piner, R.D., Zhu, J., Xu, F., Hong, S.H. and Mirkin, C.A. (1999) *Science*, **283**, 661.

2 Ginger, D.S., Zhang, H. and Mirkin, C.A. (2004) *Angewandte Chemie - International Edition*, **43**, 30.

3 Salaita, K., Wang, Y.H. and Mirkin, C.A. (2007) *Nature Nanotechnology*, **2**, 145.

4 Haaheim, J. and Nafday, O.A. (2008) *Scanning*, **30**, 137.

5 Hong, S.H., Zhu, J. and Mirkin, C.A. (1999) *Science*, **286**, 523.

6 Noy, A., Miller, A.E., Klare, J.E., Weeks, B.L., Woods, B.W. and DeYoreo, J.J. (2002) *Nano Letters*, **2**, 109.

7 Hampton, J.R., Dameron, A.A. and Weiss, P.S. (2005) *The Journal of Physical Chemistry B*, **109**, 23118.

8 Haaheim, J., Eby, R., Nelson, M., Fragala, J., Rosner, B., Zhang, H. and Athas, G. (2005) *Ultramicroscopy*, **103**, 117.

9 Sheehan, P.E. and Whitman, L.J. (2002) *Physical Review Letters*, **88**, 156104.

10 Rozhok, S., Piner, R. and Mirkin, C.A. (2003) *The Journal of Physical Chemistry B*, **107**, 751.

11 Ahn, Y., Hong, S. and Jang, J. (2006) *The Journal of Physical Chemistry B*, **110**, 4270.

12 Manandhar, P., Jang, J., Schatz, G.C., Ratner, M.A. and Hong, S. (2003) *Physical Review Letters*, **90**, 4115505

13 Lee, N.K. and Hong, S.H. (2006) *Journal of Chemical Physics*, **124**, 11471

14 Jang, J.Y., Hong, S.H., Schatz, G.C. and Ratner, M.A. (2001) *Journal of Chemical Physics*, **115**, 2721.

15 Nafday, O.A., Vaughn, M.W. and Weeks, B.L. (2006) *Journal of Chemical Physics*, **125**, 144703

16 Kooi, S.E., Baker, L.A., Sheehan, P.E. and Whitman, L.J. (2004) *Advanced Materials*, **16**, 1013.

17 Sheu, J.T., Wu, C.H. and Chao, T.S. (2006) *Japanese Journal of Applied Physics Part 1 - Regular Papers Short Notes and Review Papers*, **45**, 3693.

18 Jung, H., Kulkarni, R. and Collier, C.P. (2003) *Journal of the American Chemical Society*, **125**, 12096.

19 Ivanisevic, A. and Mirkin, C.A. (2001) *Journal of the American Chemical Society*, **123**, 7887.

20 Salazar, R.B., Shovsky, A., Schonherr, H. and Vancso, G.J. (2006) *Small*, **2**, 1274.

21 Cho, Y. and Ivanisevic, A. (2004) *The Journal of Physical Chemistry B*, **108**, 15223.

22 Demers, L.M., Ginger, D.S., Park, S.J., Li, Z., Chung, S.W. and Mirkin, C.A. (2002) *Science*, **296**, 1836.

23 Jaschke, M. and Butt, H.J. (1995) *Langmuir*, **11**, 1061.

24 Lim, J.H. and Mirkin, C.A. (2002) *Advanced Materials*, **14**, 1474.

25 Yu, M., Nyamjav, D. and Ivanisevic, A. (2005) *Journal of Materials Chemistry*, **15**, 649.

26 Su, M., Liu, X.G., Li, S.Y., Dravid, V.P. and Mirkin, C.A. (2002) *Journal of the American Chemical Society*, **124**, 1560.

27 Fu, L., Liu, X.G., Zhang, Y., Dravid, V.P. and Mirkin, C.A. (2003) *Nano Letters*, **3**, 757.

28 Su, M. and Dravid, V.P. (2002) *Applied Physics Letters*, **80**, 4434.

29 Gundiah, G., John, N.S., Thomas, P.J., Kulkarni, G.U., Rao, C.N.R. and Heun, S. (2004) *Applied Physics Letters*, **84**, 5341.

30 Wang, Y., Zhang, Y., Li, B., Lu, J.H. and Hu, J. (2007) *Applied Physics Letters*, **90**, 133102.

31 Ding, L., Li, Y., Chu, H.B., Li, X.M. and Liu, J. (2005) *The Journal of Physical Chemistry B*, **109**, 22337.

32 Smith, J.C., Lee, K.B., Wang, Q., Finn, M.G., Johnson, J.E., Mrksich, M. and Mirkin, C.A. (2003) *Nano Letters*, **3**, 883.

33 Jung, H., Dalal, C.K., Kuntz, S., Shah, R. and Collier, C.P. (2004) *Nano Letters*, **4**, 2171.

34 John, N.S. and Kulkarni, G.U. (2007) *Journal of Nanoscience and Nanotechnology*, **7**, 977.

35 Su, M., Pan, Z.X., Dravid, V.P. and Thundat, T. (2005) *Langmuir*, **21**, 10902.

36 Jang, J.Y., Schatz, G.C. and Ratner, M.A. (2002) *Journal of Chemical Physics*, **116**, 3875.

37 Weeks, B.L., Vaughn, M.W. and DeYoreo, J.J. (2005) *Langmuir*, **21**, 8096.

38 Weeks, B.L. and DeYoreo, J.J. (2006) *The Journal of Physical Chemistry B*, **110**, 10231.

39 Sheehan, P.E., Whitman, L.J., King, W.P. and Nelson, B.A. (2004) *Applied Physics Letters*, **85**, 1589.

40 Yang, M., Sheehan, P.E., King, W.P. and Whitman, L.J. (2006) *Journal of the American Chemical Society*, **128**, 6774.

41 Huang, L., Chang, Y.H., Kakkassery, J.J. and Mirkin, C.A. (2006) *The Journal of Physical Chemistry B*, **110**, 20756.

42 Nelson, B.A., King, W.P., Laracuente, A.R., Sheehan, P.E. and Whitman, L.J. (2006) *Applied Physics Letters*, 88, 033104.

43 Dong, L.X., Tao, X.Y., Zhang, L., Zhang, X.B. and Nelson, B.J. (2007) *Nano Letters*, **7**, 58.

44 Li, Y., Maynor, B.W. and Liu, J. (2001) *Journal of the American Chemical Society*, **123**, 2105.

45 Cai, Y.G. and Ocko, B.M. (2005) *Journal of the American Chemical Society*, **127**, 16287.

46 Zhang, M., Bullen, D., Chung, S.W., Hong, S., Ryu, K.S., Fan, Z.F., Mirkin, C.A. and Liu, C. (2002) *Nanotechnology*, **13**, 212.

47 Salaita, K., Lee, S.W., Wang, X.F., Huang, L., Dellinger, T.M., Liu, C. and Mirkin, C.A. (2005) *Small*, **1**, 940.

48 Salaita, K., Wang, Y.H., Fragala, J., Vega, R.A., Liu, C. and Mirkin, C.A. (2006) *Angewandte Chemie - International Edition*, **45**, 7220.

49 Lenhert, S., Sun, P., Wang, Y.H., Fuchs, H. and Mirkin, C.A. (2007) *Small*, **3**, 71.

50 Bullen, D., Chung, S.W., Wang, X.F., Zou, J., Mirkin, C.A. and Liu, C. (2004) *Applied Physics Letters*, **84**, 789.

51 Bullen, D. and Liu, C. (2006) *Sensors and Actuators A - Physical*, **125**, 504.

52 Lee, K.B., Lim, J.H. and Mirkin, C.A. (2003) *Journal of the American Chemical Society*, **125**, 5588.

53 Ryu, K.S., Wang, X.F., Shaikh, K., Bullen, D., Goluch, E., Zou, J., Liu, C. and Mirkin, C.A. (2004) *Applied Physics Letters*, **85**, 136.

54 Banerjee, D., Amro, N.A., Disawal, S. and Fragala, J. (2005) *Journal of Microlithography, Microfabrication and Microsystems*, **4**, 230.

55 Li, S.F., Shaikh, K.A., Szegedi, S., Goluch, E. and Liu, C. (2006) *Applied Physics Letters*, 89, 173125.

56 Kim, K.H., Moldovan, N. and Espinosa, H.D. (2005) *Small*, **1**, 632.

57 Moldovan, N., Kim, K.H. and Espinosa, H.D. (2006) *Journal of Micromechanics and Microengineering*, **16**, 1935.

58 Lewis, A., Kheifetz, Y., Shambrodt, E., Radko, A., Khatchatryan, E. and Sukenik, C. (1999) *Applied Physics Letters*, **75**, 2689.

59 Taha, H., Marks, R.S., Gheber, L.A., Rousso, I., Newman, J., Sukenik, C. and Lewis, A. (2003) *Applied Physics Letters*, **83**, 1041.

60 Bruckbauer, A., Ying, L.M., Rothery, A.M., Zhou, D.J., Shevchuk, A.I., Abell, C., Korchev, Y.E. and Klenerman, D. (2002) *Journal of the American Chemical Society*, **124**, 8810.

61 Bruckbauer, A., Zhou, D.J., Ying, L.M., Korchev, Y.E., Abell, C. and Klenerman, D. (2003) *Journal of the American Chemical Society*, **125**, 9834.

62 Suryavanshi, A.P. and Yu, M.F. (2007) *Nanotechnology*, 18, 105305.

63 Hong, S.H., Zhu, J. and Mirkin, C.A. (1999) *Langmuir*, **15**, 7897.

64 Agarwal, G., Sowards, L.A., Naik, R.R. and Stone, M.O. (2003) *Journal of the American Chemical Society*, **125**, 580.

65 Liu, X.G., Zhang, Y., Goswami, D.K., Okasinski, J.S., Salaita, K., Sun, P., Bedzyk, M.J. and Mirkin, C.A. (2005) *Science*, **307**, 1763.

66 Jiang, H.Z. and Stupp, S.I. (2005) *Langmuir*, **21**, 5242.

67 Zheng, Z.K., Yang, M.L., Liu, Y.Q. and Zhang, B.L. (2006) *Nanotechnology*, **17**, 5378.

68 Cho, Y. and Ivanisevic, A. (2006) *Langmuir*, **22**, 8670.

69 Cho, Y. and Ivanisevic, A. (2005) *The Journal of Physical Chemistry B*, **109**, 6225.

70 Vega, R.A., Maspoch, D., Shen, C.K.F., Kakkassery, J.J., Chen, B.J., Lamb, R.A. and Mirkin, C.A. (2006) *Chembiochem: A European Journal of Chemical Biology*, **7**, 1653.

71 Lynch, M., Mosher, C., Huff, J., Nettikadan, S., Johnson, J. and Henderson, E. (2004) *Proteomics*, **4**, 1695.

72 Hyun, J., Ahn, S.J., Lee, W.K., Chilkoti, A. and Zauscher, S. (2002) *Nano Letters*, **2**, 1203.

73 Weinberger, D.A., Hong, S.G., Mirkin, C.A., Wessels, B.W. and Higgins, T.B. (2000) *Advanced Materials*, **12**, 1600.

74 Demers, L.M. and Mirkin, C.A. (2001) *Angewandte Chemie - International Edition*, **40**, 3069.

75 Rozhok, S., Shen, C.K.F., Littler, P.L.H., Fan, Z.F., Liu, C., Mirkin, C.A. and Holz, R.C. (2005) *Small*, **1**, 445.

76 Nyamjav, D. and Ivanisevic, A. (2003) *Advanced Materials*, **15**, 1805.
77 Lee, K.B., Park, S.J., Mirkin, C.A., Smith, J.C. and Mrksich, M. (2002) *Science*, **295**, 1702.
78 Lee, S.W., Oh, B.K., Sanedrin, R.G., Salaita, K., Fujigaya, T. and Mirkin, C.A. (2006) *Advanced Materials*, **18**, 1133.
79 Wang, Y.H., Maspoch, D., Zou, S.L., Schatz, G.C., Smalley, R.E. and Mirkin, C.A. (2006) *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 2026.
80 Zhang, H., Amro, N.A., Disawal, S., Elghanian, R., Shile, R. and Fragala, J. (2007) *Small*, **3**, 81.
81 Zhang, H. and Mirkin, C.A. (2004) *Chemistry of Materials*, **16**, 1480.
82 Ivanisevic, A., McCumber, K.V. and Mirkin, C.A. (2002) *Journal of the American Chemical Society*, **124**, 11997.
83 Wu, S.Y., Berkenbosch, R., Lui, A. and Green, J.B.D. (2006) *Analyst*, **131**, 1213.
84 Coffey, D.C. and Ginger, D.S. (2005) *Journal of the American Chemical Society*, **127**, 4564.
85 Salaita, K., Amarnath, A., Maspoch, D., Higgins, T.B. and Mirkin, C.A. (2005) *Journal of the American Chemical Society*, **127**, 11283.
86 Demers, L.M., Park, S.J., Taton, T.A., Li, Z. and Mirkin, C.A. (2001) *Angewandte Chemie - International Edition*, **40**, 3071.
87 Plutowski, U., Jester, S.S., Lenhert, S., Kappes, M.M. and Richert, C. (2007) *Advanced Materials*, **19**, 1951.
88 Kwak, S.K., Lee, G.S., Ahn, D.J. and Choi, J.W. (2004) *Materials Science and Engineering C - Biomimetic Materials Sensors and Systems*, **24**, 151.
89 Valiokas, R., Vaitekonis, A., Klenkar, G., Trinkunas, G. and Liedberg, B. (2006) *Langmuir*, **22**, 3456.
90 Lee, K.B., Kim, E.Y., Mirkin, C.A. and Wolinsky, S.M. (2004) *Nano Letters*, **4**, 1869.
91 Nam, J.M., Han, S.W., Lee, K.B., Liu, X.G., Ratner, M.A. and Mirkin, C.A. (2004) *Angewandte Chemie - International Edition*, **43**, 1246.
92 Cheung, C.L., Camarero, J.A., Woods, B.W., Lin, T.W., Johnson, J.E. and De Yoreo, J.J. (2003) *Journal of the American Chemical Society*, **125**, 6848.
93 Vega, R.A., Maspoch, D., Salaita, K. and Mirkin, C.A. (2005) *Angewandte Chemie - International Edition*, **44**, 6013.
94 Vega, R.A., Shen, C.K.F., Maspoch, D., Robach, J.G., Lamb, R.A. and Mirkin, C.A. (2007) *Small*, **3**, 1482.

6
Scanning Ion Conductance Microscopy of Cellular and Artificial Membranes

Matthias Böcker, Harald Fuchs, and Tilman E. Schäffer

6.1
Introduction

Eukaryotic cells are enclosed by a plasma membrane, which creates an internal environment that is separated from the outside. The membrane defines a physical border and is impermeable to macromolecules. Integral proteins in the membrane play an essential role for inter- and transcellular processes [1, 2]. In order to understand the properties of membranes and membrane proteins, it is important that cell biology, medicine and pharmacology gain insight into the complex barrier-crossing transport mechanisms. In particular, knowledge concerning the permeability of barriers for substances such as drugs is of great relevance.

Special electrochemical and microscopic methods are used to study the ion-permeability of barrier-forming cell structures. For example, transepithelial electrical resistance (TER)-spectroscopy provides information about the barrier properties of cell layers [3–5]. The development of artificial membranes was an important step in the characterization of membranes [6, 7]. One advantage of artificial membranes is the possibility of inserting selected proteins into the membrane, which in turn creates the possibility of performing single-channel measurements on these selected proteins [8].

For the microscopic characterization of local sample properties, scanning probe microscopes have been developed in various forms. Scanning probe microscopes are based on a small, locally confined probe that is sensitive to various types of physical quantity. To date, several instruments have been developed for the characterization of different sample properties, although only a few are suitable for application in an aqueous environment – an essential requirement when analyzing biological samples under native conditions. Perhaps the most prominent member of the group is the atomic force microscope [9], which allows the creation of high-resolution topographical images of biological samples in buffer solutions [10]. The atomic force microscope employs the mechanical interaction between a sharp tip and the sample surface under investigation on the nanometer scale. In addition to

Nanotechnology. Volume 6: Nanoprobes. Edited by Harald Fuchs

ISBN: 978-3-527-31733-2

topography, it is possible to measure mechanical sample properties such as elasticity [11]. Many analyses have already been conducted on cellular membranes [12, 13], artificial membranes such as solid supported membranes [14, 15] and membrane proteins [16–19].

Unfortunately, the investigation of soft and fragile samples often proves to be problematic due to mechanical interactions between the tip and the sample. In the case of cells, a force-induced deformation reduces the resolution of atomic force microscopy (AFM) imaging, and native conditions are therefore not always reproduced [20]. Additionally, the sample can be damaged irreversibly or be compressed, leading to errors, for example, in the measurement of sample height [21]. In particular, the investigation of free-standing, pore-suspending artificial membranes with AFM has proved difficult, with very few successful measurements on such membranes having been reported [22–25]. All such reports refer to the problem of interaction forces between the atomic force microscope tip and the suspended membrane, which often leads to rupture of the membrane, even at minimal imaging forces.

6.1.1 Scanning Ion Conductance Microscopy

Scanning ion conductance microscopy (SICM), as invented by Hansma *et al*. in 1989 [26, 27], is based on the measurement of an ion current through a small aperture, which is usually formed by a nanopipette. In order to provide a medium for ion conduction, the nanopipette is filled with an electrolyte. A silver/silver chloride (Ag/AgCl) electrode is placed inside the pipette (the pipette electrode), while the sample is placed in a dish that is filled typically with the same electrolyte. A second Ag/AgCl electrode is placed inside the electrolyte in the dish (the bath electrode). By applying a voltage between both electrodes and recording the ion current through the pipette (typically on the nanoampère scale), locally resolved images of sample topography can be generated when scanning the sample. The advantage of the SICM is that no mechanical forces are necessary for imaging a sample surface.

In SICM there is a heavy dependence of the ion current on the distance between the pipette tip and sample surface for generating the feedback signal for scanning. This distance dependence is based on a 'current squeezing effect', which allows the presence of the sample surface to be sensed at a distance where no mechanical interactions between tip and sample occur. In this way, soft and delicate samples such as living cells can be imaged in a 'noncontact' configuration [28, 29]. In addition, single ion channels on living cells can be recorded at specified positions with this technique [30]. Improved imaging techniques have been developed based on the principle of an ac measurement. For example, the tip–sample distance can be modulated, thereby modulating the measured ion current; the amplitude of the modulated current can then be used for feedback control [31–33]. The technique of SICM has proved useful in obtaining high-resolution images of fine surface structures such as microvilli [34, 35] or membrane proteins [36].

For many applications, however, it is important to measure the ion current independently of the sample topography, and for this purpose several different extensions to SICM have been developed. In one revision, the scanning ion conductance microscope was combined with an atomic force microscope [37, 38]. For this, a bent nanopipette [39, 40], coated with a reflective metal layer, was used as the atomic force microscope 'tip' and scanned over the sample surface. A laser beam was then focused onto the bent pipette and the reflected light projected onto a split photo-diode. The mechanical deflection of the pipette provided the feedback signal for topography imaging, while the ion current was recorded simultaneously. As an option, the bent pipette was driven in the tapping-mode [41, 42], which simplified the imaging of soft samples in liquid solutions.

Another modification involved the combination of SICM with scanning near-field optical microscopy (SNOM). This combined microscope allowed living cells to be investigated and topography images with additional optical information to be recorded [43–45]. Yet another extension of SICM was the combination with shear force microscopy [46, 47], a technique that is also used in conjunction with the scanning near-field optical microscope [48, 49]. In SICM with complementary shear force distance control, the pipette is transversally oscillated at a mechanical resonant frequency. The sample topography is then detected by shear forces between the pipette tip and sample surface, causing a reduction in the oscillation amplitude.

6.2 Methods

6.2.1 The Basic Set-Up

In a basic SICM set-up a nanopipette with a small tip opening diameter is positioned close to a sample surface (Figure 6.1a). The nanopipette is filled with an electrolyte and the sample is placed in an electrolyte-filled dish. Typically, the electrolytes in the nanopipette and in the dish are identical, so that no osmotic flow in or out of the pipette occurs. For the current measurement a voltage is applied between two silver/silver chloride (Ag/AgCl) electrodes. Ag/AgCl electrodes have electrochemical properties that make them well suited to applications in SICM. For example, they have a very small equilibrium constant at room temperature so that only a small amount of Ag^+-ions exists in the electrolyte. Additionally, they are easily fabricated, for example by the electrolytic deposition of silver chloride on a silver wire. One of the electrodes is placed inside the pipette (the pipette electrode), while the other electrode is place inside the electrolyte in the dish (the bath electrode). The ion current from the pipette electrode through the pipette, and through its small tip opening to the bath electrode, is measured with a high-impedance current amplifier. The applied voltage is in the range of some hundred millivolts, and the measured ion current is in the nano- and picoampère range. The closer the tip is to the sample surface, the more the

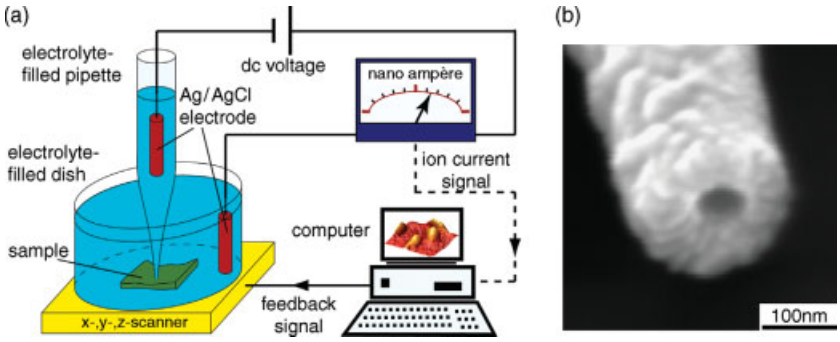


Figure 6.1 (a) Schematic of the scanning ion conductance microscope. A dc voltage between two Ag/AgCl electrodes induces an ion current through the pipette. The ion current signal is passed to a computer, which generates a feedback signal that controls the position of a *z*-scanner. Images are generated by scanning the sample in *x*,*y*-direction; (b) Scanning electron microscopy (SEM) image of the tip of a nanopipette. The inner opening diameter is 60 nm. For SEM imaging the pipette was sputter-coated with a 10 nm-thick aluminum layer.

ion current is restricted ('squeezed') by a narrow gap formed between the pipette tip and the sample surface [26, 50]. The measured ion current is therefore indicative of the tip–sample distance. The ion current signal is passed to a computer, where a feedback signal is generated and used to drive an *x*-, *y*-, *z*-scanner, consisting of piezoelectric actuators (abbreviated as piezos). Depending on the pipette geometry and on the imaging parameters used, topographical imaging without mechanical interaction between tip and sample is possible. Hence, noncontact images of sample topography can be recorded, which is especially valuable for the imaging of soft biological samples such as living cells and artificial membranes.

6.2.2 Nanopipettes

In order to deliver an ion current through a small opening, pipettes proved to be a practical solution. Pipette fabrication is a well-known technique, and pipettes have a wide range of applications in extracellular physiology, for example, in patch–clamp recording [51, 52]. Usually, pipettes are drawn (using a 'pipette puller') from glass capillaries with an initial diameter of 1–2 mm. The principle of a pipette puller is based on local heating of the capillary with a heated coil or a laser beam. While heating, a pulling force is applied to both ends of the capillary such that, when the melting point of the glass is reached, the pulling force leads to a separation of both ends of the capillary, thereby forming fine pipettes. The most commonly used glass is borosilicate; this softens at 825 °C and, as it is pulled, maintains its ratio of inner to outer diameter over the total drawn-out length [53]. A wide variety of opening diameters can be generated with borosilicate glass pipettes, down to some tens of nanometers (Figure 6.1b). Finer pipettes for higher-resolution imaging can be produced from quartz capillaries, with inner tip opening diameters of approximately 12 nm [36].

6.3
Description of Current–Distance Behavior

When placing the electrolyte-filled nanopipette with the pipette electrode into an electrolyte-filled dish with the bath electrode and applying a voltage U between the electrodes, an ion current is measured. When the tip of the pipette is far away from the sample surface the current is maximal (saturation current I_{sat}):

$$I_{sat} = \frac{U}{R_p}. \tag{6.1}$$

For calculating the pipette resistance R_p a simple analytical model can be used [46]. The geometry of the pipette is assumed to be conical (Figure 6.2a). The resistance can be approximated through sectioning the cone into successive disks of infinitesimal thicknesses, leading to

$$R_p = \frac{1}{\kappa} \cdot \frac{L_p}{\pi r_p r_i} \tag{6.2}$$

where κ is the specific conductivity of the electrolyte, L_p is the length of the drawn-out end of the pipette, r_p is the inner radius of the capillary, and r_i is the inner diameter of the tip opening. For typical borosilicate glass pipettes and electrolytes, R_p is in the range of 50–200 MΩ. The resistance of the electrolyte bath inside the dish is on the order of kilo-ohms, so that the total resistance of the circuit can be approximated by R_p. With this, the inner tip opening diameter can be estimated as

$$r_i \approx \frac{I_{sat}}{U} \cdot \frac{L_P}{\pi \kappa r_P}. \tag{6.3}$$

For standard 1 mm outer diameter borosilicate glass pipettes, for example, $r_p = 0.29$ mm. In many cases, a phosphate-buffered saline (PBS) solution with 137 mM NaCl ($\kappa \approx 1.3$ Ωm) is used as electrolyte. L_P is typically in the range of 5 mm, and $U = 200$ mV. For example, from a measured current $I_{sat} = 1$ nA, an inner tip opening radius of $r_i \approx 21$ nm is deduced.

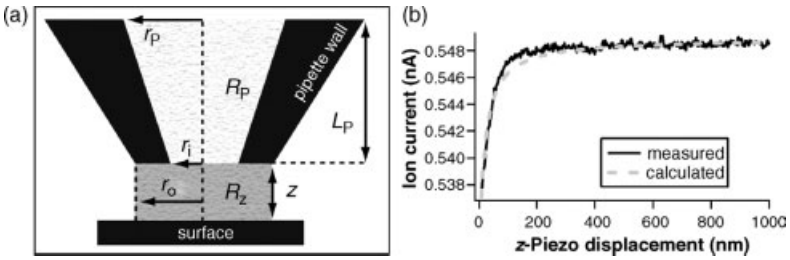


Figure 6.2 (a) Schematic of a simple model for the current–distance behavior of a cone-shaped pipette; (b) Measured and calculated ion current versus z-piezo displacement. Measurement were made with a borosilicate glass pipette over a mica surface, applying a voltage of $U = 200$ mV.

When the tip is at a distance to the sample surface that is comparable to the tip opening diameter, the measured ion current shows a heavy dependence on the tip–sample distance. This is because the current then has to 'squeeze' through the narrow gap between the pipette wall and the sample surface, which leads to an increase in the resistance. Nitz *et al.* [46] constructed an analytical model for the distance-dependent resistance R_z, obtaining

$$R_z(z) \approx \frac{3}{2\pi} \cdot \frac{\ln(r_o/r_i)}{\kappa z}, \tag{6.4}$$

where r_o is the outer tip diameter of the pipette and z is the tip–sample distance. This leads to a distance-dependent ion current of

$$I(z) \approx I_{sat}\left(1 + \frac{z_0}{z}\right)^{-1}, \tag{6.5}$$

with

$$z_0 = \frac{3}{2} \cdot \frac{r_p r_i \ln(r_o/r_i)}{L_p}. \tag{6.6}$$

The ratio of r_i to r_o usually stays constant while pulling the pipette from the borosilicate capillary [53]. In Figure 6.2b, a measurement of ion current versus z-piezo extension (the tip–sample distance with an unknown offset) is displayed (black trace). The range of the z-piezo extension is 1 µm. At a large tip–sample distance (right-hand side in Figure 6.2b), the ion current is independent of distance, yielding a saturation current of $I_{sat} = 0.549$ nA. By using Equation 6.3, r_i can be approximated as 12 nm. The calculated ion current $I(z)$ obtained from the analytical model (Equation 6.5) is also displayed (gray, dashed trace); a horizontal offset was applied here to optimize the match. The analytical model matches the measured data well. From the offsetting procedure, a tip–sample distance of 18 nm is found at the leftmost position in the graph (at a z-piezo displacement of 0 nm).

6.4 Imaging with SICM

6.4.1 Modulated Scan Technique

The dependence of the measured ion current on tip–sample distance can be utilized in a feedback loop to keep the tip–surface distance constant while scanning, thus allowing topographical images of a sample surface to be recorded. In practice, however, this imaging mode has proved vulnerable to current drift. One method of reducing the influence of current drift is to apply short voltage or current pulses [54, 55] instead of a constant voltage between the pipette and bath electrode.

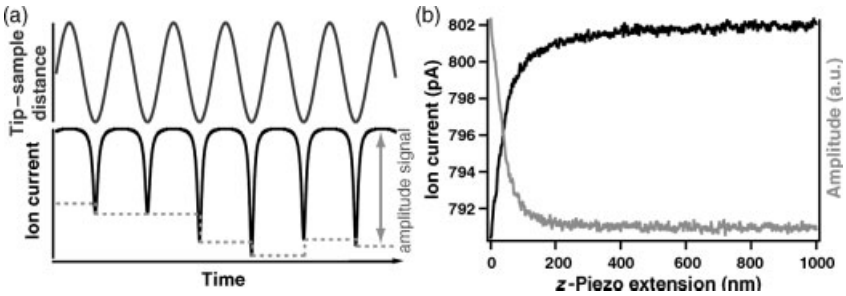


Figure 6.3 (a) Schematic diagram to illustrate the principle of the modulated scan technique. Upper trace: sinusoidal modulation of the tip–sample distance. Lower trace: the resulting ion current. The current decreases periodically at the points of closest approach between pipette tip and sample. The value of the maximal current drop provides the amplitude signal which is used for feedback; (b) Measured ion current (black trace) and ion current amplitude (gray trace) versus *z*-piezo extension while modulating the tip–sample distance with an amplitude of 120 nm at a frequency of 800 Hz.

Another method is based on modulating the tip–sample distance [31–33]. This distance is modulated sinusoidally with an amplitude of up to some hundred nanometers and with a frequency of a several hundred Hertz (Figure 6.3a, upper graph). For large tip–sample distances, this modulation does not have any influence on the ion current. For small tip–sample distances, the behavior of the ion current is displayed in Figure 6.3a (lower graph). At distances furthest from the surface, the ion current reaches the saturation current (plateaus in the graph). When approaching the sample surface during the modulation cycle, the influence of the squeezing effect becomes greater so that the current decreases. At the point of closest approach the current is smallest, but increases again when the tip retracts from the surface. An amplitude signal can be generated either by using a lock-in amplifier or a trigger-based method [47]. The amplitude of the current signal is used as input to a feedback loop, which regulates the mean tip–sample distance to keep the amplitude signal constant.

Measurement of the ion current (black, left axis) and of the amplitude signal (gray, right axis) versus *z*-piezo extension is shown in Figure 6.3b. The ion current shows the distance-dependent behavior (as discussed above), while the amplitude signal is constant for large tip–sample distances (right-hand side in Figure 6.3b) and increases for decreasing tip–sample distances. The amplitude signal is less sensitive to current drift than the ion current, and may also help minimizing the risk of lateral forces being applied to the sample.

6.4.2 Cellular Membranes

Use of the modulated scan technique allows the gentle imaging of delicate biological samples such as living cells, and also allows the recording of well-resolved images of fine surface structures such as microvilli [34] or membrane proteins [36]. Living,

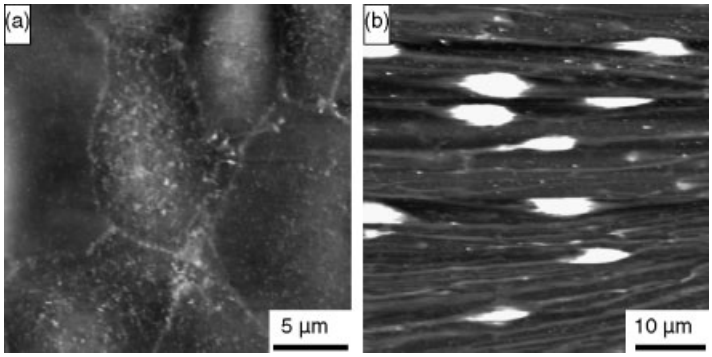


Figure 6.4 (a) Scanning ion conductance microscopy (SCIM) topographic image of living MDCK-II cells in phosphate-buffered saline (PBS) solution. The gray scale ranges from black to white over 1.8 µm; (b) SICM topographic image of living rat Schwann cells in PBS solution. The grayscale ranges from black to white over 5 µm.

confluent MDCK-II cells, for example, exhibit a large cell body and locally raised cell–cell contacts (Figure 6.4a); substructures on the cell membrane are also visible. Another example is the imaging of living Schwann cells (Figure 6.4b); these occur in the peripheral nervous system and are essential for generating the myelin sheath and maintaining the metabolism and integrity of the nerve. The SICM image shows the typical spindle-shaped body of the Schwann cell. On the basis of the excellent long-term stability provided by modulated scanning techniques it is possible to image specific surface areas, continuously, for several hours and thus to observe dynamic processes occurring on the cell membranes [35, 56]. As an example, Gorelik *et al.* were able to study the mechanism by which aldosterone activates sodium reabsorption via the epithelial sodium channel [57].

6.4.3
Artificial Membranes

The study of artificial membranes represents a new application area for SICM. The investigation of pore-suspending membranes with other scanning probe techniques such as AFM has proved difficult due to mechanical interactions that can cause damage to the membrane. In contrast, the noncontact imaging characteristic of SICM allows the mapping of such pore-suspending membranes without mechanical interaction between the pipette tip and the sample surface.

In a study of black lipid membranes (BLMs), highly ordered porous silicon proved to be a useful substrate for spanning the membrane over the pores, although for this purpose the substrate must first be functionalized. In the Müller–Rudin technique [6, 7], the membrane is applied in a solvent locally to the substrate; this causes it to spread over the surface, delivering a membrane monolayer (the spreading process is shown in Figure 6.5a). The scan was started directly after applying the membrane to the surface (scan direction: downwards). Initially, all of the pores were

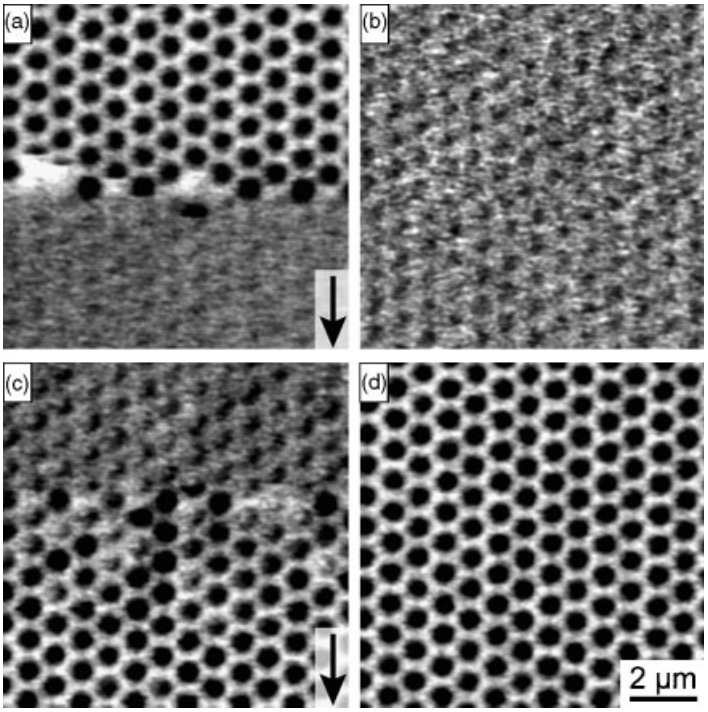


Figure 6.5 Scanning ion conductance microscopy topographic images of a highly ordered porous silicon substrate at different stages of coverage with a lipid membrane. (a) Topography imaged directly after applying the membrane solution to the silicon substrate (scan direction: downwards). In the lower section of the image, the membrane is already suspended over the pores; (b) After spreading over the surface the membrane is suspended over all pores and was stable for hours; (c) Applying a small amount of detergent (Tween 20) to the PBS solution destroyed the membrane (scan direction: downwards). In the upper section of the image, the membrane was still present, while in the lower section it was already destroyed; (d) Finally, the porous silicon substrate was totally free of membrane again. The scan rate was 5 s per line with a resolution of 256×256 data points. The grayscale ranges from black to white over 120 nm.

open (Figure 6.5a, upper half), but when reaching the vertical center of the image the first suspended pores became visible, with only suspended pores being observed in the lower section of the image. This effect was due to the solvent reaching the scan area while scanning, thus suspending the pores with membrane. The same area, after re-scanning, is shown in Figure 6.5b. Here, all the pores are suspended with membrane, although the porous structure of the substrate can still be recognized by the small depressions in the membrane surface. This state remained stable over several hours, after which time a droplet of a membrane-dissolving detergent was applied to the electrolyte (Figure 6.5c). Initially, the membrane remained unaffected (upper section), but after some minutes some pores began to re-open (middle section), and a few minutes later only open pores remained (lower section). A re-scan of the area showed only open pores (Figure 6.5d). Taken together, these measurements demonstrated the gentle imaging character of SICM.

6.4.4 SICM with Shear Force Distance Control

The fact that SICM can be used for imaging biological samples, with a resolution in the nanometer range, makes it an interesting candidate for combination with other scanning microscopy techniques. Of particular interest is the measurement of ion current independent of sample topography, and for this purpose the nanopipette can be used as a shear force sensor. Shear forces are well known from SNOM set-ups, where they serve to keep the optical fiber at a constant distance from the sample during scanning [48, 49]. In a shear force configuration, the probe is vertically oriented with respect to the sample surface, while a dither piezo excites a transverse mechanical vibration of the probe. The amplitude of the vibration depends heavily on the tip–sample distance: at small distances, the shear forces between the tip and sample reduce the vibration amplitude, which therefore can be used as a measure of tip–sample distance. To date, several methods have been established for detecting the vibration amplitude, including optical readout [48] and the use of a piezoelectric tuning fork sensor [58]. Although the latter method faced problems when adapted for use in liquids (due to electrical short circuits), several solutions were described, including a custom piezoelectric detection design [59], an electrically insulating layer [60] and a diving bell concept [61].

In a combined shear force and scanning ion conductance microscope, the nanopipette acts both as probe for the ion current and as probe for the shear force measurement. For the detection of vibration amplitude an optical readout was used [46], as this system functions equally well in liquid and in air [62]. An improved optical detection design was based on the use of a pair of periscopes (Figure 6.6) [47]. Here, a collimated laser beam from a laser diode passes down through one periscope tube. Inside the tube, the beam is reflected by a mirror and passes through a lens that provides the interface to the liquid and focuses the beam onto the thin end of the pipette. In the second periscope tube, a two-segment photo-diode detects the light scattered by the pipette, resulting in a vibration amplitude signal. With this periscope-based detection system the vibration amplitude can be detected close to the pipette tip, where the sensitivity is highest. Furthermore, the second tube can be positioned to detect either the transmitted or the reflected light from the pipette. The optical reflectivity of the pipette can be increased by coating with a thin metal layer. In order to induce pipette vibrations a dither piezo is used. By using the vibration amplitude as input to the feedback loop controlling the tip–sample distance, the sample topography can be imaged. The simultaneously recorded ion current then yields a complementary image of ion current at a constant tip–sample distance.

The combined scanning ion conductance and shear force microscope can be used for the investigation of local variations in ion conductance of biological specimens. This is of special interest for research in the field of the barrier-forming structures such as endothelial or epithelial cell layers. In multicellular organisms, these structures form the interface between different fluid compartments and play an important role in inter- and transcellular processes [1, 2]. Gaining insight into the complex barrier-crossing transport mechanisms is a common interest of cell biology,

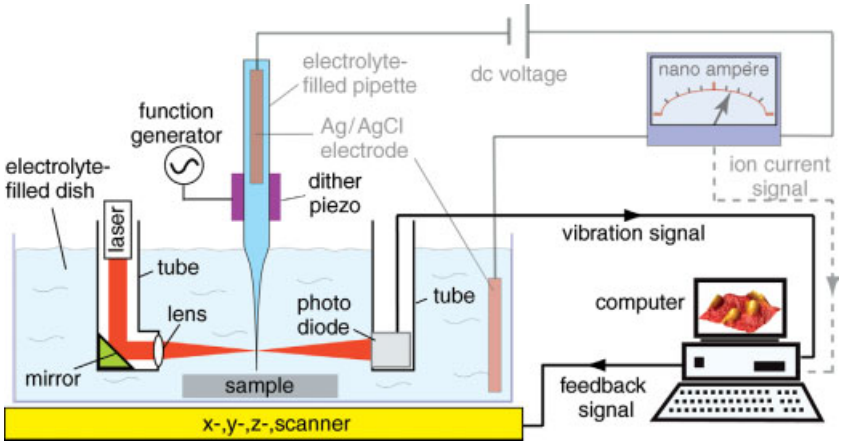


Figure 6.6 (a) Schematic of the combined ion current and shear force measurement set-up. The components for the ion current measurement are grayed out. The pipette is vibrated by a dither piezo parallel to the surface. A periscope design based on two tubes is used to measure the vibration amplitude optically. The first tube is used to focus a laser beam onto the pipette tip; the second tube contains a split photo-diode, which detects the modulated beam and thereby the vibration amplitude of the pipette. The vibration signal provides a complementary input to the computer.

medicine and pharmacology, as malfunctioning of these barriers may have pathological implications.

Special electrochemical and microscopic methods are required to study the ion-permeability of barrier-forming cell structures. For example, experimental techniques such as the measurement of TER may provide valuable information about the barrier properties of cell layers [63, 64]. In addition to such integrating measurements of the total cell layer impedance, the combined scanning ion conductance and shear force microscope can provide further insight into cellular transport mechanisms, as it allows the simultaneous recording of topographic data and local ion conductance with high lateral resolution. One such example is the investigation of the functionality of tight junctions between living MDCK-II cells (Figure 6.7). Here, the shear force image shows the topography of the cells, while the ion current image reveals lines of increased conductance at the position of the cell–cell contacts.

6.5 Outlook

The gentle character of SICM provides for a broad range of possible applications. Examples include the possible study of proteins in pore-suspending membranes, as well as combinations with other imaging techniques. By using the ion current signal of SICM to maintain a constant tip–sample distance, complementary signals such as

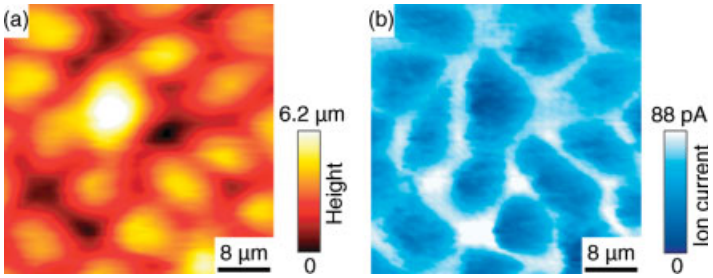


Figure 6.7 (a) Shear force topographic image of living MDCK-II cells; (b) Simultaneously recorded ion current image. The color shows the variation of the measured ion current while scanning.

local optical intensity can be recorded simultaneously. It has been shown that the end of a tapered nanopipette can serve as a near-field light source for SNOM [32, 43, 65], this being achieved by coupling a laser light into the nanopipette via an optical fiber. Coating the outside of the nanopipette with a reflective metal layer then helped to confine the laser light to the aperture (i.e. the tapered end of the nanopipette). Provided that the sample and substrate were transparent, the SNOM signal could then be collected through an objective and detected using a photomultiplier located beneath the SICM set-up. In this way, living cells have been successfully investigated using a combined SICM/SNOM set-up. An alternative use of the SICM probe as a confined light source for SNOM was suggested by Bruckbauer *et al.* [44, 45]. This method was based on the fluorescence that occurred when a calcium indicator (with which the nanopipette is filled) binds with calcium in the sample solution and is illuminated with laser light. The mixing zone where the fluorescent complex forms serves as a localized light source. Another viable combination was that of scanning confocal microscopy (SCM) [66], where the set-up comprised an inverted light microscope fully configured for SCM, on which the SICM set-up was placed. During lateral scanning the vertical position of the sample was controlled by SICM and, as a result, the optical confocal volume, which was located just beneath the end of the nanopipette, followed the topography of the sample. This allowed fluorescence images of a surface to be recorded simultaneously with topographic data.

Acknowledgments

The authors thank Boris Anczykowski, Yuri Korchev, Andrew Shevchuk, Roger Proksch, Eva Schmitt and Claudia Steinem for their stimulating discussions and support. The Schwann cells were a gift from Ilka Kleffner and Peter Young of the University Clinic Münster, while the MDCK-II cells were kindly provided by Joachim Wegener of the Institute of Biochemistry at the University of Münster. The authors are grateful to Asylum Research for support. In addition, the DFG is gratefully acknowledged for financial support (SCHA 1264/1 and STE 884/5). T.E.S. thanks the Gemeinnützige Hertie-Stiftung/Stifterverband für die Deutsche Wissenschaft for support.

References

1 Powell, D.W. (1981) *The American Journal of Physiology*, **241**, G275.
2 Simionescu, M. and Simionescu, N. (1986) *Annual Review of Physiology*, **48**, 279.
3 Cereijido, M., Gonzalez-Mariscal, L., Contreras, R.G., Gallardo, J.M., Garcia-Villegas, R. and Valdes, J. (1993) *Journal of Cell Science - Supplement*, **17**, 127.
4 Diamond, J.M. (1977) *The Physiologist*, **20**, 10.
5 Wegener, J., Sieber, M. and Galla, H.J. (1996) *Journal of Biochemical and Biophysical Methods*, **32**, 151.
6 Müller, P., Rudin, H.T., Tien, H.T. and Wescott, W.C. (1963) *The Journal of Physical Chemistry B*, **67**, 534.
7 Montal, M. and Müller, P. (1972) *Proceedings of the National Academy of Sciences of the United States of America*, **69**, 3561.
8 Schmitt, E.K., Vrouenraets, M. and Steinem, C. (2006) *Biophysical Journal*, **91**, 2163.
9 Binnig, G., Quate, C.F. and Gerber, C. (1986) *Physical Review Letters*, **56**, 930.
10 Drake, B., Prater, C.B., Weisenhorn, A.L., Gould, S.A., Albrecht, T.R., Quate, C.F., Cannell, D.S., Hansma, H.G. and Hansma, P.K. (1989) *Science*, **243**, 1586.
11 Radmacher, M., Fritz, M., Kacher, C.M., Cleveland, J.P. and Hansma, P.K. (1996) *Biophysical Journal*, **70**, 556.
12 Butt, H.J., Wolff, E.K., Gould, S.A., Dixon Northern, B., Peterson, C.M. and Hansma, P.K. (1990) *Journal of Structural Biology*, **105**, 54.
13 Yamashina, S. and Katsumata, O. (2000) *Journal of Electron Microscopy*, **49**, 445.
14 Mou, J., Yang, J. and Shao, Z. (1994) *Biochemistry*, **33**, 4439.
15 Hui, S.W., Viswanathan, R., Zasadzinski, J.A. and Israelachvili, J.N. (1995) *Biophysical Journal*, **68**, 171.
16 Butt, H.J., Downing, K.H. and Hansma, P.K. (1990) *Biophysical Journal*, **58**, 1473.
17 Hoh, J.H., Sosinsky, G.E., Revel, J.P. and Hansma, P.K. (1993) *Biophysical Journal*, **65**, 149.
18 Janshoff, A., Ross, M., Gerke, V. and Steinem, C. (2001) *ChemBioChem*, **2**, 587.
19 Mueller, H., Butt, H.-J. and Bamberg, E. (2000) *The Journal of Physical Chemistry B*, **104**, 4552.
20 Hansma, H.G. and Hoh, J.H. (1994) *Annual Review of Biophysics and Biomolecular Structure*, **23**, 115.
21 Jiao, Y. and Schäffer, T.E. (2004) *Langmuir*, **20**, 10038.
22 Hennesthal, C. and Steinem, C. (2000) *Journal of the American Chemical Society*, **122**, 8085.
23 Hennesthal, C., Drexler, J. and Steinem, C. (2002) *ChemPhysChem*, **3**, 885.
24 Goncalves, R.P., Agnus, G., Sens, P., Houssin, C., Bartenlian, B. and Scheuring, S. (2006) *Nature Methods*, **3**, 1007.
25 Ovalle-Garcia, E. and Ortega-Blake, I. (2007) *Applied Physics Letters*, **91**, 093901.
26 Hansma, P.K., Drake, B., Marti, O., Gould, S.A. and Prater, C.B. (1989) *Science*, **243**, 641.
27 Prater, C.B., Drake, B., Gould, S.A.C., Hansma, H.G. and Hansma, P.K. (1990) *Scanning*, **12**, 50.
28 Korchev, Y.E., Bashford, C.L., Milovanovic, M., Vodyanoy, I. and Lab, M.J. (1997) *Biophysical Journal*, **73**, 653.
29 Korchev, Y.E., Milovanovic, M., Bashford, C.L., Bennett, D.C., Sviderskaya, E.V., Vodyanoy, I. and Lab, M.J. (1997) *Journal of Microscopy*, **188**, 17.
30 Korchev, Y.E., Negulyaev, Y.A., Edwards, C.R., Vodyanoy, I. and Lab, M.J. (2000) *Nature Cell Biology*, **2**, 616.
31 Pastre, D., Iwamoto, H., Liu, J., Szabo, G. and Shao, Z. (2001) *Ultramicroscopy*, **90**, 13.
32 Mannelquist, A., Iwamoto, H., Szabo, G. and Shao, Z.F. (2001) *Applied Physics Letters*, **78**, 2076.
33 Shevchuk, A.I., Gorelik, J., Harding, S.E., Lab, M.J., Klenerman, D. and Korchev, Y.E. (2001) *Biophysical Journal*, **81**, 1759.
34 Gorelik, J., Gu, Y., Spohr, H.A., Shevchuk, A.I., Lab, M.J., Harding, S.E., Edwards, C.R., Whitaker, M., Moss, G.W., Benton,

D.C., Sanchez, D., Darszon, A., Vodyanoy, I., Klenerman, D. and Korchev, Y.E. (2002) *Biophysical Journal*, **83**, 3296.

35 Gorelik, J., Shevchuk, A.I., Frolenkov, G.I., Diakonov, I.A., Lab, M.J., Kros, C.J., Richardson, G.P., Vodyanoy, I., Edwards, C.R., Klenerman, D. and Korchev, Y.E. (2003) *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 5819.

36 Shevchuk, A.I., Frolenkov, G.I., Sanchez, D., James, P.S., Freedman, N., Lab, M.J., Jones, R., Klenerman, D. and Korchev, Y.E. (2006) *Angewandte Chemie - International Edition*, **45**, 2212.

37 Proksch, R., Lal, R., Hansma, P.K., Morse, D. and Stucky, G. (1996) *Biophysical Journal*, **71**, 2155.

38 Schäffer, T.E., IonescuZanetti, C., Proksch, R., Fritz, M., Walters, D.A., Almqvist, N., Zaremba, C.M., Belcher, A.M., Smith, B.L., Stucky, G.D., Morse, D.E. and Hansma, P.K. (1997) *Chemistry of Materials*, **9**, 1731.

39 Shalom, S., Lieberman, K., Lewis, A. and Cohen, S.R. (1992) *Review of Scientific Instruments*, **63**, 4061.

40 Lewis, A., Taha, H., Strinkovski, A., Manevitch, A., Khatchatouriants, A., Dekhter, R. and Ammann, E. (2003) *Nature Biotechnology*, **21**, 1377.

41 Hansma, P.K., Cleveland, J.P., Radmacher, M., Walters, D.A., Hillner, P.E., Bezanilla, M., Fritz, M., Vie, D., Hansma, H.G., Prater, C.B., Massie, J., Fukunaga, L., Gurley, J. and Elings, V. (1994) *Applied Physics Letters*, **64**, 1738.

42 Putman, C.A.J., Werf, K.O.V.d., Grooth, B.G.D., Hulst, N.F.V. and Greve, J. (1994) *Applied Physics Letters*, **64**, 2454.

43 Korchev, Y.E., Raval, M., Lab, M.J., Gorelik, J., Edwards, C.R., Rayment, T. and Klenerman, D. (2000) *Biophysical Journal*, **78**, 2675.

44 Bruckbauer, A., Ying, L.M., Rothery, A.M., Korchev, Y.E. and Klenerman, D. (2002) *Analytical Chemistry*, **74**, 2612.

45 Rothery, A.M., Gorelik, J., Bruckbauer, A., Yu, W., Korchev, Y.E. and Klenerman, D. (2003) *Journal of Microscopy*, **209**, 94.

46 Nitz, H., Kamp, J. and Fuchs, H. (1998) *Probe Microscopy*, **1**, 187.

47 Böcker, M., Anczykowski, B., Wegener, J. and Schäffer, T.E. (2007) *Nanotechnology*, **18**, 145505.

48 Betzig, E., Finn, P.L. and Weiner, J.S. (1992) *Applied Physics Letters*, **60**, 2484–2486.

49 Toledo-Crow, R., Yang, P.C., Chen, Y. and Vaez-Iravani, M. (1992) *Applied Physics Letters*, **60**, 2957.

50 Bard, A.J., Denuault, G., Lee, C., Mandler, D. and Wipf, D.O. (1990) *Accounts of Chemical Research*, **23**, 357.

51 Hille, B. (1992) *Ionic Channels of Excitable Membranes*, 2nd edn, Sinauer Associates, Sunderland, Mass.

52 Sakmann, B. and Neher, E. (1995) *Single-Channel Recording*, 2nd edn, Springer, Heidelberg.

53 Brown, K.T. and Flaming, D.G. (1986) *Advanced Micropipette Techniques for Cell Physiology*, John Wiley & Sons, New York.

54 Happel, P., Hoffmann, G., Mann, S.A. and Dietzel, I.D. (2003) *Journal of Microscopy*, **212**, 144.

55 Mann, S.A., Hoffmann, G., Hengstenberg, A., Schuhmann, W. and Dietzel, I.D. (2002) *Journal of Neuroscience Methods*, **116**, 113.

56 Gorelik, J., Zhang, A., Shevchuk, A., Frolenkov, G.I., Sanchez, D., Lab, M.J., Vodyanoy, I., W, E.C.R., Klenerman, D. and Korchev, Y.E. (2002) *Molecular and Cellular Endocrinology*, **217**, 101.

57 Gorelik, J., Zhang, Y., Sanchez, D., Shevchuk, A., Frolenkov, G., Lab, M., Klenerman, D., Edwards, C. and Korchev, Y. (2005) *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 15000.

58 Karrai, K. and Grober, R.D. (1995) *Applied Physics Letters*, **66**, 1842.

59 Brunner, R., Hering, O., Marti, O. and Hollricher, O. (1997) *Applied Physics Letters*, **71**, 3628.

60 Rensen, W.H.J., van Hulst, N.F. and Kammer, S.B. (2000) *Applied Physics Letters*, **77**, 1557.

61 Koopman, M., de Bakker, B.I., Garcia-Parajo, M.F. and van Hulst, N.F. (2003) *Applied Physics Letters*, **83**, 5083.

62 Lambelet, P., Pfeffer, M., Sayah, A. and Marquis-Weible, F. (1998) *Ultramicroscopy*, **71**, 117.

63 Wegener, J., Abrams, D., Willenbrink, W., Galla, H.J. and Janshoff, A. (2004) *Biotechniques*, **37**, 590.

64 Wegener, J., Zink, S., Rösen, P. and Galla, H.J. (1999) *European Journal of Physiology*, **437**, 925.

65 Mannelquist, A., Iwamoto, H., Szabo, G. and Shao, Z. (2002) *Journal of Microscopy*, **205**, 53.

66 Gorelik, J., Shevchuk, A., Ramalho, M., Elliott, M., Lei, C., Higgins, C.F., Lab, M.J., Klenerman, D., Krauzewicz, N. and Korchev, Y. (2002) *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 16018.

7
Nanoanalysis by Atom Probe Tomography

Guido Schmitz

7.1
Introduction

The emergence of nanotechnology is closely related to progress in microscopy. Certainly, the existence of atoms as indivisible units of matter may be postulated, as did Greek philosophers several centuries BC (Demokritos 460-371). Also the important properties of small clusters of these elementary units may be deduced by a series of clever physical experiments and suitable reasoning. But it is almost impossible to master the technology of nanostructured devices and to establish their mass production without being able to monitor – to 'see' – the real structure of the fabricated devices. Thus, any important step in reducing the size scale of a technology will require the development of a suitable means of microscopy. Since the structural width of modern devices scales down to only a few nanometers, much interest persists in imaging techniques of atomic resolution. Furthermore, to go beyond mere imaging, chemical analysis of the structure with atom-by-atom accuracy and, perhaps, in three-dimensional (3-D) spatial resolution in desired.

With transmission electron microscopy (TEM), scanning probe microscopy (SPM) and field ion microscopy (FIM), three major branches of microscopy are currently able to achieve atomic resolution in everyday practical studies. Interestingly, the first demonstration of imaging individual atoms was achieved in 1951 [1], using FIM. Since that time, whilst both TEM and SPM have found widespread applications, FIM techniques have remained in a tiny niche, with very few laboratories being able to master and develop the related methods. This is all the more surprising as the fundamental process of FIM – the field evaporation of atoms – offers the exciting possibility to perform a chemical analysis simply by counting the individual atoms.

Instruments which use this approach are referred to as atom probes (APs), and have been used for about 40 years. Unfortunately, their detector and computing possibilities remained rather limited until the 1980s, with the technique being reserved to specialized laboratories. However, based on the recent substantial progress in instrumentation this situation has changed significantly and, indeed,

Nanotechnology. Volume 6: Nanoprobes. Edited by Harald Fuchs

ISBN: 978-3-527-31733-2

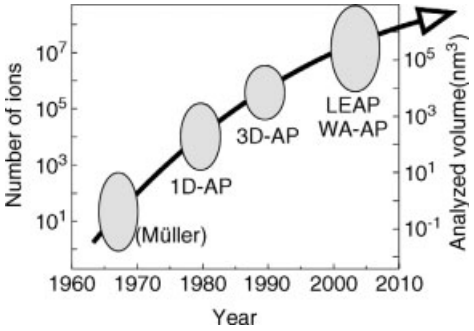


Figure 7.1 Evolution of the atom probe method in terms of analyzed number of atoms or total volume per measurement. For an explanation of the abbreviations, see Section 7.2.

is about to undergo further dramatic development. Depending on their standpoint, some research groups have claimed this to be a 'revolution' [2], while even those less euphoric scientists have admitted 'quite remarkable progress' over the past 20 years. By exploiting the achievements of fast electronics and modern computing, a real 3-D atomic reconstruction of the analyzed volume has become possible, with the maximum number of identified atoms being pushed beyond the hundred million benchmark (see Figure 7.1). Today, atom probe tomography (APT) is capable of providing the chemical analysis of a functional nanodevice as a complete unit, with atomic accuracy.

A second branch of innovation has emerged with the development of efficient pulsed laser sources. By assisting in the process of field evaporation by short laser pulses (of down to a few hundred femtoseconds duration), the practical analysis of nonconductive materials has become a realistic perspective. Recent successful measurements of former difficult classes of materials, such as semi-conductors and ceramics, are encouraging and have initiated intensive methodical research. The concept of laser assistance dates back almost 30 years when the technical limitations of the lasers available prevented their widespread use. As a consequence, the idea was buried among many other interests and, only very recently, has experienced a renaissance.

In the following sections we will examine atom probe tomography in greater detail which, with its increasingly widespread application, can no longer be neglected as an important branch of analytical microscopy. As the typical reader will most likely have only a basic knowledge of the subject, we will first describe the functional principles of the method, the basic algorithms of the 3-D reconstruction, and its accuracy. The instrumental technique and practical specimen preparation will then be considered. Applications of the method will be illustrated by some case studies which address actual problems of thin-film nanoscience. The final section is devoted to the most recent trends in atom probe tomography, including the use of pulsed laser sources to extend the method to complex materials, semiconductors, silicides and oxides as are found in microelectronic devices. It should be noted that the selected

examples were biased by the author's personal interests, and are not aimed to provide a complete overview of the extensive studies conducted in recent years. Those readers seeking additional information should consult recent reviews [2, 3] and textbooks and other reports on the atom probe method [4, 5] and FIM [6].

7.2 Historical Development

Atom probe tomography applies the principle of FIM, and represents the latest progress in this area. The method dates back to the pioneering studies of E. W. Müller, who invented the field ion microscope in 1951 [1] after years of experimentation with electron emission microscopy for which the resolution was limited by the finite thermal energy of the electrons. With FIM, Müller was able to demonstrate atomic resolution images of tungsten surfaces as long ago as 1957 [7]. Although capable of achieving image magnifications in the range of one million – and thereby of atomic resolution – a field ion microscope is a surprisingly simple instrument, when compared to the complex electron optics of electron microscopes. Usually, no imaging lens is needed here. Owing to its simple projective geometry, the instrument does not suffer from the problems of stability associated with electron microscopy.

In 1965, Müller and colleagues were the first to combine FIM with time-of-flight (ToF) mass spectrometry, thus creating the so-called one-dimensional atom probe (1D-AP), the first tool to be used for quantitative chemical analysis in the nanometer range [8]. Rapid progress in detector technology during the 1980s led to the creation of single-ion detectors with sufficient spatial resolution and high detection rates. Important milestones here were the introduction of microchannel plates, of CCD cameras, and of rapid charge-to-digital converters which allowed picosecond time measurements. With this equipment, the early atom probe of Müller experienced a remarkable improvement and, by combining chemical identification by ToF mass spectrometry with the spatial information of the atom position, the numerical 3-D reconstruction of the spatial arrangement of the atomic species became possible. The atom probe had truly advanced to become a modern tool of real 3-D analysis!

Meanwhile, a variety of instrumental concepts of three-dimensional atom probes (3D-APs) were designed and put into operation. The first effectively functioning instrument was the 'position-sensitive atom probe' (PoSAP), which was described in 1986 by Cerezo and Smith [9]. A second, improved, instrument which could handle data from multiple atoms in parallel, the tomographic atom probe (TAP), was presented later (in 1993) by Blavette and coworkers [10]. In fact, as this instrument became more popular and was used in many laboratories worldwide, it lent its name to the general term for the method, namely atom probe tomography (APT). Today, the technical development of APT is in a continual state of flux, with new instruments being introduced on a regular basis. The field of view, and in turn the size of the investigated volume, was significantly improved essentially by reducing the specimen–detector separation, which led to the wide- angle tomographic atom probe (WATAP). Likewise, the addition of a laser beam line led to the method of laser-assisted

field evaporation (LATAP) while, more recently, different energy focusing devices have been developed in order to improve the mass resolution [11, 12].

One remarkable achievement here has been that of the local electrode atom probe (LEAP) [13]. By placing a micrometer-sized electrode in front of an array of microtips, this instrument moderates the serious restrictions in specimen geometry, namely that a needle of high aspect ratio is needed. At the same time, the total number of atoms analyzed per measurement – and thus the size of the reconstructed volume – is increased by one to two orders of magnitude.

7.3 The Physical Principles of the Method

7.3.1 Field Ionization and Evaporation

All FIM techniques utilize the fact that electrical fields are concentrated at tips of sharp curvature. By supplying moderate voltages, enormous field strengths in the range of some $10\,V\,nm^{-1}$ are easily obtained at the apex of nanometer-sized tips, whereas fields of such magnitude could be never obtained with macroscopic geometries. Thus, a typical field ion microscope consists of an ultra-high-vacuum (UHV) chamber with a specimen stage which holds the sample tip, a high-voltage supply and a viewing screen with the capability of imaging the ion impacts (Figure 7.2). A positive potential is supplied to the metallic specimen, while the entrance face of the screen is kept at ground. In order to reduce thermal energies, a

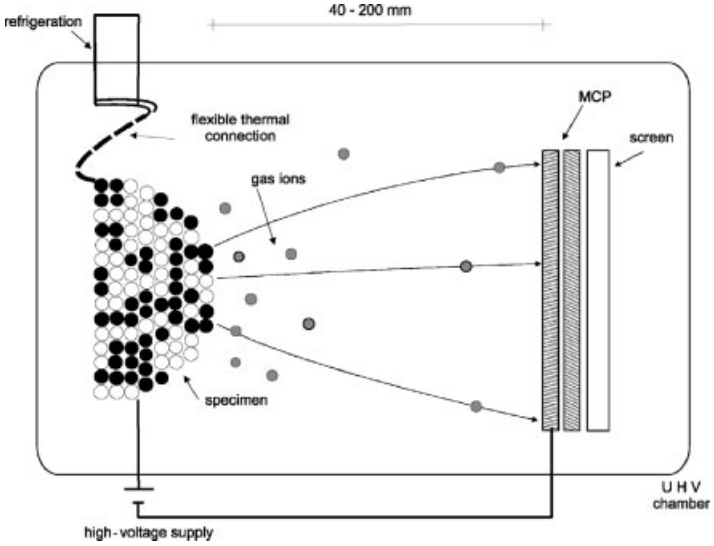


Figure 7.2 Schematic representation of a field ion microscope. MCP = multi-channel plate.

cryostat is required to cool the tip to 20–50 K. The field at the tip surface is controlled by the supplied voltage according to the relationship

$$F = \frac{V}{\beta \cdot R}, \tag{7.1}$$

in which V denotes the voltage supplied to the tip, and β denotes a dimensionless factor that varies with the exact geometry of the tip but is found in the range of 5 to 10. With increasing distance from the tip surface the field decays logarithmically, which means that the dominant drop of the field appears on the first millimeter from the tip surface.

In order to produce a field ion micrograph, an imaging gas (usually He, Ne, or a mixture of both) is introduced into the vacuum chamber. Close to the apex of the sample, the gas atoms are polarized and drawn towards the surface by the inhomogeneous field around the tip. Provided that there is sufficient field strength and a suitable distance between the gas atom and surface, a finite probability exists that an electron will tunnel from the gas atom into the band structure of the specimen. The potential well for electron transfer by tunneling is sensitively controlled by the local field strength. As a consequence, the ionization rate of the gas atoms is a function of: (i) the tip voltage; and (ii) the surface topography. After being ionized, the positively charged particle is accelerated towards the imaging screen where, by means of a multichannel plate and a phosphorus anode, the ion impact produces a visible light flash.

The trajectory of the ionized gas atom is determined by the shape of the electric field. As their thermal energy is negligible compared to the energy gain within the field, the ions are practically starting at rest. As a consequence, there is a one-to-one correspondence between the location of ionization at the tip surface and the impact position on the screen. Because of the discrete atomic structure of the sample, the local field at the surface varies in correlation to the surface corrugation. In particular, the edges of atomic terraces are protruding features and thus, are regions of elevated field strength and pronounced ionization rate. Therefore, the protruding edges of atomic terraces in crystalline structures are imaged as bright concentric rings that surround low-indexed pole directions of the crystal. This is illustrated in Figure 7.3, which shows a comparison of a field ion micrograph and the corresponding ball model of the imaged structure.

In order to achieve a clear field ion micrograph, the so-called 'best imaging field' must be established at the tip surface. To a good approximation, this field strength is only a function of the imaging gas (e.g. 35 V nm^{-1} and 44 V nm^{-1} for Ne and He, respectively). By increasing the field strength beyond this point, an alternative process of 'field evaporation' is observed. As soon as the field reaches a threshold which is characteristic of the sample material, atoms of the tip themselves are ionized, desorbed from the surface, and accelerated towards the screen thereby following very similar trajectories as the former gas atoms.

It is important to understand that the electrical field does not simply tear away the atoms, causing significant damage to the surface structure. Rather, the process remains controlled by a finite energy barrier, which must be overcome by thermal

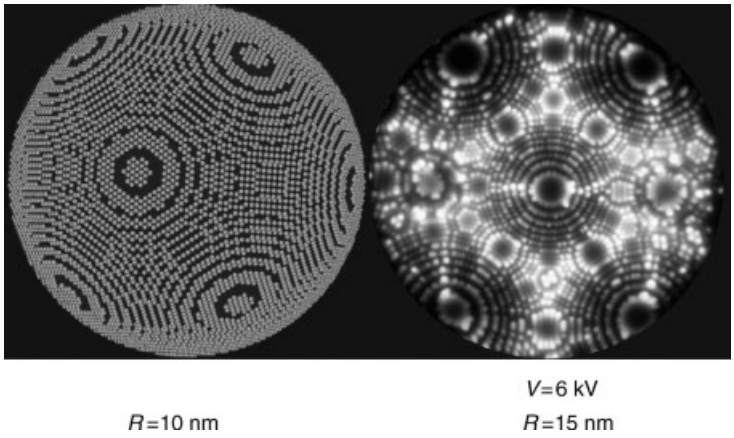


Figure 7.3 Ball model of a hemi-spherical apex (left) in comparison to an experimental field ion micrograph. Protruding atoms are represented by bright dots.The structure of the experimental FIM image at the right, comprising of concentric rings, is a natural consequence of the crystalline periodicity and the apex geometry. (Illustration courtesy of V. Vovk, University of Münster.)

activation or by quantum mechanic tunneling at very low temperatures. In other words, there is a clear justification in using the term 'evaporation', as it resembles the situation of vaporization in thermal equilibrium. This can be explained by the most simple model suitable for understanding field evaporation.

In Figure 7.4, the one-dimensional potential curves of a surface atom in its neutral and its *n*-fold charged ionic state, are plotted against the distance to the sample surface. Transferring an atom into the *n*-fold charged state requires an ionization energy, I_n, while placing the free electrons back into the band structure of the metallic

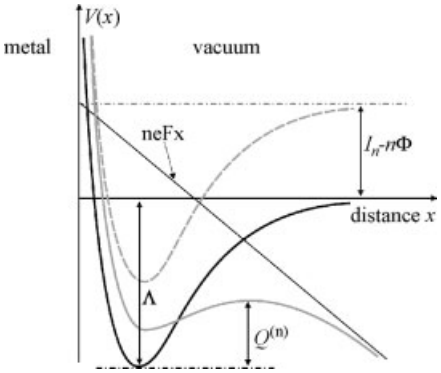


Figure 7.4 Potential curves of neutral atom (black) and ion (gray) close to the surface of the tip. The ionic curve is shown with (solid) and without (dashed) affecting field. For details of further variables, see the text.

sample delivers a payback of n times the work function, Φ. Thus, without the application of a field, the ionic potential experiences a constant shift of

$$\Delta V^{(n)} = I_n - n\Phi, \tag{7.2}$$

with respect to the potential of the neutral atom. If in addition the electrical field F is supplied, the ionic potential steadily decreases with increasing distance to the specimen surface, so that both potential curves necessarily intersect. It is also assumed in the so-called 'image hump model' that the ionic potential develops an intermediate maximum, which represents the important activation barrier. If short-ranged repulsive interactions are neglected, then the ionic potential can be approximated by

$$V(x) = -\frac{n^2 e^2}{16\pi\varepsilon_0 x} - neFx + \Delta V^{(n)}. \tag{7.3}$$

The first term on the right-hand side of the equation is caused by the image force of the metallic surface, while the second term is due to the influence of the field. Straightforward calculation yields the maximum at the hump as

$$V_{\max} = -\sqrt{\frac{n^3 e^3 F}{4\pi\varepsilon_0}} + \Delta V^{(n)}. \tag{7.4}$$

so that the activation barrier reads

$$Q^{(n)} = \Lambda + I_n - n\Phi - \sqrt{\frac{n^3 e^3 F}{4\pi\varepsilon_0}} =: Q_0^{(n)} - \alpha \cdot F^{1/2}. \tag{7.5}$$

In Equation 7.5, Λ denotes the sublimation energy of the sample material. If this activation barrier is overcome by thermal excitation, the temperature dependence of the evaporation rate is expected to follow an Arrhenius relationship:

$$\nu_{\text{evap}} = \nu_0 \exp\left(-\frac{Q_0^{(n)} - \alpha \cdot F^{1/2}}{k_B T}\right). \tag{7.6}$$

Although Equation 7.6 is derived under rather simplifying assumptions, it describes the evaporation behavior at least in a qualitative sense correctly. Some aspects – for example, the characteristic evaporation thresholds of different materials – are even in surprisingly good quantitative agreement with experimental observations. For metals, this evaporation threshold is found to lie in the range of 20–60 V nm^{-1}, a field strength which is easily achieved with specimens of approximately 30 nm curvature radius. Several modifications have been suggested to describe the evaporation, including the so-called 'charge exchange model' and also quantum mechanical concepts. These deliver partly different exponents of the field dependence in the numerator of the argument of the exponential in Equation 7.6. For the range of practical interest, the logarithm of evaporation rate varies almost linearly

with the applied field [14], so that Equation 7.5 is frequently replaced by the empirical relationship

$$Q^{(n)} \approx Q_0^{(n)} \left(1 - \frac{E}{E_0^{(n)}}\right), \tag{7.7}$$

in which $E_0^{(n)}$ denotes the field strength of vanishing barrier.

However, with all of these variations the obvious interpretation of Equation 7.6 is preserved: The desorption rate of the sample atoms is sensitively controlled by adjusting the field strength and, to a lesser extent, by the variation of temperature. In particular, the rate can be maintained at such a low level that the evaporation process can be studied in an atom-by-atom manner.

In order to perform a chemical analysis, a time- and position-sensitive detector system is placed opposite the sample, instead of a viewing screen (see Section 7.4.1). A positive dc voltage is then supplied to the tip, this being slightly too low to affect measurable field evaporation. Short high-voltage pulses of only a few nanoseconds duration are superposed to the base voltage to trigger the field evaporation. Measuring the time between the triggering pulse and the detection of a species allows identification of the evaporated particles by means of ToF spectroscopy. Typically, the pulse height is adjusted so that an event is detected after only about 1% of the pulses. Under these circumstances, the probability of evaporating multiple events comprising several atoms (which the detector system may not be able to split correctly into individual species) becomes negligibly small. In this way the sample atoms can be identified and counted, one-by-one.

For practical measurements, the choice of correct evaporation parameters, base and pulse voltages, pulse frequency and sample temperature, is an art which can only be mastered with profound experience. By continuously desorbing atoms, the samples become increasingly blunted. In order to maintain constant evaporation conditions, the tip voltage must be increased steadily during the measurement. In most cases, this is achieved under computer control, so that a constant evaporation rate in terms of the number of atoms per pulse is preserved. In order to avoid early specimen fracture, a rather low pulse amplitude and not too-low temperatures would be preferred. A low pulse amplitude would also improve the resolution of the mass spectra. However, on the other hand too-low pulses and high temperatures corroborate the accuracy of analysis, as alloy components with a low evaporation threshold may desorb in between the pulses and so become lost in the composition statistics. Suitable compromises are typically found at specimen temperatures between 30 and 50 K, and with a pulse fraction of $V_{\text{pulse}}/V_{\text{d.c.}} = 20\%$.

7.3.2
Ion Trajectories and Image Magnification

In order to understand the properties of field ion micrographs and the quality of the volume reconstruction, the ion trajectories must be discussed in more detail. An idealized specimen may be represented by the geometric model of a truncated cone closed by a hemispherical cap, as sketched in Figure 7.5 (the field lines and model

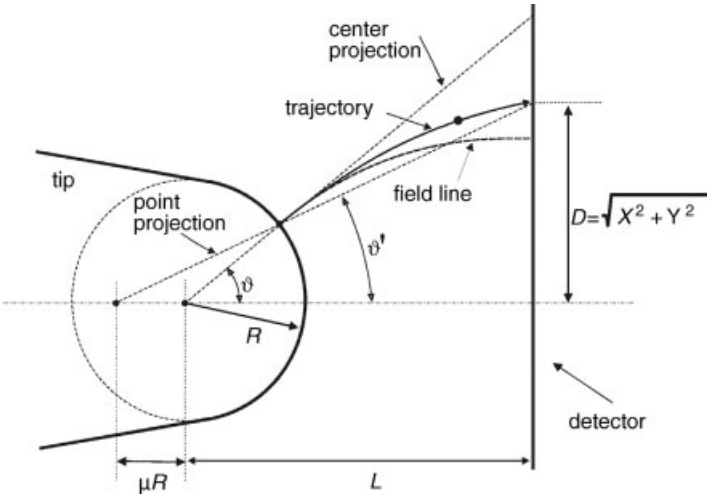


Figure 7.5 Electrical field and trajectory of an evaporated atom. The impact position and initial location at the tip surface are related by a point projection. The center of projection is shifted relative to the center of the spherical cap by μR. As an alternative, the ratio between polar angle ϑ and projection angle ϑ' may be used for evaluation.

trajectories are shown here for clarity). Owing to axial symmetry, the potential and the field can be considered in a two-dimensional space with r and z, the coordinates perpendicular and parallel to the rotational axis of symmetry, respectively. For convenience, the expression for the electrical potential Φ is split into the absolute tip voltage V and a spatial distribution function φ:

$$\Phi(r,z) = V \cdot \varphi(r,z) \tag{7.8}$$

while the equations of motion can be written in accordance to classical mechanics as:

$$\frac{d^2r}{dt^2} = -\frac{neV}{m}\frac{\partial\varphi}{\partial r}; \quad \frac{d^2z}{dt^2} = -\frac{neV}{m}\frac{\partial\varphi}{\partial z}, \tag{7.9}$$

where n and m denote the charge state and mass of the ion, respectively. Without any further calculation, it is seen that the acceleration in both coordinate directions depends on mass, charge state and voltage, in the same manner. Thus, when the ion is initially at rest, the shape of the trajectory becomes independent of all these variables. At given tip geometry, different species will follow the same path, and only the required flight time will vary and be characteristic for the given charge state and mass. This has the important consequence that the fundamental imaging relationship between the tip surface and the detector is universal for all species and that, besides minor modifications due to a slight difference in initial position, the imaging gas and specimen atoms will follow the same trajectory.

Since the ions start at rest, their motion follows initially the field lines, and they leave the spherical apex in radial direction. Later, after having gained considerable

kinetic energy, the trajectory becomes almost straight and deviates from the field lines only because of the forces of inertia. Let us define (as in Figure 7.5) the initial position of the ion by the polar angle ϑ, and the imaged position at the detector by the smaller angle ϑ'. In order to determine the original position at the tip's surface, only the function between both angles must be known. This function may vary from tip to tip, although from a practical point of view a simple proportionality holds, which is conveniently described by an imaging compression factor:

$$\kappa := \vartheta'/\vartheta. \tag{7.10}$$

This factor can be calibrated by means of field ion micrographs of single crystalline specimens, such as depicted in Figure 7.3. The angle between the different pole directions, and thus the polar angle ϑ, is known from crystallography, while the detection angle ϑ' is determined from the position of the pole in the FIM micrograph and the flight distance L between tip and screen. Typical data determined for an electropolished tungsten tip are shown in Figure 7.6. Clearly, the linear relationship of Equation 7.10 is well fulfilled; a compression factor of $\kappa = 0.54$ is determined for this exemplary specimen.

It is straightforward enough to quantify the magnification of the analytical microscope on the basis of the geometric model of Figure 7.5. The polar distance in the image $D = L \cdot \sin\vartheta' \approx L \cdot \vartheta'$ must be compared to the distance at the hemispherical cap $d \approx R \cdot \vartheta$, where R denotes the current radius of the apex. Thus, the magnification is given by

$$M := \frac{D}{d} = \frac{L \cdot \kappa}{R}. \tag{7.11}$$

Recalling that a typical tip radius amounts to about 30 nm, and that the distance between the detector and tip may reach 50 cm, a magnification of 10^7 is easily obtained.

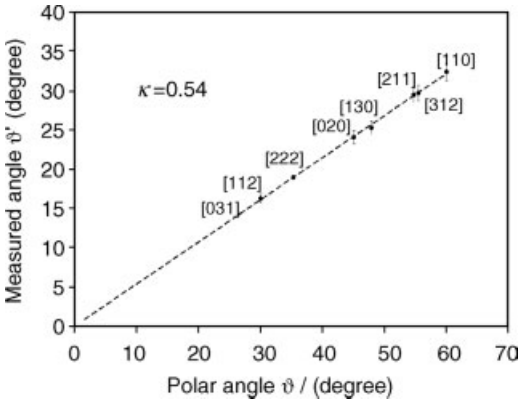


Figure 7.6 The relationship between measured angle ϑ' and polar angle ϑ, as determined for electropolished tungsten tips. The polar angle has been determined from crystallography. Tip axis aligned parallel to [011]. (Illustration courtesy of P. Stender, University of Münster.)

In Equation 7.11 an interesting detail is noteworthy. The magnification of the microscope depends on the tip radius of the specimen or, in other words, the specimen itself represents the essential lens of the microscope. Therefore, APT will only function in a reliable manner, if the specimens are prepared carefully and are notably reproducible in shape. Furthermore, the tip radius increases during the measurement, as the specimen field-evaporates continuously and, consequently, the magnification will decrease during the measurement. In order to reconstruct the spatial arrangement of the atoms after the measurement, the evolution of the radius must be recorded or estimated in a suitable manner.

The introduced geometric model of the tip neglects the roughness of the surface on the atomic scale. In reality, the edges of the atomic terraces and faceting of the spherical surface (low-indexed surface orientations become more pronounced in size due to anisotropy of the evaporation probability) will lead to slight modifications of the trajectories (this point is discussed further in Section 7.3.4).

7.3.3
Tomographic Reconstruction

During each measurement, several million events, the respective flight times and impact positions are recorded. From these raw data the original spatial arrangement of the atomic species is reconstructed by efficient, yet surprisingly simple, algorithms. In order to reduce the mathematical effort, we assume the specimen axis to be aligned perpendicular to the detector plane. The outlined scheme follows the studies of Bas *et al.* [15], and the general case of taking into account a relative rotation between tip and detector can be treated in an analogous manner (appropriate formulas are available in Refs [5, 16]).

The evaluation of data is conveniently subdivided into three steps:

- The specific mass m/n is calculated from the ToF.
- The lateral position at the tip surface is calculated from the impact position at the detector.
- The depth scale along the symmetry axis of the specimen is determined from the data sequence.

In the following section we use the geometric parameters as defined in Figure 7.5.

As the field lines are concentrated at the tip apex, the ions gain the major fraction of their kinetic energy during only the first millimeter of their trajectory. Later, the motion is almost straight and uniform, so that from conservation of energy the specific mass is calculated to sufficient approximation by:

$$\frac{m}{n} = \frac{2t_{\mathrm{ToF}}^2 e(V_{\mathrm{tip}} + V_{\mathrm{pulse}})}{L^2 + X^2 + Y^2}. \tag{7.12}$$

From the geometric detection angle

$$\tan\vartheta' = \sqrt{X^2 + Y^2}/L, \tag{7.13}$$

the Cartesian coordinates of the position at the tip's surface can be determined by

$$\begin{aligned} x &= \frac{X}{D} R \sin\vartheta = \frac{X}{D} R \sin(\vartheta'/\kappa) \\ y &= \frac{Y}{D} R \sin(\vartheta'/\kappa) \\ \tilde{z} &= R(1-\cos(\vartheta'/\kappa)). \end{aligned} \tag{7.14}$$

if the image compression factor κ has been calibrated before. In Equation 7.14 the axial coordinate $\tilde{z}$ has been marked by a tilde to express that this axial position is only preliminary, as it is still given relative to the position of the tip front z_0. As this reference point shifts during the measurement, we must correct the depth position in a final evaluation step. With each evaporated atom, the specimen is eroded by one atomic volume; thus, the number of detected atoms represents a natural depth scale. To establish this scale, the actual image magnification, which relates the sensitive area of the detector to the investigated area at the apex, must be taken into account. By expressing all of this in a differential equation, we obtain

$$dz_0 = \frac{\Omega}{\rho A_{\text{measured}}} dN = \frac{\Omega \cdot M^2}{\rho A_{\text{detector}}} dN, \tag{7.15}$$

where Ω and N denote the average volume per atom and the number of detected atoms, respectively. The factor ρ takes into account the limited detection probability of the detector ($\rho \approx 0.5$) and the magnification M is calculated by means of Equation 7.11. By applying Equation 7.15, the total shift of the tip front relative to its initial position at the start of the measurement is found by integration, and the final z-coordinate results from summing both contributions: $z = z_0 + \tilde{z}$.

In order to evaluate Equation 7.14 or 7.15, the instantaneous tip radius R must be known. If the evaporation properties of the investigated material are reasonably homogeneous, this radius is concluded from the total voltage (dc plus pulse voltage). The preset evaporation rate since is known to be obtained at the critical field strength E_{evap} of the investigated material, we can use by inversion of Equation 7.1.

$$R = \frac{U_{\text{tot}}}{\beta \cdot E_{\text{evap}}} \tag{7.16}$$

to determine the actual radius. However, this scheme is only feasible if the evaporation properties of the sample are reasonably constant. In heterogeneous specimens, for example the thin-film layer type, the radius must be concluded from geometric considerations (see e.g. Ref. [17]).

A typical reconstruction, calculated by the outlined formulas, is shown in Figure 7.7. In this case, the analyzed volume stems from a Cu/Co multilayer specimen. The position and chemical identity of each detected atom is marked by a color-coded dot. With modern wide-angle instruments, the lateral width of the reconstructed volumes reaches about 50 nm, but with blunted tips even 100 nm can be achieved. In view of the rather simple algorithms used, it comes as a surprise that even lattice planes of the

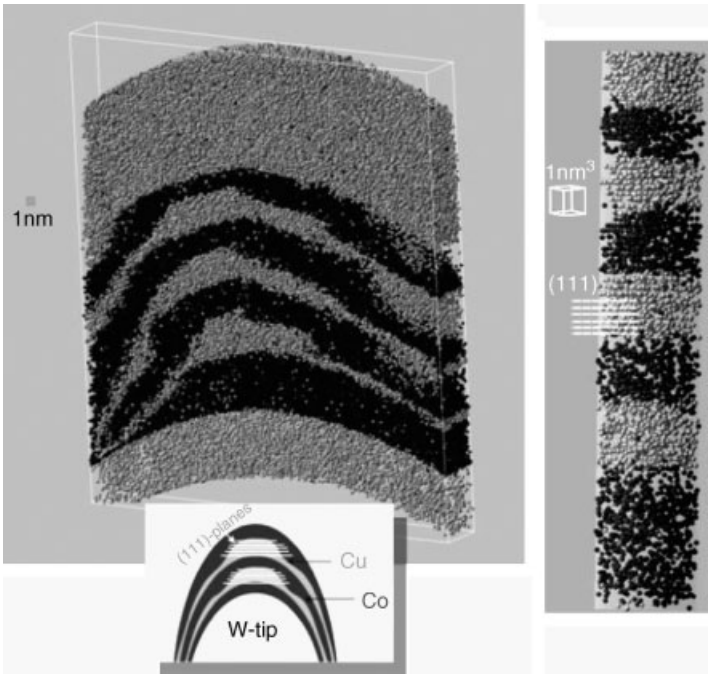


Figure 7.7 Atom probe tomography. Positions of individual atoms are represented by color-coded dots in a perspective representation. This is part of a larger data set of a Cu/Co multilayer after 60 min annealing at 450 °C. The detail on the right-hand side documents lattice plane resolution. (Measurements performed with the WATAP at University of Münster by V. Vovk [60].)

crystalline structure are resolved, as shown in the detail of Figure 7.7. This very welcome feature is explained by the physics of the evaporation process. As atoms at the edges of the atomic terraces have the highest evaporation probability, low-indexed lattice planes that are aligned parallel to the specimen surface have the tendency to desorb in a layer-by-layer mode. Resolved lattice planes demonstrate the outstanding spatial resolution of the method and allow the calibration of important parameters of the reconstruction algorithm. If the critical evaporation field strength, the field and image compression factors β, κ, and the detection probability are chosen correctly, indeed the correct lattice spacing expected for the material is reproduced.

After having reconstructed the spatial arrangement of the atoms, various averages, composition profiles, 2-D compositional maps and iso-concentration surfaces may be derived by sorting and counting the species in suitable subvolumes. Different algorithms were proposed, and are indeed in use, to detect precipitates and the shape of interfaces automatically [18]. The further analysis by Fourier methods [19, 20] or the calculation of pair correlation functions [21] similar to image analysis in the 2-D world of electron microscopy, has also been demonstrated.

7.3.4
Accuracy of the Tomographic Reconstruction

As discussed earlier (see Section 7.3.2), the sample itself represents the critical 'lens' in the projective geometry of the atom probe. Thus, it is not surprising, that a well-prepared shape of the specimen, usually produced by continuous field evaporation prior to the measurement, is the most critical issue in practical work. If the process of 'field development' is conducted too rapidly or insufficiently, then a variety of artifacts may be induced. Of particular danger here is the partial fracture of a specimen before or during measurement, as this usually produces surface topologies that are unsuitable for a reliable spatial reconstruction. In addition, the depth scale will be erroneously calibrated due to the intermediate loss of material. Yet, even if the experimentalist obeys all rules of good experimental practice, the positioning of the atoms cannot be perfect – at least as long as the evaluation scheme outlined in Section 7.3.3 is used. Apart from large angle corrections, this scheme is currently the 'state of the art'.

As illustrated in Figure 7.7 (an even clearer example is shown in Figure 7.20b), it is quite common to reproduce a set of lattice planes in volume reconstructions of pure metals. Usually, these planes are aligned almost parallel to the tip surface. In exceptional cases, several different lattice sets, inclined to each other, could be detected at the same time. By careful Fourier analysis of such examples, the accuracy of the atomic positions in the reconstruction was quantified [19, 20]. By deriving static Debye–Waller factors from the intensity of higher order Fourier components, the standard deviation of the individual atoms from their ideal lattice positions was determined at the example of a pure iron sample to $\sigma_{\perp} = 0.03$ nm in the direction perpendicular to the local tip surface, and to $\sigma_{\parallel} = 0.1 \ldots 0.15$ nm in lateral direction [20]. The strong anisotropy in resolution is a characteristic of the atom probe method. It should be noted here that this outstanding high accuracy has been observed under 'best-case' conditions – that is, the investigation of a pure, coarse-grained metal at rather low temperature. Frankly, a microscopic analysis of such a specimen is useless. In relevant cases, of heterogeneous samples measurements are significantly less accurate. In particular, it is impossible to exploit the impressive depth resolution, way better than 1 Å, in order to determine the shift of interstitial defects out of the host lattice planes. Owing to the principle of depth scaling, these defects will be assigned to either of the neighboring lattice planes. Even worse, the localization of these defects is determined by their relative evaporation threshold rather than their original physical position in between the host lattice.

It is instructive to consider the origin of these inaccuracies. The reconstruction scheme outlined in Section 7.3.3 is based on two important assumptions: (i) the tip apex may be represented by a perfect sphere; and (ii) atoms evaporate individually in a predictable layer-by-layer sequence. The former is the prerequisite for accurate lateral positioning, while the latter is critical for exact depth scaling. It is clear that already with pure metals neither assumption is perfectly met, as the atomic scale roughness and surface facets (which are unavoidable due to anisotropy of the evaporation threshold) are neglected. With alloys, the situation becomes even worse.

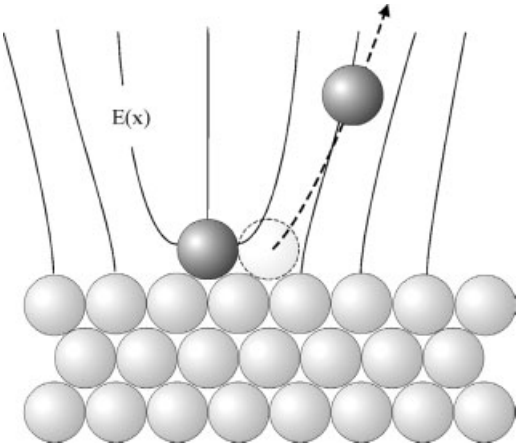


Figure 7.8 Schematic field distribution at a low indexed lattice plane just before being completely evaporated. The last but one atom is accelerated by a field which is significantly deformed by the last atom.

As the various components evaporate at quite different field levels, the sequence of evaporation can be severely disturbed in this case.

Furthermore, in heterogeneous microstructures, so-called 'local magnification effects' prevent the simple mapping of the projective geometry. For example, the embedded particles of a material of the higher evaporation threshold tend to protrude from the tip surface, which leads locally to an increased curvature of the surface and thus to an inhomogeneous magnification by the microscope. The undesired consequence is that the particles appear artificially broadened and that their atomic density is underestimated in the reconstructions. Worse still, due to overlapping of the trajectories artificial mixing of the materials from both sides of the particle interface is erroneously measured.

In order to understand why positioning in lateral direction is less accurate, we consider the situation sketched in Figure 7.8, when only very few atoms are left shortly before evaporating a further low-indexed lattice plane completely. In this case, the electrical field that controls the trajectory of the atom next in evaporation line will definitely deviate from the idealized field surrounding a spherical surface, due to local disturbance of the remaining protruding atoms. Thus, in order to calculate the trajectories exactly, the structure must be known in advance. We are faced with an implicit problem *when* the atoms should be localized exactly on the atomic scale. The solution of this problem remains a goal for the future. Meanwhile, in order to achieve a sound evaluation of atom probe data, the only promising strategy is to simulate the evaporation sequence of hypothetical specimen structures and to compare the simulated reconstructions with those of real experiments. Such a procedure is quite analogous to the normal practice in high-resolution electron microscopy, where experimental images are compared to those simulated from hypothetic structures until a good match is found.

For that task, Vurpillot and coworkers [22, 23] derived a simulation scheme that allowed an investigation of the spatial accuracy of tomography and the influence of

heterogeneous evaporation on theoretical grounds. As the simplest geometric model, which still reflects microscopic features on the atomic scale, these authors suggested constructing the apex from a simple cubic arrangement of Wigner–Seitz cells. In a first step, the electrical field surrounding the model tip is calculated by solving numerically Poisson's equation for the electrical potential by means of a finite element method. The electrical field strength at the locations of the surface atoms is determined, and the position of the highest field strength identified. In a second step, the atom at this specific position (meaning the corresponding Wigner–Seitz cell) is removed from the apex model and the field is recalculated for the new configuration. In a final step, the trajectory of the removed atom is calculated between the tip surface and detector in accordance with classical mechanics which considers the acceleration of the ion within the electrical field. In the case of alloys or heterogeneous systems, the different evaporation probabilities of the atomic species must be taken into account. Thus, before the position of highest field is selected in the first step, the electrical field is scaled artificially by a factor varying from atom to atom in order to reflect the respective evaporation probability.

By repeating this scheme recursively to predict the impact positions of several thousand atoms, a part of a measurement can be simulated quite realistically. The general features of the experimental data are well reproduced, although an artificially short flight distance between the tip and the detector, and rather small specimens of approximately 20 nm radius, had to be used to limit the computational effort. The statistical nature of thermal activation has also been neglected. In Figure 7.9, the

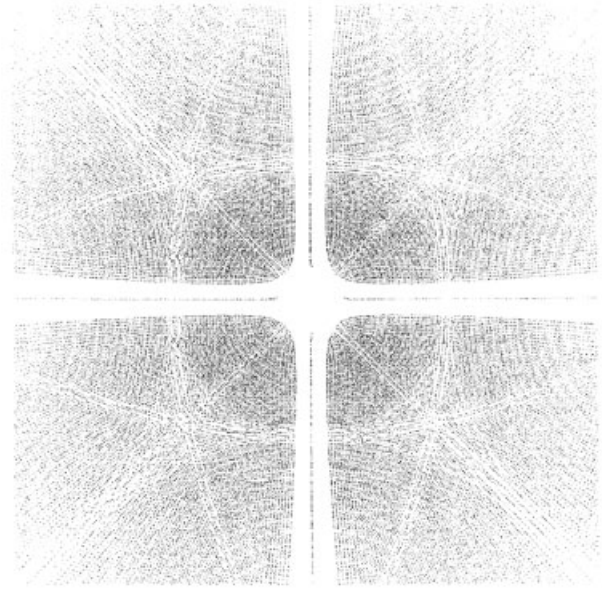


Figure 7.9 Simulated impact positions on the detector area, during evaporation of a few lattice planes of a simple cubic model alloy. The tip axis is aligned along [001]. (Reproduced with permission from Ref. [22].)

impact positions of several thousands of atoms on the detector are shown, as calculated from the simulated evaporation of a few lattice planes. In contrast to the behavior of an ideal microscope, the spatial density of the events is by no means homogeneous. Rather, lines of significant redistribution of the atoms are seen, which are related to low-indexed zones of the crystal structure. Clearly, this redistribution is related to the situation sketched in Figure 7.8. The last few atoms sticking on a flat, low-indexed surface are affected by severe field distortions and therefore, their trajectory is significantly disturbed in comparison to a simple point projection.

Deviations of the trajectories induced by the local surface topology on the atomic scale are the main limiting factor for the instrument's lateral resolution along the tip surface. Although various ideas have been proposed to correct for the pronounced deviations at zone lines in pure samples, a practicable improved reconstruction algorithm has not yet been presented. In the case of statistical alloys, the situation is particularly difficult, as the chemical neighborhood of the evaporated atom is not known. One may imagine iterative algorithms to refine the reconstruction. But even the atom probe detects only 60% of the atoms (see Section 7.4.1). Thus, the local chemistry is never known completely. The effect of disordered alloys becomes particularly clear if, in simulated measurements, the atom positions of a disordered alloy are compared to those of a pure tip of identical geometry. An example is shown in Figure 7.10. Owing to the different evaporation probabilities of the two species, the evaporation sequence is disturbed and the local fields are distorted in the case of an alloy in a different manner as compared to a pure specimen. Clearly, the reconstructed positions of the atoms do not agree. However, as can be seen from Figure 7.10, most positions agree within about one lattice constant. Only a minor fraction of atoms originating from zone line positions are shifted by much larger amounts, up to five lattice constants. In this way, the simulation indicates the lateral

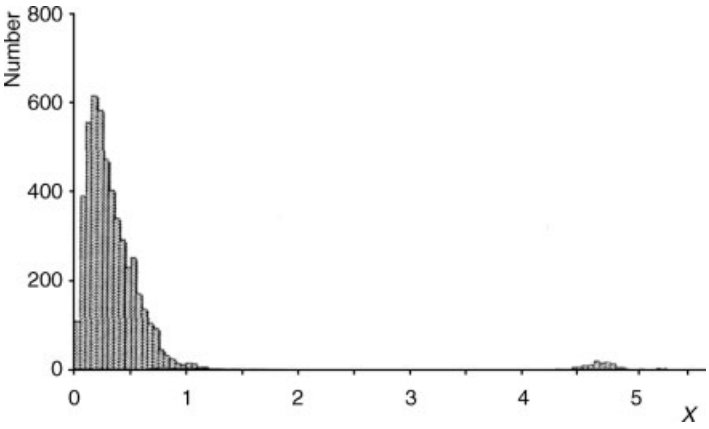


Figure 7.10 Statistics of reconstructed atom positions of an AB model alloy with respect to their ideal position in an homogeneous specimen (relative shift x in units of the lattice constant). (Reproduced with permission from Ref. [22].)

accuracy of the tomography to be slightly better than a lattice constant, as long as poles and zone lines are avoided for analysis.

7.4
Experimental Realization of Measurements

7.4.1
Position-Sensitive Ion Detector Systems

The rapid progress of APT in recent years has been made possible only by the remarkable evolution of spatially resolving detector systems with single-ion sensitivity. During the past two decades, several detector concepts have been proposed and put into operation. Those systems currently in use will be discussed at this point.

All available detector concepts are based on a stack of two to three multichannel plates (MCP). An MCP represents a secondary electron multiplier with many independent channels working in parallel. The device is composed of thousands of small glass tubes, each approximately 25 μm in diameter, and packed in parallel alignment to form a plate of about 1 mm thickness (see Figure 7.11). The front and reverse sides of the plate are coated with thin metallic films which serve as electrodes to supply a voltage in the 1 kV range. An ion which hits the inner wall of such a glass tube will produce a few secondary electrons that are accelerated by the supplied field. On their way towards the reverse side, they impact the glass wall several times and produce further secondary electrons. This cascade process finally produces a cloud of about 10^4 electrons per ion. If a consecutive stack of two or three MCPs is used instead of a single unit, the individual amplification factors will be multiplied so that a single ion hitting the front side with sufficient energy will produce finally a cloud of about

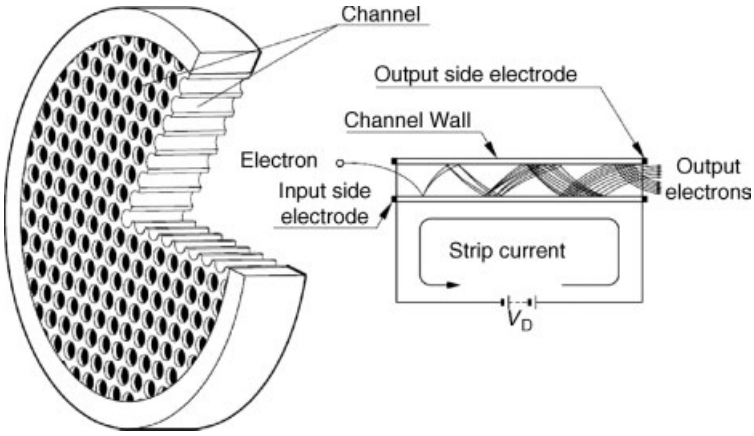


Figure 7.11 Functional principle of a multi-channel plate (MCP). The MCP is an arrangement of thousands of secondary electron multipliers working in parallel. Each of the electron multipliers is made from a tiny glass tube, 25 μm in diameter.

10^8 elementary charges, which is sufficient for further electronic evaluation. An MCP is a rather rapidly operating device; indeed, with optimized electronic circuitry the raise time of a single ion pulse is in the range of 100 ps. The spatial resolution of the MCP is determined by the dimension of the glass tubes. One important drawback here is that the detection efficiency of an MCP is way below 100%. Hence, only those charged particles which penetrate into one of the tiny tubes will induce the described avalanche process, while those hitting the massive front side are simply reflected. Due to mechanical requirements, channel plates cannot be produced with an open area fraction significantly larger than 60%. Attempts to improve the detection probability by placing additional electron mirrors in front of the channel plate have not been sufficiently successful so as to be used in modern-day instruments. Thus, based on principle, the presently available atom probes do count only half of the atoms of the analyzed volume.

For imaging purposes in FIM, it is sufficient to place a phosphorus screen behind the exit face of the MCP, so that each electron cloud produces a short light flash. Several attempts have been made to record these light flashes by means of a gated CCD camera in order to determine also quantitative positional information; examples include 'Optical PoSAP' [24] and 'Optical TAP' [25]. However, both systems suffered from a rather slow read-out of the camera, which severely limited the practical pulse frequency. Therefore, state-of-the-art instruments usually evaluate the charge clouds using methods mostly developed by nuclear physicists. For this, a multiple anode array is placed behind the MCP and connected to sensitive preamplifiers and fast converters in order to transform analogous charge information into digital data. Various concepts can be distinguished by their different layouts of the anode array and the complexity of the electronics. Historically, the first functioning system was the PoSAP, which was built around a 'wedge and strip anode' [26]. The name of this anode is self-explanatory, based on the sketch in Figure 7.12a. The geometry is designed in such a way that the relative fractions of the total charge measured on the three electrodes Q_X, $Q_Y$and Q_Z, vary with the position of the electron cloud. For the layout shown, the position may be calculated in a straightforward manner by

$$X \propto \frac{Q_X}{Q_X + Q_Y + Q_Z}; \quad Y \propto \frac{Q_Y}{Q_Y + Q_Z}. \tag{7.17}$$

Since only three independent anodes are used, the required electronics is reasonably simple. However, the layout has the important drawback that the total area of each electrode – and thus the respective capacities – are quite large and the drain of charges after the impact takes a considerable time. If a second ions hits the detector within this time gap, then both events cannot be separated; consequently, the operator is forced to use very low data rates.

Following a significant effort to improve the electronic instrumentation, a square array layout of many smaller electrodes was realized shortly afterwards, with the original TAP detector (see Figure 7.12b) [10]. By choosing the correct distance between the MCP and anode array, the electron cloud will always spread over at least three or four electrodes, so that its central position can be determined by charge

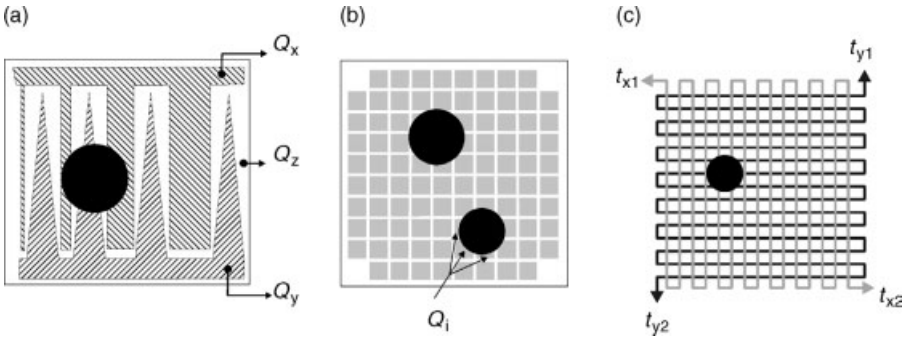


Figure 7.12 Anode layouts for the read-out of the positional information of a channel plate. (a) Wedge and strip detector; (b) TAP detector; (c) Delay line detector. Here, instead of the double-wire Lecher lines, only a single wire is shown to illustrate the principle. The extension of the electron cloud is symbolized by black circles.

weighting. Due to its parallel design the detector has some capability to separate multiple events, which allows much higher data rates than with the PoSAP detector.

The latest instruments (those constructed after 2003) mainly apply the delay line principle, which was first proposed in 1987 [27]. Here, instead of flat electrode areas, two independent double wire spirals are used that are wound along the *X* and *Y* axes of the detector, as shown schematically in Figure 7.12c. Each double wire represents a Lecher line, on which the pulse signal propagates with the velocity of light. As opposed to previous concepts, the spatial information is not determined from charge weighting but rather from time measurements, so that no expensive measurement of analogue signals is required. Each ending of the Lecher line is connected to a separate channel of a fast time-to-digital converter (TDC) with sub-nanosecond resolution. If the ion impacts at the center of the detector, the pulse signals will propagate symmetrically to both ends of the Lecher line, and will therefore reach the TDC at exactly the same time; in contrast, for an asymmetric impact position the two time signals will differ considerably. The sum and difference of the two time signals of such a Lecher line are presented in Figure 7.13; these were collected for many independent impacts on a circular detector of 120 mm diameter and a spire spacing of the anode of 1 mm. The sum of both time signals represents an instrumental constant, and corresponds approximately to the propagating time from one end of the line to the other. With 150 ns, this time interval is easily measurable. The time difference between both signals is proportional to the position according to

$$X = v_p\left(t_x^{(l)} - t_x^{(r)}\right), \tag{7.18}$$

where the calibration parameter v_p denotes the propagation velocity along the spiral axis, which amounts to about 0.4 mm ns^{-1}. An analogous equation holds for the *Y* direction. With modern computer electronics, a time resolution below 100 ps – and

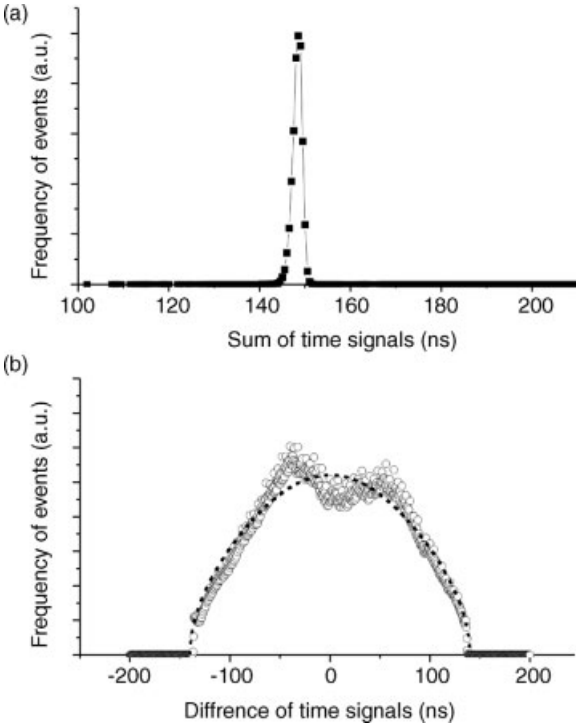


Figure 7.13 Evaluation of time signals of the delay line detector. (a) The sum of both signals is a constant which may be used to correlate time signals to individual events; (b) The difference is a direct measure of the position. The dashed line represents the spatial distribution expected for a circular detector in wide-angle configuration. Experimental data deviate due to anisotropic evaporation of the crystalline sample. (Illustration courtesy of P. Stender, University of Münster.)

thus a positional accuracy of better than 0.1 mm – is achievable. As a fast TDC is already required for ToF mass spectrometry, delay line detectors represent a very economic solution. Furthermore, the delay line principle allows very high data rates, so that evaporation pulses may be applied with frequencies of up to 300 000 Hz. In order to improve the multi-hit capability, a hexagon anode design [28] and quantitative evaluation of the pulse shape by use of fast oscilloscope electronics has been realized [29]. However, it is not yet clear whether this additional effort in instrumentation will provide an advantage in practical terms.

7.4.2
Instrumental Design of 3-D Atom Probes

In principle, a 3-D or tomographic atom probe consists of the same components as the FIM shown in Figure 7.2. Only the viewing screen must be replaced with one of the above-discussed position-sensitive detector systems, and the high-voltage supply

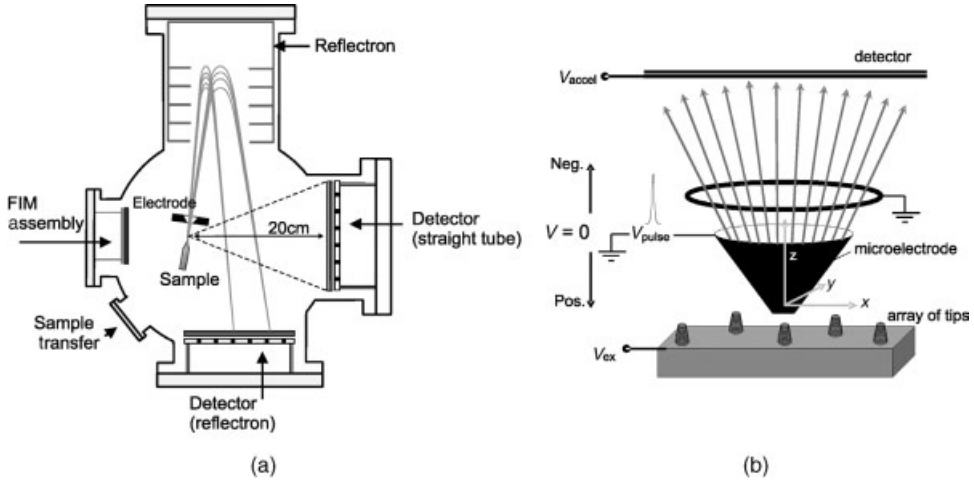


Figure 7.14 The design of the tomographic atom probe operated at the University of Münster [31]. The dedicated chamber layout allows switching between different geometries: (i) the straight flight tube yields a short flight path to optimize the open angle of the instrument; (ii) the reflectron arrangement yields an improved mass resolution; (b) The principle of a microelectrode atom probe. An extraction electrode with an open diameter of approximately 10 µm is placed close to a microtip to concentrate the electric field. (For an example, see Ref. [13].)

must be extended by the required voltage pulsing. By using the voltage pulse as start signal for an accurate time measurement, and the detector signal for stopping, the ToF of the ions can be measured in straightforward manner. However, the flight distance between the tip and detector must be adapted to obtain the desired mass resolution. During the past two decades, this flight distance has been steadily decreased in response to the improving time resolution of available measurement electronics. Recently, the flight path was reduced to 10–20 cm in so-called 'wide-angle instruments' [30, 31]. With detector diameters of up to 120 mm, this yields a geometrical aperture of $2\vartheta' \approx 44°$, which means an even larger effective opening in the range of $2\vartheta \approx 60°$ due to the curvature of the ion trajectories (as explained in Section 7.3.2). A schematic representation of a modern conventional 3-D atom probe is shown in Figure 7.14a. In contrast to the concepts discussed above, two modifications should be noted:

- Instead of a straight flight tube, a reflectron geometry is used with an ion mirror that leads to parabola-shaped ion trajectories. This geometry compensates for fluctuations in the kinetic energy of individual ions. A faster ion will penetrate deeper into the mirror field, and so will have a longer flight path. If the length and voltage of the reflectron are adjusted correctly, then ions of identical mass but with slightly varying initial velocity will hit the detector after identical flight times. In this way, the mass resolution is significantly improved. Originally, reflectrons were designed with a homogeneous mirror field. With such a device, a perfect time focusing is paid by trajectory deviations which can only be tolerated for small aperture angles. Therefore, a conventional reflectron is not compatible with the

wide-angle concept. Very recently dedicated reflectrons have been designed with curved electrode shapes, which eliminate this problem [12]. When using a reflectron, the typical mass resolution of a 3-D atom probe can be expected to be $\Delta m/m = 1/500...1/2000$, whereas without such a device the resolution is limited to $\Delta m/m \approx 1/100$ [all data full width at half maximum (FWHM), determined at an effective mass of $m = 30$].

- The evaporation trigger is supplied as a negative pulse to an extraction electrode in front of the tip, which allows shorter and better-defined pulse shapes. Recently, the use of an extraction electrode has created important progress towards miniaturization. If a large number of atoms were to be evaporated and measured in a reasonable duration of the experiment, the pulse frequencies would need to be as high as possible. However, this strategy finds a natural limit, as no practicable means are available to produce a 5 kV nanosecond pulse with frequencies exceeding 20 kHz. A very intelligent method to circumvent this technical problem is to use a micrometer-sized electrode, as suggested in 1994 by Nishikawa and Kimoto [32]. By placing a tiny extraction electrode of some 10 µm bore size close to the tip, the β-parameter of Equation 7.1 is reduced by a factor of 2 to 3, as exemplified by the experimental data in Figure 7.15. In consequence, much lower voltage pulses are required, which today can be produced with repetition rates far in excess of 100 kHz. After solving any related technical problems, the concept has been put into operation during the past few years. Meanwhile, these instruments are functioning well and available commercially [33]. Their efficiency of analysis is impressive [34]; a data rate higher than 10 000 atoms per second is obtainable and data sets with more than 10^8 atoms have been routinely achieved. Beside the high

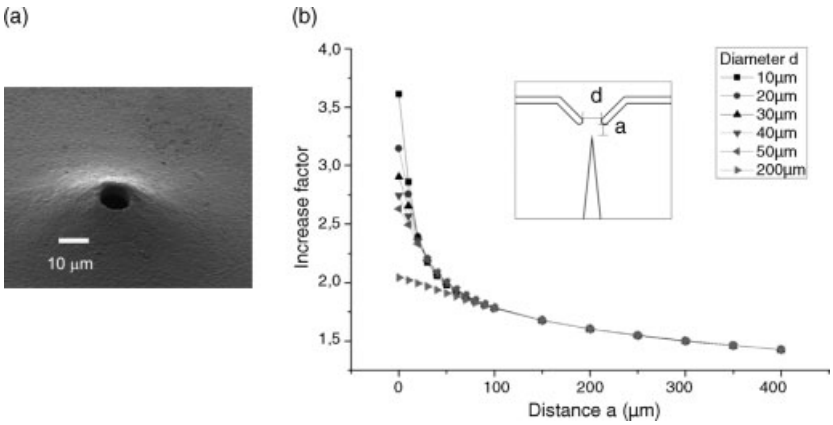


Figure 7.15 The increase of field by a microelectrode in front of the tip allows the voltage to be reduced by a factor of two to three at a constant tip radius: (a) SEM micrograph of an electro-plated Ni microelectrode (b) The relative increase of field at constant voltage for microelectrodes of 20 to 200 µm is plotted against the spacing between tip apex and electrode. (Illustration courtesy of Ralf Schlesiger, University of Münster [88].)

data rate, these instruments have an important advantage in specimen preparation. As indicated in Figure 7.14b, an array of microtips may be used rather than a single tip made from a supporting wire. As the requirement for a large aspect ratio is considerably relaxed by the microelectrode geometry, the tip array may be conveniently produced by sputtering through a suitable mask. Very recently, In As nanowires, grown naturally by using nanopatterned Au catalysts, were directly analyzed using a tomographic atom probe equipped with such a microelectrode [35].

7.4.3 Specimen Preparation

As APT requires a dedicated needle-shaped sample geometry, obtaining suitable specimens is a delicate matter. Although the art of preparation has undergone considerable progress, a large proportion of desired measurements fail due to this issue. Traditionally, the required needles were produced by the electropolishing of thin metallic wires under *in situ* observation by means of optical microscopy. At present, thin films and other complex nanostructure are the focus of interest, of which usually no conventional wire is available.

For studies on reactions in thin-film materials, layer systems have been deposited onto tungsten tips, which serve as a substrate. In order to achieve an optimum shape, the freshly prepared tungsten substrates are field-developed prior to deposition. As considerable stress is induced by the electrical field during the measurement, many investigations are prevented by insufficient mechanical stability of the interface between substrate and coating. Thus, this interface requires special care. Occasionally, an interlayer (often chromium) is used as an adhesion aid, while ion-beam cleaning of the substrates immediately before deposition has also been shown as advantageous [36]. In Figure 7.16a, a thin-film specimen produced in this way,

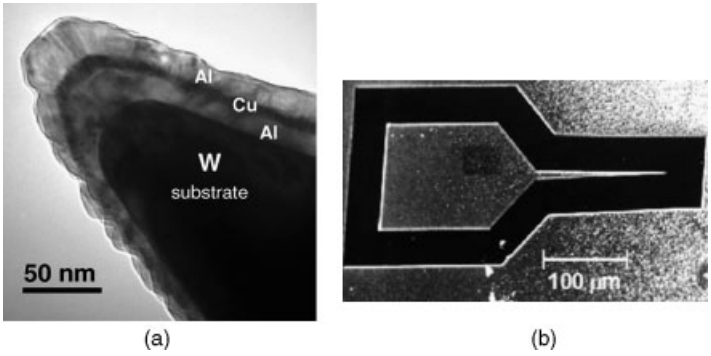


Figure 7.16 (a) Example of an Al/Cu/Al trilayer on a tungsten substrate tip, deposited using an ion beam sputtering technique. (Illustration courtesy of C. Ene, Göttingen.); (b) Scanning electron micrograph of a layer specimen prepared by electron-beam lithography. (Illustration courtesy of J. Schleiwies, Göttingen.)

namely an Al/Cu/Al trilayer deposited on tungsten, is presented. In this geometry, the main analysis direction is aligned normal to the interfaces, so that these specimens are especially suited for the investigation of reactive diffusion by local depth profiles of composition. However, when interpreting the analysis results it must be borne in mind that these films are deposited onto a curved surface. Usually, the curvature induces a rather small grain size, so that the microstructure is not directly comparable to that of thin films deposited onto conventional planar wafer substrates [37].

If this variation in microstructure cannot be tolerated – for example, because the properties of technical devices should be characterized – the tips may be cut by either lithography [38] or focused ion beam techniques. For the former a planar layer system is first coated with a suitable photo resist, and exposed to electron-beam lithography and developed chemically to obtain a suitable etching mask. A typical sample is shown in Figure 7.16b. The tip is attached to a 'handle' which is about 100 µm in length, with the wedge-shaped needle pointing to the right-side taper to a width below 100 nm at the apex. After etching by sputtering, the tips are removed from the substrate and glued to a supporting wire. In this geometry, the interesting interfaces are aligned parallel to the main analysis direction; thus, the method is well suited to investigations of interfacial roughness or pin holes in multilayers of small periodicity.

With the emergence of focused ion beam (FIB) facilities, equipped with a Ga beam of 5 to 30 keV, the preparation of difficult nanometric geometries has been revolutionized. Thus, it is no wonder that this technique is being used increasingly to prepare the required needles. Following an original proposition by Larson [39], thin films are deposited on top of flat-ended Si posts for that purpose. The width of the rectangular prism-shaped posts is chosen to be about $10 \times 10\,\mu m^2$ in cross-section – sufficiently large to mimic realistically the deposition conditions of a larger planar substrate, yet at the same time thin enough to limit the required beam time of the FIB. As an alternative, the cutting of a thin bar directly from a massive volume has also become usual practice as the intensity of available Ga beams has been further improved. Originating from transmission electron microscopy (TEM) studies [40], this 'lift-out' technique has been recently adapted to the needs of APT [34]. The essential preparation steps are shown in Figure 7.17. With any FIB method, the final step to produce a sharp tip is always an annular milling from the front face with a continuously decreasing size of the circular mask, as indicated in Figure 7.17c. It goes without saying that great care must be taken to avoid irradiation damage of the part which is to be analyzed. At 30 keV, the Ga ions of the cutting beam can be expected to penetrate at least 20 nm into the sample, and therefore suitable metallic coatings must be used to cover sensitive areas. Then, on completion of the procedure the energy of the beam should be reduced as much as is practicable.

The FIB has certainly revolutionized the preparation task, particularly if the tips are to be cut from chemically or mechanically difficult materials such as multilayers and semiconductors or ceramics. On the other hand, it cannot be overlooked that 'beam time' on these machines is still a rather limited resource. In this context, it should be noted that although many insulated measurements of FIB-prepared samples using

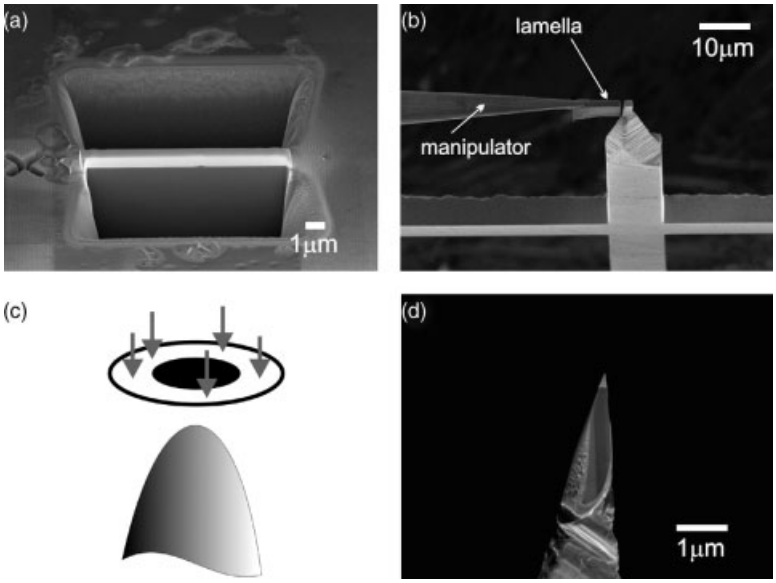


Figure 7.17 Stages of focused ion beam preparation using the lift-out procedure. (a) Cutting of lamella perpendicular to the substrate. The lamella edges are covered with deposited Pt; (b) Lamella moved to a supporting post (e.g. Cu). A part of the lamella had been fixed to the post and the remaining cut by the ion beam; (c) Annular milling from the front with continuously decreasing aperture size; (d) The final tip sample. (Illustration courtesy Ralf Schlesiger, University of Münster [86].)

APT are reported today, very few experimental series have been reported that characterize the different stages of a physical process. This may be seen as an indicator that the reliability and efficiency of the FIB procedures still require significant improvement.

7.5 Exemplary Studies Using Atom Probe Tomography

The application of APT to the physics of reactions is demonstrated here with some studies with nanosized, man-made geometries. As an analysis with an atom probe requires appreciable effort, the method should be used especially in those cases where its particular advantages can be best utilized. To summarize, the outstanding features of APT include:

- A spatial resolution of chemical analysis of a few Ångstroms. While theoretical performance data expect a similar resolution to be achieved with analytical TEM [electron energy loss spectroscopy (EELS) or energy-dispersive X-ray spectroscopy (EDS)], a comparison of composition profiles at interfaces reveals that APT has a significantly higher discriminating power.

- A standard-free chemical analysis with identical sensitivity of all elements across the periodic table. Whilst in TEM studies a variety of species cannot be measured due to physical limitation [e.g. low-mass elements in energy-dispersive X-ray (EDX) or peak overlap and low intensity in EELS], APT will always reveal a chemical contrast that even allows different isotopes to be separated.
- A real chemical mapping in three dimensions. While the 2-D image projection required in high-resolution TEM studies hinders the clear characterization of complex morphologies, APT is especially suited to investigate curved and buried interfaces in nanocrystalline matter.

The following examples have been selected to illustrate, in which way these advantages may be decisive in the success of experimental studies.

7.5.1 Nucleation of the First Product Phase

Owing to the technological trend towards miniaturization, the very early stages of reactive diffusion at thin-film interfaces have shifted into the focus of materials research. Frequently, it is argued that the thermodynamic driving force of forming a first reaction product is usually so high that the critical thickness of nucleation ranges down below the size of a lattice constant. Thus, nucleation should not be a rate-controlling step at all. However, initial evidence that this may not be true came from calorimetric studies of reactive diffusion in metallic thin films. Although only a single product forms, double-peaked heat releases were observed [41–43]. This experimental finding was interpreted as a two-stage mechanism [44]. In the first stage, nuclei form at the initial interface and grow quickly in lateral directions, while in the second stage heat release is attributed to parabolic thickness growth by volume diffusion. However, the process of nucleation remained quite unclear. Several mechanisms have been proposed to explain the apparent reduction of driving force in the presence of a sharp composition gradient [45–47]. In common, they predict that the composition gradient at the interface must first decrease by interdiffusion to a critical level before nucleation of the first intermetallic compound becomes possible. However, no clear experimental verification of this interpretation was provided.

A recent nanoanalytical study [48] was aimed at enlightening details of the early nucleation stages. For these experiments, bilayers of Co and Al, each 20–30 nm thick, were deposited on tungsten substrate tips, as described in Section 7.4.3. Two examples of 3-D reconstructions of the Co/Al interface are shown in Figure 7.18. Although the layers were deposited on curved surfaces, the initial interface appeared practically flat, as the radius of curvature was still significantly larger than the width of the analyzed volumes. This flat interface is preserved for short annealing so that the earliest reaction stages may be characterized by 1-D composition profiles determined normal to the interface, as shown in Figure 7.19. Due to the outstanding resolution, minor chemical modifications at the interface become noticeable. After a 5 min period of heat treatment at 300 °C, the zone of chemical transition at the interface

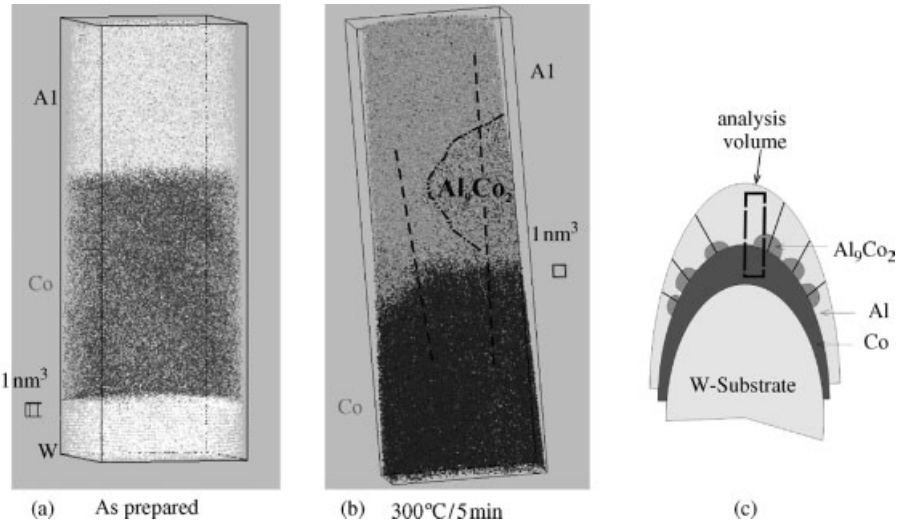


Figure 7.18 (a, b) Atomic reconstructions of the Al/Co interface in the as-prepared state (a) and after annealing at 300 °C for 5 min; (b) Positions of individual atoms are marked by gray coded dots; (c) Sketch of specimen geometry. (After [48]).

broadened to 3.5 nm, indicating a significant mixing of the components. However, the composition profile in this annealing state was well fitted by an error function. Thus, it follows that, up to this stage, only interdiffusion rather than formation of a new intermetallic product has taken place.

The first nucleation of a new phase is observed only in a small fraction of measurements at this annealing stage. In these cases, globular particles are detected at the interface towards the Al side (see Figure 7.18b). The fact that these particles

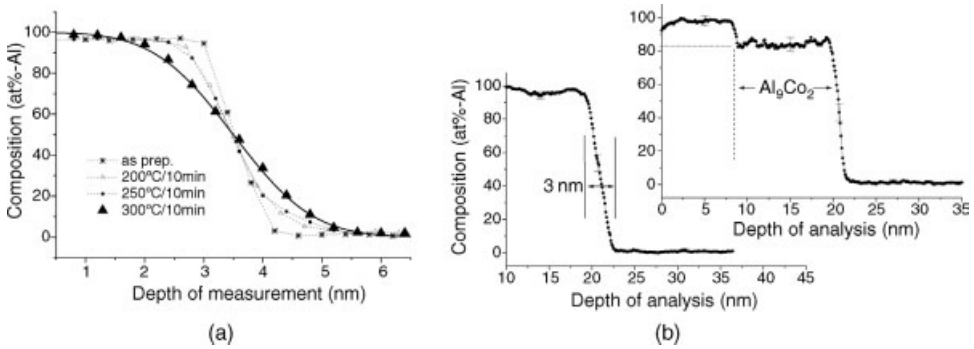


Figure 7.19 (a) Composition profiles determined perpendicular to the initial Al/Co interface after different annealing treatments; (b) Composition profiles determined along the left and right dashed lines in Figure 7.18b, demonstrating interdiffusion and phase formation, respectively. (After [48]).

appear only in part of the measurements after identical annealing conditions emphasizes the statistical nature of a nucleation process. Furthermore, it becomes clear from the volume reconstructions that nucleation takes place at heterogeneous sites at the interface, as sketched in Figure 7.18c. A composition profile across the newly formed phase identifies the product as Al_9Co_2 (see Figure 7.19b), while a profile determined across the remaining unreacted interface confirms again the interdiffusion on a depth of 3 to 4 nm as described previously.

The set of composition profiles in Figure 7.19, which could be obtained in this way only by APT, provide a clear demonstration that significant interdiffusion takes place before nucleation of the product. Furthermore, the experimental data quantify the critical diffusion depth before onset of nucleation to 3.5 nm (Al/Co at 300 °C).

If the theoretical nucleation thickness of the intermetallic product is estimated by the balance between the volume driving force and interfacial energy, a value of $d = 2\sigma/g_v = 0.2$ nm is expected. In view of this small value, it is very surprising, that the intermetallic Al_9Co_2 is only formed after the intermixed zone has already reached a thickness of 3.5 nm. However, based on the presented measurements, a theoretical study [49] was able to demonstrate that this behavior is clearly consistent with a polymorphic nucleation mechanism. Such a mechanism assumes that the nucleus of the new phase is produced by transforming the lattice structure, without modifying the local composition [45]. As any nucleus must have a minimum size in order to overcome the nucleation barrier, the ideal stoichiometric composition is only established in the center of the nucleus, whereas towards its boundaries the composition must deviate due to the existing concentration gradient. In other words, nucleation is only probable within a thin-layer fraction of the total diffusion zone, and the thickness of this layer shrinks with the chemical sharpness of the interface. In consequence, high concentration gradients will prevent nucleation.

In the cited theoretical study [49], the free energy ΔG required to form a nucleus under the constraint of a sharp concentration gradient was calculated. Numerical results are presented in Figure 7.20 for three different widths of the interdiffusion

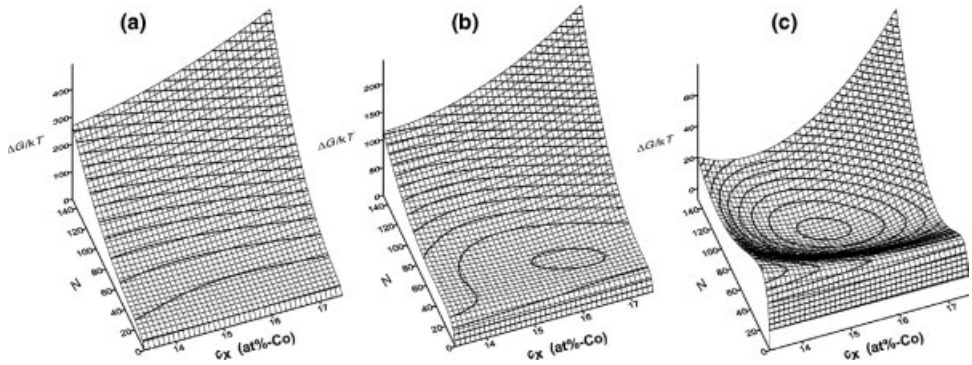


Figure 7.20 Surface $\Delta G(N, c_x)$ for the polymorphic transformation of a cubic volume into the Al_9Co_2 phase inside diffusion fields of width: (a) 3.0 nm; (b) 3.5 nm; and (c) 4.0 nm. Thermodynamic functions evaluated at a temperature of 573 K. (After [49]).

zone, namely 3.0, 3.5 and 4.0 nm. (In these plots, the size of the nucleus is expressed by the number N of atoms within the nucleus. The concentration c_x represents the composition at the center of nucleus.) For a diffusion width of 3.0 nm, the Gibbs free energy still increases monotonously with its size; thus, nucleation is forbidden. At 3.5 nm width, the situation has already slightly changed, with a weak local minimum appearing in the energy landscape. For 4.0 nm width, this minimum has become pronounced and its magnitude negative, which means that a product particle may now be formed under gain of energy. A nucleation barrier of 25 kT is determined from the energy landscape, which is a quite realistic value to obtain reasonable nucleation rates. Thus, the critical interdiffusion width is predicted to be slightly larger than 3.5 nm, in remarkably good agreement with the experimental observation by APT.

The same considerations were also performed for other suggested nucleation mechanisms. For example, calculation of the so-called 'transversal mode' [46] yielded a critical diffusion width smaller than 1.0 nm, in obvious contrast to the measurements. In conclusion, the atom probe analysis yielded convincing evidence for the interdiffusion process taking place before nucleation of the first product, although the depth of mixing ranged down to only a few nanometers. By using APT it has been possible to determine the critical diffusion depth with better than 1 nm accuracy, which allows a distinction to be made between different suggested nucleation modes.

Other studies on reactive metallic model systems made particular use of the calibration-free analysis of the atom probe. This chemical accuracy becomes especially important if metastable phases that spread over only two to three lattice constants are to be identified. Under these circumstances, a structural image is rarely achieved by high-resolution (HR) TEM, although the phases may be already clearly characterized by the local level of composition determined with APT, as for example demonstrated in studies on Ni/Al [50] and Ag/Al [51].

7.5.2
Thermal Stability of Giant Magnetoresistance Sensor Layers

In recent years, reading heads based on the giant magnetoresistance (GMR) effect have made possible a dramatic increase in magnetic recording density. As the periodicity of the required multilayers ranges down to a few nanometers, their spatially resolved analysis is a challenge, even with APT. Hence, several groups have used APT to characterize the as-produced state of GMR sensor layers [52–55]. Co/Cu and Cu/$Ni_{79}Fe_{21}$ are two of the most often-used metallic systems. The soft magnetic alloy $Ni_{79}Fe_{21}$ (Permalloy, Py) is especially suited for position sensors, as the effect of hysteresis is very low, while Cu/Co is mostly used for reading heads in data storage. For many potential applications, including for example angular sensors in motor vehicles, the thermal stability of the device is an important issue. It is well known that the Cu/Py system is much more sensitive to a thermal load than is Cu/Co. Whereas for the latter the GMR amplitude remains stable up to 400 °C, in the former case the GMR effect begins to degenerate at temperatures of only 150 °C [56]. As both metallic systems are immiscible in a thermodynamic sense, and furthermore all diffusivities

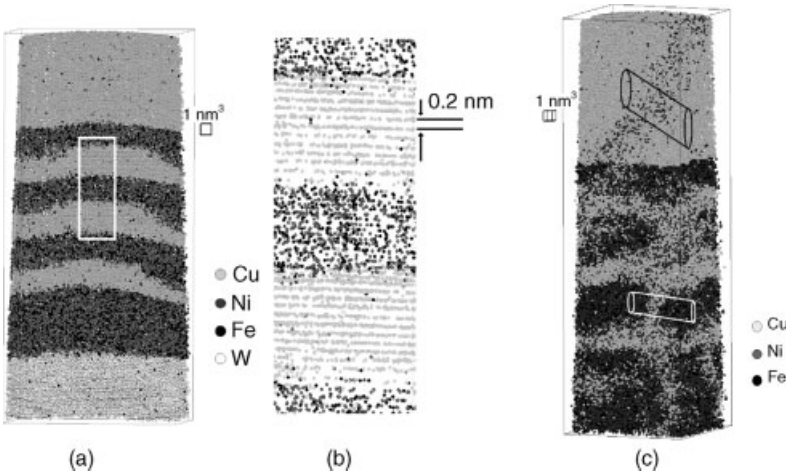


Figure 7.21 Atom probe tomography at a Cu/NiFe giant magnetoresistance structure. (a,b) In the as-prepared state; (c) After 30 min annealing at 400 °C. The detail (b) demonstrates lattice plane resolution and the outstanding chemical accuracy that allows the reproduction of sharp transitions in concentration from plane to plane. (After [58]).

are quite comparable, the low-temperature degradation of Cu/Py is surprising. In general, different mechanisms have been proposed as being responsible for GMR degradation: Van Loyen *et al*. [57] argued that at least two effects should contribute, namely grain boundary diffusion and inter- or demixing at the interface. In contrast, Hecker *et al*. [56] concluded that the alloying tendency of Ni and Cu above 250 °C controlled the decay of GMR in the Py systems.

In an extended APT study [58], $Cu_{2.5nm}/Py_{2.5nm}$ multilayers were deposited onto substrate tips and annealed in an UHV furnace. A typical volume reconstruction of an as-prepared state is shown in Figure 7.21a. The resolution is sufficient to distinguish individual (111) planes of Cu (Figure 7.21b), so that the dimensions of the reconstructed volume could be calibrated exactly. If the microstructures of the as-prepared state were compared to those after annealing up to 350 °C, then no remarkable difference was seen at first sight. Notably, the integrity of the multilayer was preserved. This was all the more striking as the magnetoresistivity vanished completely at the lower temperature of 150 °C. Only after annealing at even higher temperatures – at which the GMR effect also vanished in the more stable Cu/Co structure (400 °C) – was any clear indication of grain boundary transport observed (as shown in Figure 7.20c).

As an advantage of the 3-D experimental data, concentration profiles can be evaluated by defining analysis cylinders of smaller diameter and exactly aligning them perpendicular to the interfaces. In this way, artifacts due to local curvature or roughness of interfaces can be mostly excluded. An exemplary profile obtained after annealing at 350 °C is presented in Figure 7.22a, where the sharpness of the chemical

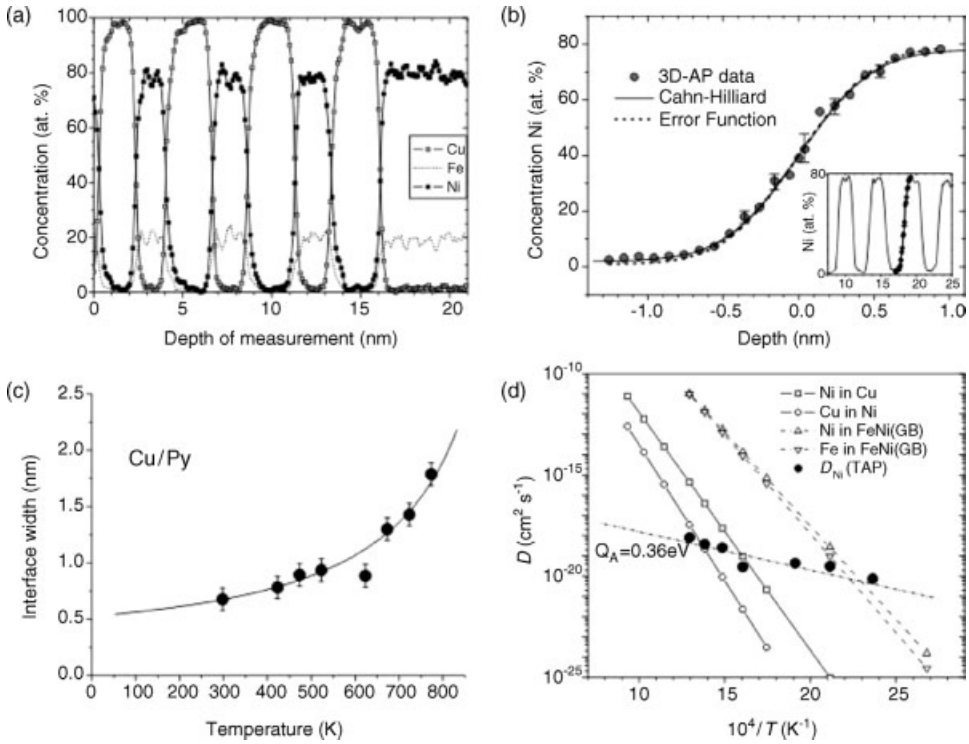


Figure 7.22 Nanoanalysis of Cu/Py multilayers. (a) Composition profile after 20 min annealing at 350 °C; (b) Error-function shape of chemical transition at interface; (c) Development of interfacial width with annealing temperature solid line represents prediction by Cahn–Hilliard theory. (d) Interdiffusion coefficient determined from this width in comparison to published data. (Reproduced with permission from Ref. [59]; © Carl Hanser Verlag 176.)

transition from almost 0% to 100 at% Cu on the length of two lattice plane distances, should be noted. Also of note, it is not possible to achieve such selectivity with analytical TEM.

Furthermore, the shape of the chemical transition can be characterized in detail and compared with expected model curves, as for example for interdiffusion (error-function) and interfacial thermodynamics (Cahn–Hilliard) shown in Figure 7.22b. Due to the outstanding sensitivity of APT, it is possible to note the smallest modifications of interfacial chemistry with temperature. The width of the chemical transition, defined exactly between 10% and 90% of the concentration amplitude, was used as a characteristic parameter. Clearly, although only on the order of 1 nm, the width is proven to increase significantly during low-temperature annealing (see Figure 7.22c). The formal description by an error function leads to the assumption that the width of the transition is controlled by kinetics of interdiffusion, as in the previously discussed study on Al/Co (see Section 7.5.1). However, the diffusion

coefficients derived in accordance with this interpretation reveal a way too-low activation energy in comparison to reported data (see Figure 7.22d). Furthermore, interdiffusion would be difficult to understand from the viewpoint of thermodynamics, as the layer system is expected to be stable up to a temperature of about 1070 K. In an analysis of these data [59], it could be shown that the temperature dependence of the interface width is rather based on interfacial thermodynamics within the frame of Cahn–Hilliard theory. The experimentally observed temperature dependence of the width of the interface agrees nicely with the related model curve, as presented in Figure 7.22c. In this way, it is possible to determine, via direct measurement with APT, the required – but experimentally almost unknown – gradient coefficients that are important in the Cahn–Hilliard theory.

By combining wide field-of-view FIM and atom probe analysis, it was also possible [58] to exclude definitely grain boundary transport as a significant reason for GMR degradation in Cu/Py multilayers. As shown in Figure 7.21c, segregation – as an indicator for grain boundary diffusion – can be demonstrated in the volume reconstructions, but clearly not at temperatures relevant to GMR breakdown. Many individual grain boundaries were investigated after annealing at up to 250 °C, but no segregation was found in any case.

For the design of GMR devices, the APT study has important consequences. Clearly, a short-ranged broadening of interfaces on the depth of 1–2 nm due to interfacial thermodynamics (which has been neglected in previous studies) has an important influence on magnetoresistivity. In view of the small multilayer periodicity of only 4 nm, this does not come as a surprise. It is, furthermore, in complete agreement with measurements of total electrical resistivity [56], which had shown that the base resistance of the multilayer, indicating complete alloying of the layers, increased significantly only at temperatures higher than 250 °C. Thus, only a weak interfacial alloying could take place at the relevant conditions below that temperature. In order to develop temperature-stable devices, it is insufficient to select only a thermodynamic stable system. In addition, the critical temperature of the respective miscibility gap should be at least threefold higher than the application temperature to suppress the described interfacial broadening. As confirmation, a similar study with APT on Cu/Co (a system distinguished by a much higher critical temperature of about 1600 K) revealed no significant broadening of the interfaces up to a temperature of 450 °C. As a consequence, the degradation of the GMR effect was considered due to grain boundary diffusion [60].

7.5.3
Influence of Grain Boundaries and Curved Interfaces

In nanocrystalline matter with grain sizes down to about 10 nm, the volume fraction attributed to the grain boundary (GB) can easily exceed 50%. With further decreasing grain size an additional necessary topological feature of the boundary arrangement – the so-called 'triple line' – may affect atomic transport. Along a triple line, three GBs merge to form a line-shaped junction defect, the structure of which is expected to differ considerably from that of ordinary GBs. Most likely, it is more disordered, so

that a rather high atomic mobility can be assumed. However, until now the measurement of triple line diffusion using analytical electron microscopy has been rare [61]. The difficulty of such measurement stems from the small effective cross-section of the defect, which requires chemical analysis of the highest possible resolution.

In atom probe experiments [62], both Au (15 nm thickness) and Cu layers (25 nm thickness) were deposited on top of tungsten tips. In order to slow down the intermixing of the soluble metals Cu and Au, a thin Co barrier (6 nm thickness) was inserted in between. Due to the strong substrate curvature, thin films deposited on the substrate tips tend to be very fine-grained, with grain sizes down to 5 nm [63]. Thus, these specimens are ideal candidates to observe the transport along topological singularities of the GB arrangement. A typical field ion micrograph of the upper Cu layer is shown in Figure 7.23a. Discontinuities in the structure of concentric rings mark grain boundaries (some of which are emphasized by dashed lines in the figure). Clearly, a polycrystalline structure with a grain size of about 15 nm has formed. Triple junctions are also seen, aligned approximately perpendicular to the surface so that they can be analyzed along the tip axis; in the figure a particularly clear junction is marked by a "T".

By selecting suitable areas for analysis, the local concentration field around the junction could be measured. The geometry of three GBs and an exemplary 2-D composition map is shown in Figure 7.23b. As the atom probe delivers 3-D data, 2-D composition maps can be evaluated in any arbitrary direction subsequent to the measurement, so that the penetration of solute into the triple line and the merging grain boundaries can be determined in a unique manner. The example 2-D map in the figure is aligned perpendicular to a triple line. Clearly, the line is locally enriched

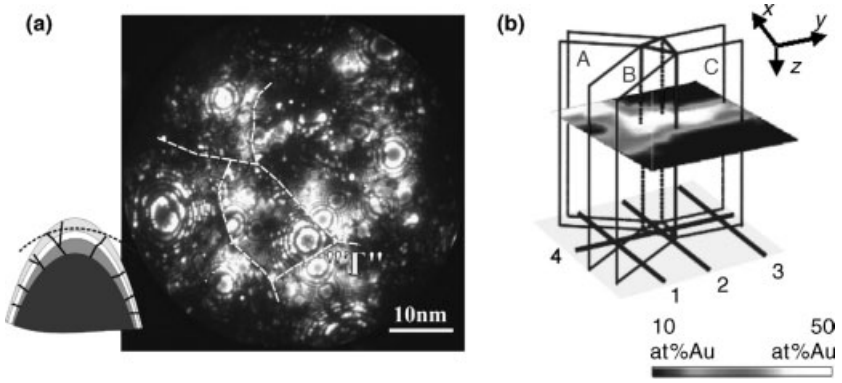


Figure 7.23 (a) Field ion micrograph of sputter-deposited Cu layer. The grain boundaries are marked by dashed lines, a triple junction by 'T'; (b) Two-dimensional composition map in gray-scale representation determined at a cross-section through a triple line and the related three boundaries (A, B, C). (After [62]).

in Au, and the three GBs are also distinguished by a measurable Au content, albeit of a somewhat lower level. The concentration fields around the triple line were evaluated by means of the approximate solution of transport equations proposed by Klinger *et al.* [64]. In the case of the triple line, the atomic transport may be understood by a three-level cascade process: (i) the material is transported along the triple line; (ii) leakage into the related three boundaries takes place; and (iii) atoms are drained from the grain boundaries into the bulk grain volume. Expressing this in a compact formula, the concentration field is described by:

$$c(x,y,z,t) = c_0 \cdot \exp\left(-\frac{\sqrt{3}\cdot\sqrt[4]{D_{GB}\cdot\delta\cdot s_{GB}}\cdot\sqrt[8]{4D_V/\pi t}}{\sqrt{D_{TL}\cdot q\cdot s_{TL}}}\cdot z\right)$$
$$\times \exp\left(-\frac{\sqrt[4]{4D_V/\pi t}}{\sqrt{D_{GB}\cdot\delta\cdot s_{GB}}}\cdot y\right) \qquad (7.19)$$
$$\times \exp\left(1-\mathrm{erf}\left[\frac{x}{2\sqrt{D_V t}}\right]\right)$$

(For a definition of the coordinates, see Figure 7.23b). If the volume diffusion coefficient D_V is known, the two other diffusion coefficients – or, more exactly, the respective transport products $p_{TL}:=D_{TL}\cdot q\cdot s_{TL}$ and $p_{GB}:=D_{GB}\cdot\delta\cdot s_{GB}$ – can be determined from logarithmic plots of composition versus penetration depth in the z- and y-directions. (δ, q, s_{TL} and s_{GB} define grain boundary width, effective cross-section of the triple line and the segregation factors of triple line and grain boundary, respectively.) Exemplary plots are presented in Figure 7.24, demonstrating the transport behavior in accordance with Equation 7.19. Indeed, by evaluating the slope of the straight lines in Figure 7.24, it was quantitatively found that D_{TL} is a factor 5600-fold larger than D_{GB} at a temperature of 295 °C, if potential segregation is neglected and $q=\delta^2$ is assumed.

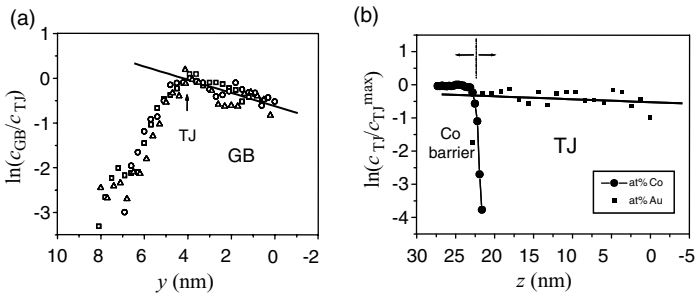


Figure 7.24 Au diffusion in Cu at 295 °C. Normalized concentration profiles along the (a) grain boundary (GB) and (b) triple junction (TJ), as determined with atom probe tomography. For details, see the text. (After [62]).

7.6
Approaching Nonconductive Materials: Pulsed Laser Atom Probe Tomography

On completing this survey of the atom probe method, an additional and quite important methodical aspect must be addressed. The following point could have been presented earlier, in Section 7.3, and frankly – from a logical standpoint – such an order would be more valid. However, by postponing it to the end of the chapter, our aim is to highlight the fact that pulsed laser evaporation represents the most active recent field in methodical research in FIM. Indeed, this technique promises not only to expose APT to a much broader range of materials, but also raises some interesting controversy on the fundamental mechanisms of field evaporation by using ultra-short light pulses of about 100 fs duration.

7.6.1
The Limitations of High-Voltage Pulsing

The conventional atom probe technique based on high-voltage pulsing has always been restricted to the investigation of sufficiently conductive materials, usually metals. Very few measurements of ceramic materials have been reported, and in all successful cases the nonconducting phases had the geometry of tiny particles or thin films embedded in a metallic matrix [65–67]. Besides the unfavorable mechanical properties of ceramics or semi-conductors, this situation is first and foremost due to the limited possibility of transferring a short pulse to the specimen surface. As illustrated in Figure 7.25, the arrangement of sample and detector represents nothing else but a resistive–capacative (RC) oscillator which is naturally limited in its transfer frequency. A few lines of calculation are sufficient to estimate the required minimum conductivity of the tip material to achieve a sufficient band width of the transfer line. According to Poisson's law, the field in a vacuum is related to a density σ of charges at the surface of the metallic electrode:

$$\sigma = \varepsilon_0 \cdot F = \varepsilon_0 \frac{V}{\beta R}, \tag{7.20}$$

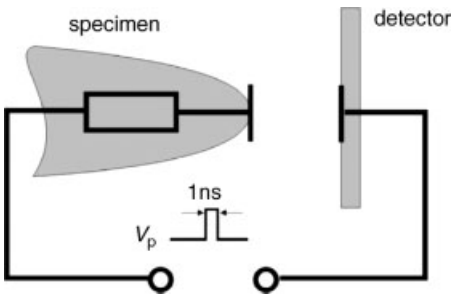


Figure 7.25 Electric equivalent circuit of the arrangement of sample and detector or vacuum chamber.

where we have made use of Equation 7.1 to formulate the second equation. Approximating the effective electrode surface by a semi-sphere, the electrical capacity of the arrangement reads

$$C = \frac{2\pi R^2 \sigma}{V} = \frac{2\pi R \varepsilon_0}{\beta}. \tag{7.21}$$

With regards to the specimen geometry ($R \approx 50\,\mathrm{nm}$), this capacity amounts to about 10^{-18} F. Likewise, the resistivity $R_{\mathrm{el.}}$ may be estimated by approximating the specimen shape with a truncated cone of length L and semi-shaft angle γ. So, one can formulate

$$R_{\mathrm{el.}} = \frac{\rho}{\pi} \int_0^L \frac{dl}{(\gamma \cdot l + R)^2} \approx \frac{\rho}{\pi} \cdot \frac{1}{\gamma \cdot R} \tag{7.22}$$

in which ρ denotes the specific resistivity of the material. In order to transfer pulses properly, the limiting frequency $\omega = 1/(R_{\mathrm{el.}}\, C)$ must reach 10^9 Hz, which requires the specific resistivity to be smaller than

$$\rho_c < \frac{\pi \gamma R}{\omega C} \approx 10^3\ \Omega\,\mathrm{cm}. \tag{7.23}$$

Clearly, the resistivity of metals is way below this limit, particularly at the low temperatures used for evaporation. However, already in the case of semi-conductors, sufficient conductivity is only achieved with very high doping. In the case of real insulators (such as oxides or nitrides) the condition of Equation 7.23 is failed by orders of magnitude in any case, leaving very little hope to perform successful measurements by electric pulsing.

Additional complications arise from the brittleness of the materials. If we multiply (for an estimate to order of magnitude) the charge density (Equation 7.20) by the field (Equation 7.1), then tensile stresses of up to GPa magnitude are predicted to be induced by the field. In the case of high-voltage pulsing, a significant fraction of this stress is supplied to the tip as a continuous sequence of mechanical shocks resembling a classical fatigue test. Considering that a stress of 1 GPa exceeds the fracture strength of many bulk materials, it is clear that specimens will fail frequently by fracture during the 'fatigue test' of the measurement, and that the risk of this failure increases with the brittleness of the tip material.

7.6.2
The Mechanism of Pulsed Laser Evaporation

With regards to Equation 7.6 (see Section 7.3), it is possible to imagine at least two alternative routes to trigger field evaporation. Either the numerator in the argument of the exponential could be reduced by a short increase of the electric field (as is done with conventional high-voltage pulsing), or the denominator may be increased by short rises in temperature. The latter can be achieved with the light flashes delivered

by pulsed laser sources, so it is not surprising that attempts to utilize this effect date back to the 1980s. The early studies with pulsed laser sources were motivated by studies of the photoionization of adsorbed species [68, 69], but shortly afterwards the potential to control field evaporation was also acknowledged and the pulsed laser atom probe (PLAP) introduced by Kellog and Tsong [70]. In these early experiments, nitrogen lasers of 1 ns pulse width (comparable to the length of high-voltage pulses) were normally used. The authors stated the following as the main advantages of the laser evaporation mode:

- As no fast charge transport is required, a low conductivity of the sample no longer represents a bottleneck to the atom probe analysis. The controlled field evaporation of intrinsic Si could be demonstrated.
- As the remaining field is supplied in dc mode, the PLAP does not suffer from the energy spread which is unavoidable under high-voltage pulsing. Thus, a much better mass resolution could be achieved, even with straight flight tubes, without using any time-focusing devices.
- The level of mechanical stress is reduced by the fraction that has been formerly induced by the high-voltage pulses. The remaining stress is produced by the constant base voltage, and thus loaded to the tip permanently. This is a much more favorable situation for the specimen life time than the above-mentioned 'fatigue test' condition.

The time resolution – and thus the accuracy of the energy measurement – was improved by Tsong and coworkers to the order of 10^{-5}. So, by careful measurement of the energy deficits of evaporating species, the mechanism of desorption could be identified [71]. For laser pulses of nanosecond duration, it could be shown that with metals the laser effect was merely thermal in nature, whereas in the case of semiconductors (which were distinguished by a significant band gap) the additional influence of direct photoionization was indicated. As the laser acted predominantly as a pulsed heat source, the question immediately arose as to whether heating of the specimen surface might diminish the spatial resolution of the atom probe analysis. However, already Tsong and coworkers had already shown that it was possible to trigger the evaporation reliably by temperature pulses below 200 K, at which surface diffusion remained frozen, at least for the refractory metals being studied.

Until 1990, the advantages of the pulsed laser mode were exploited not only in investigations of Si, SiC and SiO but also of Group III–V semiconductor materials. However, as these studies were mostly restricted to verify mass spectra and to prove the average composition of the materials, the total number of evaluated atoms was consequently very low. Except for some interesting studies on the oxidation of Si [72], very few attempts were made at a spatially resolved analysis. In this context, it is quite remarkable that reports on the different versions of 3-D atom probes which emerged at that time made no mention of any attachment of a laser beam line. Apparently, only the group at Oxford University attempted to use the PoSAP instrument with laser assistance to perform for example, a study on GaInAs quantum well structures [73]. However also in this case the fact that no volume reconstructions were presented

may be seen as an indication of the significant experimental problems that the group encountered. Compared to the vast amount of successful reports on conductive materials with high-voltage pulsing, the pulsed laser technique virtually disappeared during the 1990s. It might be speculated that this situation was essentially caused by the limited reliability of available laser systems and the involved alignment procedure.

Fortunately, this situation has changed remarkably since 2002, and many university research groups and commercial companies are now engaged in the development of atom probe instrumentation, both in terms of testing and offering instruments with the facility of laser-assisted evaporation. Today, commercial laser systems are much more stable, provide pulse frequencies, and have found vast improvements in the ease of operation. Moreover, continuous miniaturization in microelectronics has led to an enormous driving force to overcome any remaining barriers when developing new tools for high-resolution semiconductor analysis.

The first experiments with modern sub-picosecond laser pulses were reported by Gault *et al.* in 2005 [74]. These authors showed that the efficiency of laser pulsing depended on the orientation of light polarization with respect to the tip axis, and claimed this to be evidence of a direct, athermal influence of the optic wave. In other words, similar to high-voltage pulsing, the activation barrier in the numerator of Equation 7.6 should be temporarily decreased by the electrical field of the wave. This idea was very attractive, as in this way laser-triggered evaporation would become possible without heating the specimen. Moreover, since with a decreasing pulse width at constant pulse energy the electric field amplitude of the optical wave increased, this might indeed be the case. However, the average field strength of the laser beam in the experiments performed by Gault and coworkers was still in the range of only $0.1\,\mathrm{V\,nm^{-1}}$, which was much less than the effect of typical high-voltage pulses. Nonetheless, one might expect a field enhancement due to polarization of the metallic tip. In theoretical studies, enhancement factors of up to three orders of magnitude have been predicted [75–77], with the fields becoming quite capable of having a significant effect on the evaporation barrier in Equation 7.6. A further conceptual difficulty arises from the fact that optical fields oscillate with a frequency of 10^{15} Hz, while the spectrum of atomic vibration is limited at about 10^{13} Hz. So, how might an atom be excited way off resonance? The group of Gault argued that the rapidly oscillating field could be converted into a directed field for the duration of the pulse by a nonlinear response, in the same way that a diode might be used to demodulate high-frequency radio signals. Indeed, in recent theoretical calculations proof was furthered that this 'optical rectification' would indeed produce field pulses of an appropriate order of magnitude and polarity [78]. In addition, optical pump probe experiments demonstrated a rapid sub-picosecond response, which has been interpreted in the same direction [79].

On the other hand, meanwhile, it became clear that the interpretation of measurements in terms of a clear athermal triggering by Gault *et al.* [74] was too euphoric, and their evidence less clear [80]. The dependence of evaporation rate on the polarization of the wave could be naturally explained by scattering at the nanometer-sized, cylindrically shaped tip, as had been shown earlier [81]. The measurement of the

temperature rise produced by sub-picosecond pulses, although performed with a remarkable pump probe experiment that combined laser and voltage pulses [82], delivered an ambivalent result. Yet, the idea of achieving direct-field pulsing by using ultra-short laser pulses proved so attractive that considerable effort is currently being undertaken to clarify this aspect, notably by the group at the University of Rouen.

Today, several other teams are focusing on the practical consequences of laser pulsing to concrete analytical studies. With laser pulsing, several new parameters must be explored to define the optimum measurement conditions. An extended study [83] has investigated the influence of wavelength and energy density of the pulses, laser spot size, specimen geometry, and heat conductivity on the quality of mass spectra. Yet, interestingly, no significant influence of pulse width and wavelength on the quality of mass spectra was identified. The mass resolution achieved with laser pulsing is indeed significantly improved in comparison to high-voltage pulsing (up to $\Delta m/m = 1:1000$, without using a reflectron). However, significant thermal peak tails are observed in the mass spectra, with rather long laser pulses exceeding the important picosecond benchmark of electron phonon coupling, as well as with short pulses way below this threshold. Instead of the pulse width, the duration at which the ions are evaporated is essentially defined by the cooling period of the specimen after the pulse. Thus, samples of low heat conductance or of particularly long and thin geometry lead to poor mass resolution. With materials of low heat conductivity, such as stainless steels, a small heat spot achieved with an optimally focused laser beam has proved to be advantageous when achieving rapid cooling rates, and thereby good mass resolution. Remarkably, also in the case of a femtosecond laser, the tails of the mass peaks are correctly described by the principles of heat conduction, which proves that also with this type of laser a considerable temperature rise appears during the pulse.

In summary, it can be stated that today, a final conclusion of whether short pulses in the range of 100 fs provide any significant advantage in the analysis cannot be drawn. Although, in various reports, the impression is sometimes raised that femtosecond pulses are required for the reliable analysis of difficult materials (such as the measurement of oxides [84]), the opposite can be clearly demonstrated [85].

7.6.3
Application to Microelectronic Devices

Unaffected by the scientific controversy surrounding the mechanism of laser-assisted evaporation, the 3-D atom probe has advanced in recent years to a modern measurement tool, and is set to become an established component of those industrial laboratories involved with microelectronics. While the structural width of a transistor now ranges down to 35 nm, the measuring field of the atom probe has reached approximately 100 nm. This almost perfect matching of length scales has motivated the industrial application of the method. With the number of evaluated atoms exceeding 100 million, APT now permits the mapping of dopant distributions, in outstanding resolution, where the chemical sensitivity of analytical TEM is usually way too low (see e.g. [86]).

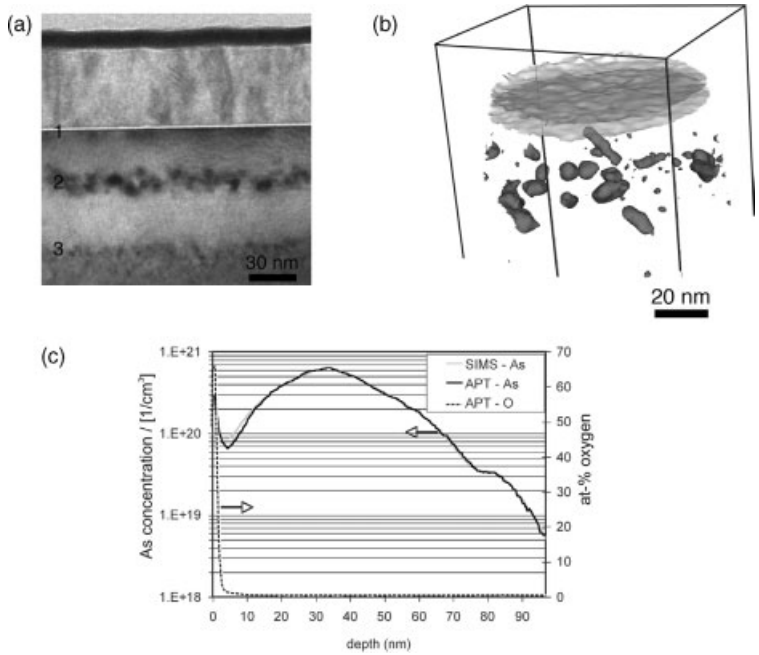


Figure 7.26 Analysis of As doping of Si by ion implantation. (a) Cross-section TEM image of the test structure showing an oxide layer (1), As-rich particles (2), and the end of damage range (3). The ion irradiation was apparent from the top; (b) Volume analysis by APT showing iso-concentration surfaces at 10 at.% oxygen (light gray) and 2 at.% As (dark gray); (c) Comparison of atom probe data with SIMS measurement. (Reproduced with permission from [87]).

These findings are illustrated by the results of the study shown in Figure 7.26. Here, a {001} Si wafer was implanted with 2×10^{15} As atoms per cm^2 with 30 keV kinetic energy. An oxide film of 2 nm thickness and 50 nm of undoped, polycrystalline Si were then deposited, to mimic the basic steps of microelectronics production. While forming the polycrystalline Si, the specimen had to be heated to 600 °C for about 30 min, so that the implanted As atoms could cluster. As shown in the TEM image (Figure 7.26a), a defect band was formed at approximately 30 nm beneath the substrate surface (this is localized in the figure by the thin light oxide layer). For atom probe analysis, tip-shaped samples were prepared by FIB milling. With pulsed laser assistance, the semiconductor sample could be properly evaporated, and from the volume reconstruction local composition maps calculated which allowed the compositional heterogeneities to be located by means of iso-concentration surfaces (Figure 7.26b). At a preset concentration level of 2 at.% As, these iso-surfaces mark As-rich clusters, which have formed by nucleation. The iso-surfaces at a level of 10 at.% oxygen were used to localize the interfaces to the oxygen layer. Quantitative composition profiles aligned perpendicular to the substrate surface could also be determined (as shown in Figure 7.26c) in

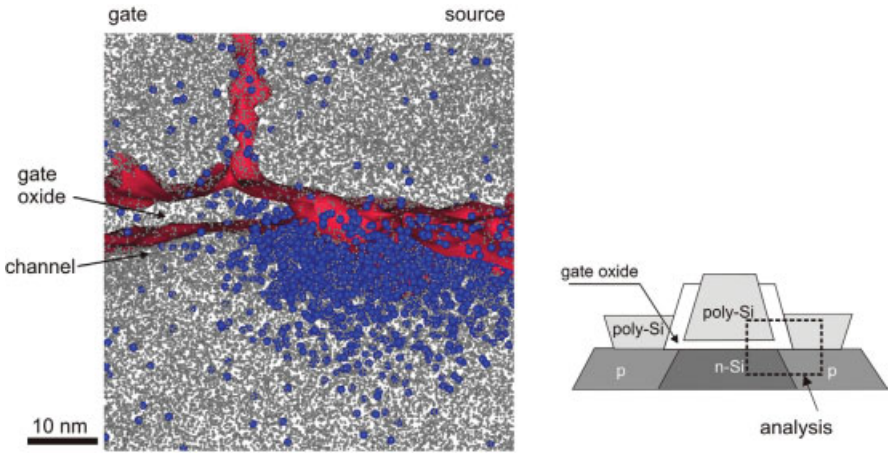


Figure 7.27 Three-dimensional reconstruction of a field effect transistor structure showing Si atoms (light gray dots) and boron dopant (blue). Iso-concentration surfaces at 3 at% oxygen localize the gate oxide film (red). (Illustration courtesy of D.J. Larson, Imago Scientific Instruments.) The schematic at the right clarifies the location of the analyzed region.

comparison to depth profiling by secondary ion mass spectrometry (SIMS), which represents the industrial standard. Clearly, the atom probe delivers quantitative data in agreement with SIMS, but more importantly the APT data proved to be more accurate. While the segregation of As to the substrate surface is indicated only faintly by SIMS, the corresponding composition peak appears very pronounced in the APT measurement, due to a significantly better spatial resolution of the latter technique. As the atom probe is equally sensitive to all chemical species, the oxygen content of the oxide is determined within the same measurement.

The final example refers to the preparation of tip samples of technical devices taken from industrial production. By using modern, dual-beam FIB, areas of interest can be selected for TAP measurement. In this way, the corner region of a field effect transistor – between the source contact and the channel beneath the gate contact – has been analyzed with regards to local dopant and oxygen distribution (see Figure 7.27). A few years ago, the achievement of such an analysis from a microchip was beyond thinking. Having mapped the spatial arrangement of the atoms, not only the depth profile beneath the source or drain contacts of the field effect transistor could be determined, but also the lateral distribution of the dopant. In this way, essential information can be obtained concerning dopant diffusion from the contact region into the channel under the gate oxide. As the structural width of transistors continues to decrease, the control of this sideward transport may in particular become a critical issue. It is likely that, in future, APT will become the most important tool for controlling this undesired process.

References

1 Müller, E.W. (1951) *Zeitschrift für Physik*, **131**, 136.

2 Kelly, T.F. and Miller, M.K. (2007) *Review of Scientific Instruments*, **78**, 031101.

3 Cerezo, A., Clifton, P.H., Galtrey, M.J., Humphreys, C.J., Kelly, T.F., Larson, D.J., Lozano-Perez, S., Marquis, E.A., Oliver, R.A., Sha, G., Thompson, K. and Zandbergen, M. (2007) *Materials Today*, **10**, 1.

4 Miller, M.K., Cerezo, A., Hetherington, M.G. and Smith, G.D.W. (1996) *Atom Probe Field Ion Microscopy*, Oxford Science Publications, Oxford.

5 Miller, M.K. (2000) *Atom Probe Tomography*, Kluwer Academic, New York.

6 Tsong, T.T. (1990) *Atom-Probe Field Ion Microscopy*, Cambridge University Press, Cambridge.

7 Müller, E.W. (1957) *Journal of Applied Physics*, **28**, 1.

8 Müller, E.W., Panitz, J.A. and McLane, S.B. (1968) *Review of Scientific Instruments*, **39**, 83.

9 Cerezo, A., Godfrey, T.J. and Smith, G.D.W. (1988) *Review of Scientific Instruments*, **59**, 862.

10 Deconihout, B., Bostel, A., Menand, A., Sarrau, J.M., Bouet, M., Chambreland, S. and Blavette, D. (1993) *Applied Surface Science*, **67**, 444.

11 Sijbrandij, S.J., Cerezo, A., Godfrey, T.J. and Smith, G.D.W. (1996) *Applied Surface Science*, **94–95**, 428.

12 Panayi, P. (2006) Great Britain Patent Application No. GB2426120A, November 15.

13 Kelly, T.F., Gribb, T.T., Olson, J.D., Martens, R.L. *et al.* (2004) *Microscopy and Microanalysis*, **10**, 373.

14 Tsong, J. (1978) *Journal of Physics F: Metal Physics*, **8**, 1349.

15 Bas, P., Bostel, A., Deconihout, B. and Blavette, D. (1995) *Applied Surface Science*, **87/88**, 298.

16 Al-Kassab, T., Wollenberger, H., Schmitz, G. and Kirchheim, R. (2003) *High Resolution Imaging and Spectrometry of Materials* (eds T. Ernst and M. Rühle), Springer, Berlin, p. 290.

17 Schmitz, G. and Howe, J.M. (2007) High Resolution Microscopy, in *Alloy Physics* (ed. W. Pfeiler), Wiley, pp. 774–860.

18 Hellman, O. and Seidman, D. (2000) *Microscopy and Microanalysis*, **6**, 437.

19 Warren, P.J., Cerezo, A. and Smith, G.D.W. (1998) *Ultramicroscopy*, **73**, 261–266.

20 Vurpillot, F., da Costa, G., Menand, A. and Blavette, D. (2001) *Journal of Microscopy*, **203**, 295.

21 Geiser, B.P., Kelley, T.F., Larson, D.J., Schneir, J. and Roberts, J.P. (2007) *Microscopy and Microanalysis*, **13**, 437–447.

22 Vurpillot, F., Bostel, A., Cadel, E. and Blavette, D. (2000) *Ultramicroscopy*, **84**, 213.

23 Vurpillot, F., Bostel, A. and Blavette, D. (2001) *Ultramicroscopy*, **89**, 137.

24 Cerezo, A., Godfrey, T.J., Hyde, J.M., Sijbrandij, S.J. and Smith, G.D.W. (1994) *Applied Surface Science*, **76/77**, 374.

25 Deconihout, B., Renaud, L., Da Costa, G., Bouet, M., Bostel, A. and Bavette, D. (1998) *Ultramicroscopy*, **73**, 253.

26 Cerezo, A., Godfrey, T.J. and Smith, G.D.W. (1988) *Review of Scientific Instruments*, **59**, 862.

27 Keller, H., Klingelhöfer, G. and Kankelheit, E. (1987) *Nuclear Instruments & Methods*, **A258**, 221.

28 Jagutzki, O., Cerezo, A., Czasch, A. *et al.* (2002) *IEEE Transactions on Nuclear Science*, **49**, 2477.

29 da Costa, G., Vurpillot, F., Bostel, A., Bouet, M. and Deconihout, B. (2005) *Review of Scientific Instruments*, **76**, 013304.

30 Deonihout, B., Vurpillot, F. and Gault, B. (2007) *Surface and Interface Analysis*, **39**, 278.

31 Stender, P., Oberdorfer, C., Artmeier, M. and Pelka, P. (2007) *Ultramicroscopy*, **107**, 726–733.

32 Nishikawa, O. and Kimoto, M. (1994) *Applied Surface Science*, **76/77**, 424.

33 Kelly, T.F., Gribb, T.T., Olson, J.D., Martens, R.L. *et al.* (2004) *Microscopy and Microanalysis*, **10**, 373.

34 Miller, M.K. and Russel, K.F. (2007) *Ultramicroscopy*, **107**, 761–766.

35 Perea, D.E., Allen, J.E., May, S.J., Wessels, B.W., Seidman, D.N. and Lauhon, L.J. (2006) *Nano Letters*, **6**, 181.

36 Schleiwies, J. and Schmitz, G. (2002) *Materials Science and Engineering A*, **327**, 94.

37 Lang, C. and Schmitz, G. (2003) *Materials Science and Engineering A*, **353**, 119.

38 Hono, K., Hasegawa, N., Okano, R., Fujimori, H. and Sakurai, T. (1993) *Applied Surface Science*, **67**, 407.

39 Larson, D. (2001) *Microscopy and Microanalysis*, **7**, 24.

40 Giannuzi, L.A. and Stevie, F.A. (1999) *Micron*, **30**, 197.

41 Michaelsen, C., Barmak, K. and Weihs, T.P. (1997) *Journal of Physics D - Applied Physics*, **30**, 3167.

42 Roy, R. and Sen, S.K. (1992) *Journal of Materials Science*, **27**, 6098.

43 Bergmann, C., Emeric, E., Clugnet, G. and Gas, P. (2001) *Defect and Diffusion Forum*, **194–199**, 1533.

44 Coffey, K.R., Clevenger, L.A., Barmak, K., Rudman, D.A. and Thompson, C.V. (1989) *Applied Physics Letters*, **55**, 852.

45 Gusak, A.M. (1990) *Ukrainian Journal of Physics*, **35**, 725.

46 Desré, P.J. and Yavari, R. (1990) *Physical Review Letters*, **64**, 1533.

47 Hodaj, F. and Desré, P.J. (1996) *Acta Materialia*, **44**, 4485.

48 Vovk, V., Schmitz, G. and Kirchheim, R. (2004) *Physical Review B - Condensed Matter*, **69**, 104102.

49 Pasichnyy, M.O., Schmitz, G., Gusak, A.M. and Vovk, V. (2005) *Physical Review B-Condensed Matter*, **72**, 014118.

50 Jeske, T. and Schmitz, G. (2001) *Scripta Materialia*, **45**, 555.

51 Schleiwies, J. and Schmitz, G. (2002) *Materials Science and Engineering A*, **327**, 94.

52 Schleiwies, J., Schmitz, G., Heitmann, S. and Hütten, A. (2001) *Applied Physics Letters*, **78**, 3439.

53 Zhou, X.W. *et al.* (2001) *Acta Materialia*, **49**, 4005.

54 Larson, D.J., Cerezo, A., Clifton, P.H., Petford-Long, A.K., Martens, R.L., Kelly, T.F. and Tabat, N. (2001) *Journal of Applied Physics*, **89**, 7517.

55 Larson, D.J. *et al.* (2003) *Physical Review B - Condensed Matter*, **67**, 144420.

56 Hecker, M., Tietjen, D., Wendrock, J., Schneider, C.M., Cramer, N. and Malinski, L. (2002) *Journal of Magnetism and Magnetic Materials*, **247**, 62.

57 van Loyen, L., Elefant, D., Tietjen, D., Schneider, C.M., Hecker, M. and Thomas, J. (2000) *Journal of Applied Physics*, **87**, 4852.

58 Ene, C.B., Schmitz, G., Kirchheim, R. and Hütten, A. (2005) *Acta Materialia*, **53**, 3383.

59 Stender, P., Ene, C.B., Galinski, H. and Schmitz, G. (2008) *International Journal of Materials Research*, **99**, 480.

60 Vovk, V. and Schmitz, G., Ultramicroscopy (inpress).

61 Bokstein, B., Ivanov, V., Oreshina, O., Pteline, A. and Peteline, S. (2001) *Materials Science and Engineering A*, **302**, 151.

62 Schmitz, G., Ene, C., Lang, C. and Vovk, V. (2006) *Advanced Science and Technology*, **46**, 126.

63 Lang, C. and Schmitz, G. (2003) *Materials Science and Engineering A*, **353**, 119.

64 Klinger, L.M., Levin, L.A. and Petelin, A.L. (1997) *Defect and Diffusion Forum*, **143–147**, 1523.

65 Shashkov, D.A. and Seidman, D.N. (1995) *Physical Review Letters*, **75**, 268.

66 Kluthe, C., Al-Kassab, T. and Kirchheim, R. (2002) *Materials Science and Engineering A*, **327**, 70.

67 Kuduz, M., Schmitz, G. and Kirchheim, R. (2004) *Ultramicroscopy*, **101**, 197.

68 Tsong, T.T., Block, J.H., Nagasaka, M. and Viswanathan, B. (1976) *Journal of Chemical Physics*, **65**, 2469.

69 Nishigaki, S., Drachsel, W. and Block, J.H. (1979) *Surface Science*, **87**, 389.

70 Kellog, G.L. and Tsong, T.T. (1980) *Journal of Applied Physics*, **51**, 1184.

71 Tsong, T.T. (1984) *Physical Review B - Condensed Matter*, **30**, 4946.
72 Grovenor, C.R.M. and Cerezo, A. (1989) *Journal of Applied Physics*, **65**, 5089.
73 Liddle, J.A., Norman, A., Cerezo, A. and Grovenor, C.R.M. (1989) *Applied Physics Letters*, **54**, 1555.
74 Gault, B., Vurpillot, F., Bostel, A., Menand, A. and Deconihout, B. (2005) *Applied Physics Letters*, **86**, 094101.
75 Novotny, L., Bian, R.X. and Xie, X.S. (1997) *Physical Review Letters*, **79**, 645.
76 Martin, O. and Girard, C. (1997) *Applied Physics Letters*, **70**, 705.
77 Martin, Y., Haffmann, H.F. and Wickramasinghe, H.K. (2001) *Journal of Applied Physics*, **89**, 5774.
78 Vella, A., Deconihout, B., Marrucci, L. and Santamato, E. (2007) *Physical Review Letters*, **99**, 046103.
79 Vella, A., Gilbert, M., Hideur, A., Vurpillot, F. and Deconihout, B. (2006) *Applied Physics Letters*, **89**, 251903.
80 Cerezo, A., Smith, G.D.W. and Clifton, P.H. (2006) *Applied Physics Letters*, **88**, 154103.
81 Robins, E.S., Lee, M.J.G. and Langlois, P. (1986) *Canadian Journal of Physics*, **64**, 111.
82 Vurpillot, F., Gault, B., Vella, A., Bouet, M. and Deconihout, B. (2006) *Applied Physics Letters*, **88**, 094105.
83 Bunton, J.H., Olson, J.D., Lenz, D.R. and Kelly, T.F. (2007) *Microscopy and Microanalysis*, **13**, 418.
84 Gault, B., Menand, A., de Geuser, F. and Deconihout, B. (2006) *Applied Physics Letters*, **88**, 114101.
85 Oberdorfer, C., Stender, P., Reinke, C. and Schmitz, G. (2007) *Microscopy and Microanalysis*, **13**, 342.
86 Kelley, T.F., Larson, D.J., Thompson, K., Alvis, R.L., Bunton, J.H., Olson, J.D. and Gorman, B.W. (2007) *Annu. Rev. Mater. Res*, **37**, 681–727.
87 Thompson, K., Flaitz, P.L., Ronsheim, P., Larson, D. and Kelley, T. F. (2007) *Science*, **317** 1370.
88 Schlesiger, R. (2008) Design und Aufbau einer Mikroelektroden Atomsonde, Dipl. Thesis, University of Münster.

8
Cryoelectron Tomography: Visualizing the Molecular Architecture of Cells*

Dennis R. Thomas and Wolfgang Baumeister

8.1
Introduction

In order to reveal the networks of macromolecular interactions which underlie higher cellular functions on a systems level, new techniques are needed. As a starting point, it is crucial to have a validated and quantified list of all components, which can be provided by using mass spectrometry-based proteomics techniques [1]. The next challenge is to analyze the interaction patterns *in situ* with increasing degrees of complexity, ranging from that of supramolecular modules, to organelles or even whole cells. Some of these systems are tightly integrated complexes, robust enough to withstand the isolation and purification procedures that are used traditionally in biochemistry. These functional modules are amenable to detailed studies with the established tools of structural biology [2]. Other complexes interact transiently and weakly to form supramolecular networks that are designed to associate or dissociate in response to specific signals. Many such putative cellular complexes have been identified through affinity-based isolation methods. The components of these complexes can be identified using mass spectrometry to provide valuable information regarding the composition of such interacting complexes. In this manner interactions can be detected, whether they are direct or indirect [3]. However, this type of approach is prone to error and does not provide information about the order or molecular details of how the components interact. As the complexity of interacting systems increases, the component lists and affinity-based interaction data no longer suffice to describe the architecture of networks [4]. Ideally, the target would be a 'snapshot' of the system in action, acquired in a nondisruptive, noninvasive manner, avoiding perturbations to the system. Thus, we can hope to obtain a detailed three-dimensional (3-D) image of the system in action at molecular resolution.

*This chapter is a modified and updated version of a previously published article: Baumeister, W. (2005) From proteomic inventory to architecture. *FEBS Letters* **579**(4), 933–7.

Nanotechnology. Volume 6: Nanoprobes. Edited by Harald Fuchs

ISBN: 978-3-527-31733-2

Electron tomography is uniquely suited to studying large pleomorphic structures and visualizing macromolecules in their functional cellular context. The development and implementation of automated low-dose data-acquisition procedures, has made it possible to study biological samples embedded in vitreous ice. Electron cryotomography provides a 3-D picture of complex systems preserved in as near to a native state as can be achieved at molecular resolution [5]. Today, we have the tools to bridge the (resolution) gap that currently exists between cellular and high-resolution molecular structural techniques. The reconstructed tomograms of organelles or cells at molecular resolution represent snapshots of their proteomes. These 3-D snapshots can be interpreted using advanced pattern recognition techniques, revealing the molecular architecture present at the instant the sample was frozen. The fitting of tomographic maps of cells with high-resolution structures of their components should ultimately enable us to generate pseudo-atomic models of large – and otherwise elusive – assemblies and networks in the act of performing their cellular role [6].

8.2
Basic Principles and Challenges of Electron Tomography

An electron micrograph is a two-dimensional (2-D) projection of the sample present on the specimen support. Much like a medical X-ray, all structural features present in three dimensions are imaged, but are seen superimposed in the 2-D image. As a consequence, the images are difficult to interpret directly. In electron tomography, the specimen holder is tilted incrementally around an axis perpendicular to the electron beam, and images are taken at each position. These images are projections of the same object as viewed from different angles. Based on principles first described by Radon [7], these projections can be combined to produce a 3-D density map (Figure 8.1). Before such a density map is calculated – most commonly by using a 'weighted back-projection' algorithm – the projection images must be all be aligned to a common origin.

In order to successfully apply electron tomography to ice-embedded specimens, it is necessary to overcome two conflicting sets of limitations which have, for more than two decades, stood in the way of its widespread application to radiation-sensitive biological samples:

- In order to obtain the most detail with the least distortion in the reconstruction, it is necessary to sample as large an angular range as possible while tilting with increments as small as possible. This strategy calls for maximizing the number of projection images used for the reconstruction.
- Biological specimens embedded in vitreous ice, are extremely sensitive to radiation damage; therefore, it is very important to minimize exposure to the beam. To acquire a tilt series of perhaps more than a hundred images, the dose per image must be limited in order to prevent radiation damage from destroying the finer details of the structure or, worse yet, rendering reconstructions useless. The problem is, as one lowers the dose per image, the signal-to-noise ratio (SNR)

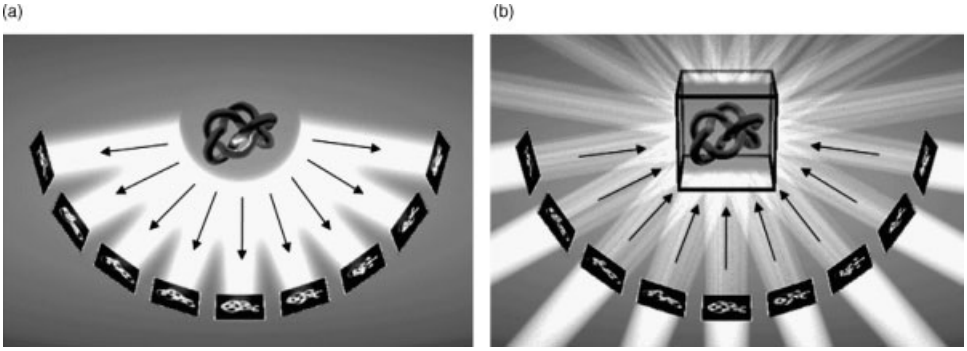


Figure 8.1 (a) Single axis tilt tomographic data acquisition. The unknown object is represented by a 'flexible knot' to emphasize the fact that electron tomography can reconstruct structures with unique topologies. A set of projection images is recorded as the object is tilted incrementally; (b) Following alignment of the projection images, the object is reconstructed generally by weighted back-projection. The 2-D projections are recombined to generate the 3-D density distribution of the object – the tomogram. The implementation of algorithms such as algebraic reconstruction techniques (ARTs) and simultaneous iterative reconstruction technique (SIRT) provide means for refining reconstructions [8, 9].

decreases. Thus, it is necessary to choose between having a better SNR but fewer projections, or more projections with a lower SNR.

Given the above considerations, we must consider how to 'spend' the allowable electron dose. Early theoretical considerations [10], corroborated by computer simulations [11], have suggested that, in principle, the electron dose needed to visualize structural features at a particular resolution limit is the same for 2-D and 3-D images containing equivalent information. In principle, if one were to average tomograms (3-D images), the dose could be distributed over as many projections as required to achieve the desired resolution, at the expense of lowering the SNR of the individual 2-D images. A consequence of combining the information from the projection images is an improvement in the SNR, similar to the improvement obtained by averaging statistically noisy images of repetitive structures. There is, nevertheless, a practical limitation, namely that the SNR of the 2-D images must be sufficient so as to permit the accurate alignment needed to establish a common framework of coordinates for all projection images.

If a constant exposure or dose per image were to be used throughout the tilt series then, as the tilt angles increased, images would be obtained with increasingly worse SNRs. This is because, as the tilt angle increases, the thickness of the specimen through which the electron beam passes also increases. One way to solve this problem would be to distribute the dose such that the dose or exposure time increased with the increasing tilt. The intended effect would be to keep the signal content of the images more or less constant. There are variations which can be used with this approach. One is to collect increasingly finer tilt increments at higher tilt angles without changing dose or, alternatively, to combine the previous two ideas and both increase the dose at higher tilt and collect finer increments. This latter approach

has the drawback of spending a great deal of the total allowed dose at high tilt angles which, due to increased specimen thickness and potential beam damage, may or may not contribute much to the information in the resulting tomogram.

Projection image alignment can be achieved using cross-correlation methods that may include area matching or feature tracking, or by the alignment of high-contrast fiducial markers, such as gold nanoparticles. The electron-dense gold particles have a relatively high SNR, even in very low-dose projections. However, gold – while not subject to radiation damage itself – is affected by the beam because the ice matrix in which it is embedded is subject to radiation-induced changes. If one compares the image of the untilted specimen at acquired at the start of the tilts series with an untilted image of the same area recorded after the tilt series has been completed, this can be observed as a movement of the gold relative to the specimen over the course of collecting a tilt series. Like radiation damage, this effect is dependent on the total dose and other factors such as specimen thickness. Thus, whether using cross-correlation only methods, or using fiducial marker-based alignment, the total electron dose which can be used is limited.

8.3 Automated Cryoelectron Tomography

Computer-controlled transmission electron microscopes first became commercially available during the late 1980s. This development, combined with the improving quality and availability of large-area charge-coupled device (CCD) cameras, made possible the development of sophisticated software that could be used to control the microscope and image-acquisition in a fully automated manner [12–15]. This software maintains the specimen centered in the field of view, by controlling the stage movement or beam shift, and determines image defocus automatically (Figure 8.2). An important development is the ability to track (center) and focus on areas some distance along the tilt axis away from the sample being imaged, thus minimizing the exposure of the area of interest. The fraction of the dose that is spent on overhead [search, centering, (auto) focusing] with automation has thus been reduced to as little as 3% of the total dose – something which is utterly impossible to achieve with manual operation [16].

These developments have greatly improved the status of electron tomography of cryopreserved specimens from that of a technique with potential to that of one beginning to produce exciting results. Starting with 'phantom cell' experiments – that is, liposomes encapsulating macromolecules [17, 18], and more recently with viruses, prokaryotic [19, 20] and eukaryotic cells [21] – we can now apply the potential of 3-D imaging to samples preserved in a close-to-life state. Vitrification by rapid freezing not only preserves the native molecular and cellular structures, but also allows 'snapshots' to be taken of dynamic events, thus freezing moments in time; an example would be trapping the short-lived open state of the acetylcholine receptor [22–24]. Vitrified samples do not suffer from the artifacts traditionally associated with chemical fixation and staining, nor with the dehydration of cellular structures. Tomograms of frozen-hydrated structures have a natural density distribution (albeit noisy), whereas staining and preservation reactions

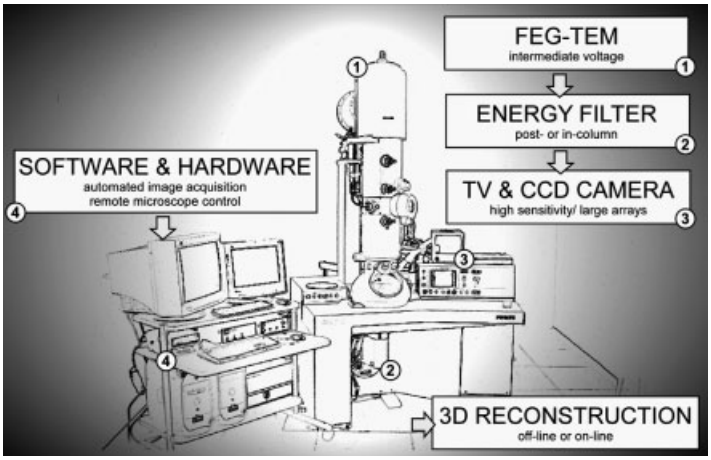


Figure 8.2 Cartoon showing where the advances in electron microscopy come into play. (1) The development of highly coherent intermediate-voltage electron sources [field emission guns (FEGs)] results in images with good signal-to-noise ratios extending to higher resolutions; (2) The application of energy filters to exclude inelastically scattered electrons from the image reduces noise in the image; (3) The acquisition of digital images on CCD cameras allows data acquisition to be automated; (4) The ability to control specimen movement, focusing, image tracking and image acquisition with computer software increases the efficiency of data collection and minimizes the electron dose spent on non-data-acquisition steps.

tend to produce artificial contrast from intricate mixtures of positive and negative staining. Unfortunately, the artifacts of staining and preservatives make the molecular interpretation of tomograms from such techniques very problematic [25].

8.4
Resolution, Signal-to-Noise Ratio and Visualization of Tomograms

Today, the development of automated procedures has made the recording of low-dose tilt series a routine procedure, with user-friendly software available for downstream processing [26, 27]. It is in fact now significantly less cumbersome and time-consuming to obtain a cryotomogram than to go through the conventional procedures of plastic embedding and sectioning biological samples. There is, however, one caveat – that sample thickness is a limitation. The thickness of whole prokaryotic cells, isolated organelles or thin (<1 μm) eukaryotic cells limits the resolution of cryotomograms to at best a range of 4–5 nm, although the prospects for improvement are good (see Refs. [26, 28]). The obstacle presented by sample thickness makes the development of reliable protocols for the sectioning of frozen-hydrated material a priority. While progress is clearly being made towards obtaining thin, artifact-free sections from samples preserved in vitreous ice, tomographic studies of thin cryosections are still not routine [29]. However, on-going improvements in instrumentation can be expected to make significant improvements in data quality. Better still, more efficient CCD detectors, in particular, would improve the resolution by

retaining the higher resolution signal that currently is lost to the modulation transfer function (MTF) of present-day CCDs. These improvements would allow tomography to enter the realm of molecular resolution (2–3 nm). In addition, with dual-axis tilting schemes, the effects of missing data (due to the restricted tilt range; usually $\pm 70^{\circ}$) could be reduced and resolution become more isotropic.

In electron tomography, as in the processing of single particles or 2-D crystals, an attainable resolution and the SNR are intricately linked. A signal may be present in a tomogram at high resolution, but detection of the information is limited by the limited degree of averaging which is inherent in the tomographic reconstruction process. As a consequence of averaging, the noise cancels (and thus is reduced in the averaging process) while the signal is summed. However, the signal can only be detected at frequencies where an acceptable SNR is achieved [30]. Electron microscopic single-particle analysis benefits greatly from the stratagem of combining a very large number of images (i.e. >10 000) of a structure viewed in random orientations, into a 3-D reconstruction, where averaging significantly improves the SNR [31]. The noise reduction resulting from averaging during reconstruction is quite limited in electron tomography (<150 images), given the uniqueness of cellular tomograms. Therefore, other means of noise reduction must be applied to the analysis of tomograms. For example, if the tilt series has been acquired in such a way that information does not extend past a certain resolution limit (say 5 nm), the tomogram may be Fourier-filtered with a cut-off at that resolution, thereby eliminating what can only be noise at frequencies higher than the cut-off frequency. Sophisticated 'de-noising' algorithms are also available; these are based on the same basic principle but employ 'diffusion' criteria based on local continuity of density [32]. Although the gain in SNR obtained from such methods is not very large, this does help in visualizing the underlying structure present in the tomograms. Nonetheless, new algorithms are required to achieve a greater noise reduction.

The interpretation of a tomogram at an ultrastructural level requires the identification of structural components – that is, membrane-bound organelles, cytoskeletal filaments or large macromolecular complexes. In the past, manual assignment has been commonly used as human pattern recognition is often superior to available segmentation algorithms; on the other hand, machine-based segmentation should, in principle, be more objective. Continuous structures are relatively easy to recognize and delineate, in spite of the low SNR present in cryotomograms. For example, visualizing the organization of the cytoskeleton in both *Spiroplasma melliferum* and *Dictyostelium discoideum*, was possible at the level of individual filaments, without the need for extensive post-processing (Figure 8.3) [21, 33].

Although averaging can obviously not be applied to tomograms of unique structures such as individual cells or organelles, such tomograms may nevertheless contain multiple copies of components such as ribosomes, chaperones or proteases. Small regions of the tomogram containing isolated complexes can be extracted from the tomogram *in silico*, and these so-called 'subtomograms' can be subjected to classification against a library of known structures (see Figure 8.4). The subtomograms can then be aligned to a common orientation and averaged within the appropriate class. The result should be an average with a better SNR and a higher

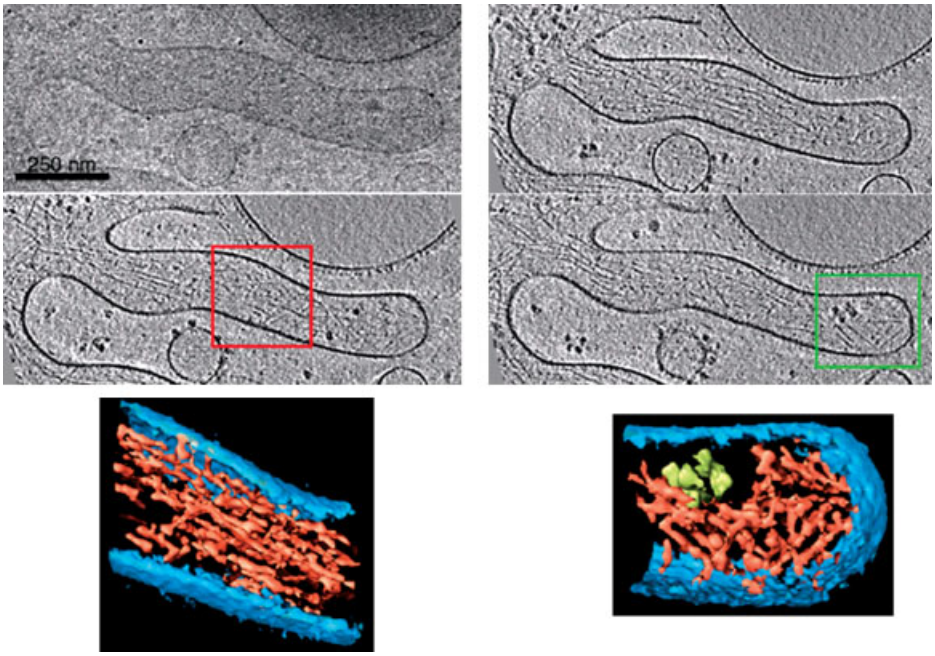


Figure 8.3 Actin filament organization in filopodia [34]. The upper left panel is a projection image from the tilt series. The next three images are sequential sections taken from the tomogram. Two segments of the filopod have been surface-rendered to reveal the organization of the actin filaments and interactions with the membrane. The red box indicates the region shown in the rendered image (lower left); the green box indicates the region rendered (lower right).

resolution. The original low-resolution tomographic image can be replaced by the average or by the higher-resolution template itself, if available. The result is a 'synthetic' tomogram with a much improved, local SNR. Such a procedure has been used in a tomographic study of enveloped *herpes simplex* virions [19] (Figure 8.5), and to also visualize nuclear pore complexes in intact nuclei [35, 36] (Figure 8.6).

8.5
Merging High Resolution with Low: The Molecular Interpretation of Cryotomograms

In tomographic reconstructions of vitrified samples, the macromolecular content of an organelle or cell is present in its native state, thereby making interpretation at the molecular level less problematic. Although the resolution of an individual tomogram may be limited, the advantage is that everything can be seen in its native context. A vast amount of information is available, as tomograms are 3-D images of the entire proteome, and should ultimately enable the spatial relationships of macromolecules to be mapped in an unperturbed cellular environment. However, the retrieval of this information is faced with major problems. First, although everything can be seen, the identification of what is seen may be difficult in the crowded macromolecular

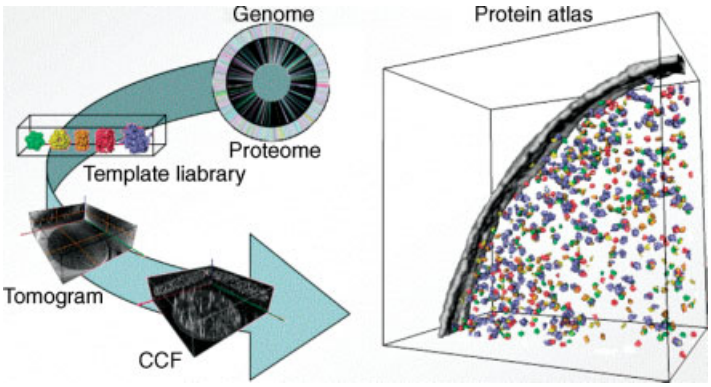


Figure 8.4 Using template matching to generate a protein atlas of a cell. Proteomics, nuclear magnetic resonance, X-ray crystallography and electron microscopy have produced a wealth of structural templates which can be used as a library for analyzing tomograms. In this example, templates from the library have been chosen as probes for the tomogram. Each template, in a number of different orientations, is cross-correlated against the tomogram. The positions of complexes appear as peaks in the 3-D cross correlation function (CCF). A box or subvolume is extracted centered on each of these peaks, resulting in a dataset of 3-D images or subvolumes. These subvolumes are then aligned in three dimensions against the library of complexes. Each volume is assigned to the class to which it correlates best, and each of the classes is averaged, increasing the signal-to-noise ratio and improving resolution. Finally, the class average or original template for the class can be substituted for the original complex in the tomogram, in the determined orientation. This produces a 3-D map of the cell with the contents identified. It is now possible to consider how the various complexes might be related in the cellular context.

environment where complexes literally touch each other [36]. Therefore, the interpretation of low-SNR cryo-tomograms is difficult and can be tedious. Furthermore, because it is not possible to tilt a full 180°, the tomographic reconstructions are distorted by missing data, and this results in a nonisotropic resolution. There are essentially only two options for identifying macromolecules in tomograms, namely *specific labeling* or *pattern recognition methods*, where complexes are matched against a library of known structures. Of course, the two approaches are not mutually exclusive.

8.5.1 Specific Labeling

Proteins exposed on the surface of cells or organelles can be labeled with antibodies or, specific ligands bound to gold nanoparticles. These labels provide indicators for the presence of a specific molecule within a broad molecular landscape. Intracellular labeling is more problematic and requires innovative approaches. These approaches could be based on noninvasive genetic manipulations generating covalent fusions with a protein such as metallothionein, which has the potential to bind heavy metals such as gold [38, 39]. Ideally, the aim would be to introduce a label in a time-resolved experiment (thereby identifying a particular event), and subsequently to remove the background, as is achieved with the ReAsH compound in fluorescence microscopy [40]. However, the achievement of labeling that can be statistically quantified is a

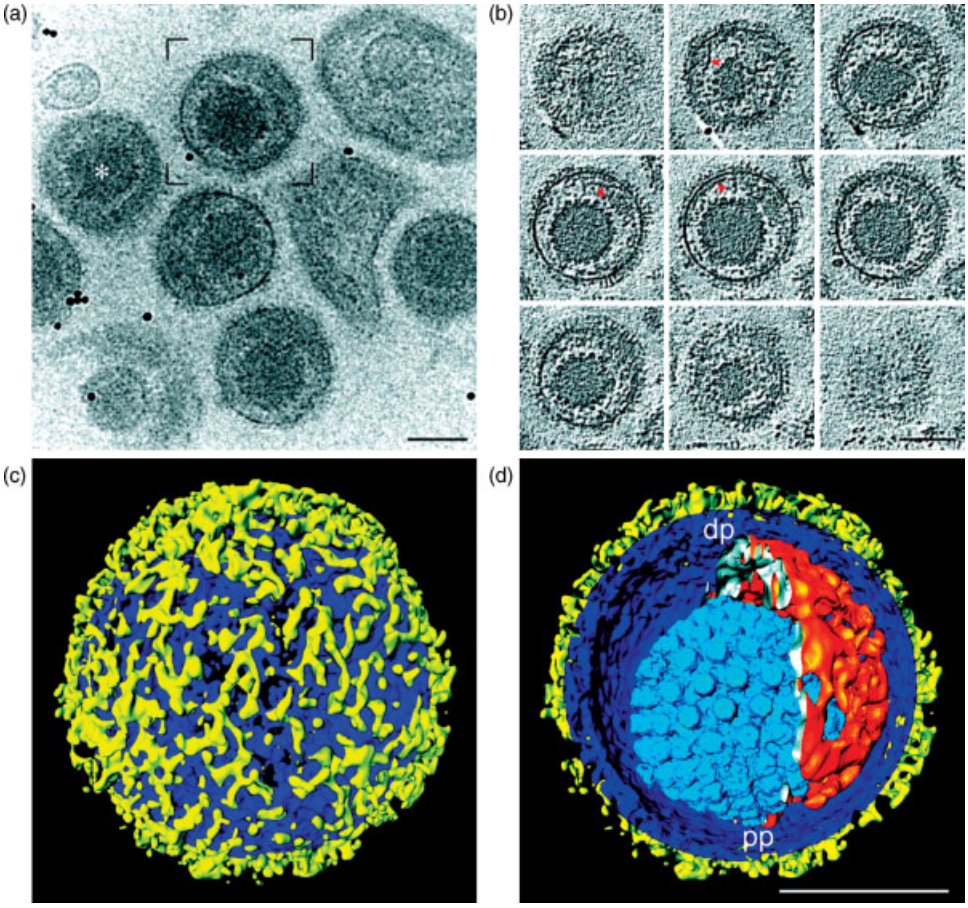


Figure 8.5 (a, b) Tomographic reconstruction of HSV-1 virions in vitreous ice (based on Ref. [19]). (a) The untilted projection from a tilt series. The black dots are 10 nm gold particles used as fiducial markers; (b) Gallery of sections from the tomogram taken from the virion framed in panel (a). Each section is an average of seven planes from the tomogram, and represents a slab which is 5.2 nm thick. The sections are separated by 15.5 nm. Red arrowheads mark filaments in the tegument. All scale bars are 100 nm. (c, d) The surface-rendered tomogram from the same virion after denoising; (c) Outer surface showing the distribution of glycoprotein spikes (yellow) protruding from the membrane (blue); (d) Cutaway view of the virion interior, showing the capsid (light blue) and the tegument 'cap' (orange) inside the envelope (blue and yellow). pp = proximal pole; dp = distal pole.

daunting task. It is also difficult to imagine that labeling can be developed such that it becomes a high-throughput technology capable of mapping entire proteomes. In order to identify every molecule of interest, the entire procedure of labeling, as well as data acquisition and reconstruction of the tilt series, must be repeated. Moreover, the unique nature of cellular tomograms makes the direct correlation of the different maps impossible, which in turn poses a major problem when deriving such maps from the molecular interaction patterns.

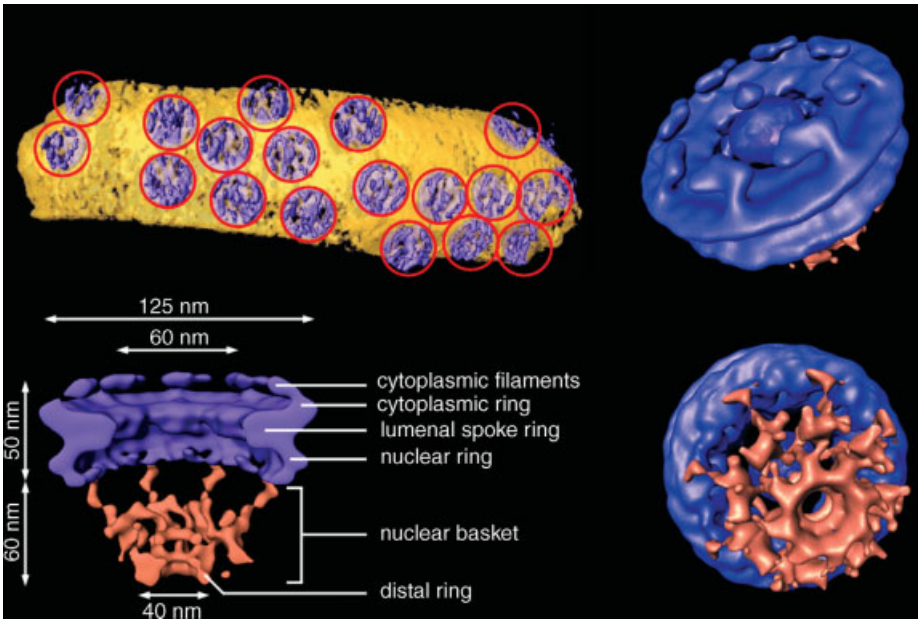


Figure 8.6 The structure of the nuclear pore complex (NPC) obtained from averaging subvolumes [35]. A surface-rendered representation of a segment of nuclear envelope (NPCs in blue, membranes in yellow). The dimensions of the rendered nuclear volume are 1680 nm × 984 nm × 558 nm. The upper right diagram shows the cytoplasmic face of the NPC, and the lower right the inward (nuclear) face. The distal ring of the basket is connected to the nuclear ring by the nuclear filaments. The lower left diagram shows a cutaway view of the NPC. The dimensions of the main features are indicated. The nuclear basket is shown in brown.

An alternative strategy is to combine fluorescent light microscopy and cryotomography. Specific fluorescent labels can be engineered into proteins of interest, or alternatively fluorescent labels can be applied and taken up by the cells. Progress has been made in the development of the cryogenic light microscope [41, 42], which can visualize fluorescence in vitrified specimens. This allows a fluorescence image of the frozen hydrated specimen to be recorded on the grid, and for any targets of interest for tomography to be identified. The specimen is then transferred directly to the electron microscope, where tomographic tilt series can be recorded of the target areas. Software-based methods can be used to align the fluorescence image with a low magnification image from the electron microscope, thus locating the desired areas in the electron microscope overview image.

8.5.2 Pattern Recognition

The identification of a single 26S proteasome in the cytoplasm of a *Dictyostelium* cell suggested that a template-matching approach could be used for mapping cellular proteomes [21]. Given that one can see 'everything' in a tomogram, it makes sense to

probe the tomogram with a library of templates derived from known structures, using intelligent pattern recognition algorithms. The intent is to locate known structures and determine the molecular context in which complexes are organized in organelles or cells, with less emphasis on the discovery of novel molecular features and more on determining whether there is some correlation in the spatial arrangement of complexes, relative to one another. To achieve this, a 'template-matching' strategy is being pursued [43, 44]. Given the array of high- to medium-resolution relevant structures that are available to be included in a template library, the systematic analysis of tomograms by scanning for the presence of these known templates is feasible. The procedure is computationally intensive, as not only must the positions matching a given template be determined, but also their spatial orientations. The tomogram must be probed with every template, and the best match determined for each complex in the tomogram. Ideally, such a multi-template search would result in a 3-D map with each low-resolution complex that was identified replaced by the higher-resolution template that it matched, in the orientation determined (Figure 8.4). Simulations and experiments with 'phantom cells' have shown that such an approach is feasible [44], and that the search results can be validated. At present, only large complexes such as the ribosome have been identified with an acceptable accuracy (> 95%) at the routinely attained resolution of 4–5 nm. Yet, an improvement in resolution to 2–3 nm would allow the accurate identification of smaller complexes [45].

8.6
Creating Template Libraries

Once the challenges of obtaining a sufficiently good resolution have been met, the next target will be to expand the libraries of available templates. Many approaches can contribute to achieve this goal. Structural genomics efforts will increase the pace at which high-resolution structures of domains, subunits or larger entities become available, and eventually will provide a comprehensive structural dictionary. Hybrid methods, combining information from different sources and of variable quality, will also play an important role [2, 4]. Electron microscopic single-particle analysis will undoubtedly continue to provide many medium-resolution structures of complexes. Clearly, the prospects for accelerating throughput by means of automated data collection and analysis show great promise [15].

Nonetheless, sample preparation continues to be the major bottleneck that slows progress in the field. Today, improvements are required in the methods used to isolate and purify proteins. This is especially true for labile macromolecular assemblies (which are probably more abundant in cells than stable complexes), with novel strategies being required. Whilst traditional biochemical methods tend to optimize for yield and/or purity, they are frequently time-consuming, and labile or transient complexes are generally lost during a traditional course of isolation. Given the modest requirements of single-particle analysis – where only very small quantities of material are needed and impurities can be removed computationally – there is no reason to

purify labile complexes to exhaustion. It has been shown recently, by using lipid monolayer techniques and taking advantage of His tags, that it is possible specifically to pick up only the desired His-tagged protein from a crude extract and directly freeze and image the specimen in vitreous ice [46]. There is, therefore, plenty of scope for innovative approaches aimed either at minimizing the purification steps, or even avoiding purification altogether by taking advantage of optimized *in vitro* translation systems.

8.7 Outlook

With cryoelectron tomography providing 3-D images at molecular resolution, and with image analysis tools at our disposal for interpreting the tomograms, we are now poised to integrate structural data into pseudo-atomic maps of organelles or cells. Whilst these maps will provide unprecedented insights into the molecular architecture that underlies cellular behavior, they will also pose new challenges. It will not be a trivial task to extract generic features from the maps, nor to derive general rules regarding the principles that govern supramolecular organization, given the stochastic nature of cellular systems as well as their dynamics. Most importantly, systems analysis will need to start taking this into account, and there will also be a need to develop statistical methods similar to those used when analyzing macroscopic social systems. Alternatively, sophisticated molecular dynamics software could be expanded to model large-scale systems.

References

1 Aebersold, R. and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature*, **422** (6928), 198–207.

2 Sali, A., Glaeser, R., Earnest, T. and Baumeister, W. (2003) From words to literature in structural proteomics. *Nature*, **422** (6928), 216–225.

3 Aloy, P., Boettcher, B., Ceulemans, H. *et al.* (2004) Structure-based assembly of protein complexes in yeast. *Science*, **303** (5666), 2026–2029.

4 Robinson, C.V., Sali, A. and Baumeister, W. (2007) The molecular sociology of the cell. *Nature*, **450** (7172), 973–982.

5 Baumeister, W., Grimm, R. and Walz, J. (1999) Electron tomography of molecules and cells. *Trends in Cell Biology*, **9** (2), 81–85.

6 Baumeister, W. (2004) Mapping molecular landscapes inside cells. *Biological Chemistry*, **385** (10), 865–872.

7 Radon, J. (1917) Über die Bestimmung von Funktionen durch ihre Integralwerte längs gewisser Mannigfaltigkeiten. Berichte über die Verhandlungen der Königlich Sächsischen Gesellschaft der Wissenschaften zu Leipzig. *Mathematisch Physikalishe Klasse*, **69**, 262–277.

8 Gordon, R., Bender, R. and Herman, G.T. (1970) Algebraic reconstruction techniques (ART) for three-dimensional electron microscopy and X-ray photography. *Journal of Theoretical Biology*, **29** (3), 471–481.

9 Gilbert, P. (1972) Iterative methods for the three-dimensional reconstruction of an

object from projections. *Journal of Theoretical Biology*, **36** (1), 105–117.

10 Hegerl, R. and Hoppe, W. (1976) Influence of electron noise on three-dimensional image reconstruction. *Zeitschrift für Naturforschung*, **A314**, 1717–1721.

11 McEwen, B.F., Downing, K.H. and Glaeser, R.M. (1995) The relevance of dose-fractionation in tomography of radiation-sensitive specimens. *Ultramicroscopy*, **60** (3), 357–373.

12 Typke, D., Dierksen, K. and Baumeister, W. (1991) Automatic electron tomography (ed. W. Bailey) *Proceedings, 49th Annual Meeting of the Electron Microscopy Society of America*, San Francisco Press, San Francisco, CA, pp. 544–545.

13 Dierksen, K., Typke, D., Hegerl, R., Koster, A.J. and Baumeister, W. (1992) Towards automatic electron tomography. *Ultramicroscopy*, **40**, 71–87.

14 Dierksen, K., Typke, D., Hegerl, R. and Baumeister, W. (1993) Towards automatic electron tomography. II. Implementation of autofocus and low-dose procedures. *Ultramicrosopy*, **49**, 109–120.

15 Carragher, B., Fellmann, D., Geurra, F. *et al.* (2004) Rapid routine structure determination of macromolecular assemblies using electron microscopy: current progress and further challenges. *Journal of Synchrotron Radiation*, **11** (Pt 1) 83–85.

16 Koster, A.J., Grimm, R., Typke, D. *et al.* (1997) Perspectives of molecular and cellular electron tomography. *Journal of Structural Biology*, **120** (3), 276–308.

17 Dierksen, K., Typke, D., Hegerl, R., Walz, J., Sackmann, E. and Baumeister, W. (1995) Three-dimensional structure of lipid vesicles embedded in vitreous ice and investigated by automated electron tomography. *Biophysical Journal*, **68** (4), 1416–1422.

18 Grimm, R., Barmann, M., Hackl, W., Typke, D., Sackmann, E. and Baumeister, W. (1997) Energy filtered electron tomography of ice-embedded actin and vesicles. *Biophysical Journal*, **72** (1), 482–489.

19 Grunewald, K., Desai, P., Winkler, D.C. *et al.* (2003) Three-dimensional structure of herpes simplex virus from cryo-electron tomography. *Science*, **302** (5649), 1396–1398.

20 Grimm, R., Singh, H., Rachel, R., Typke, D., Zillig, W. and Baumeister, W. (1998) Electron tomography of ice-embedded prokaryotic cells. *Biophysical Journal*, **74** (2 Pt 1), 1031–1042.

21 Medalia, O., Weber, I., Frangakis, A.S., Nicastro, D., Gerisch, G. and Baumeister, W. (2002) Macromolecular architecture in eukaryotic cells visualized by cryoelectron tomography. *Science*, **298** (5596), 1209–1213.

22 Dubochet, J., Adrian, M., Chang, J.J. *et al.* (1988) Cryo-electron microscopy of vitrified specimens. *Quarterly Reviews of Biophysics*, **21** (2), 129–228.

23 Berriman, J. and Unwin, N. (1994) Analysis of transient structures by cryo-microscopy combined with rapid mixing of spray droplets. *Ultramicroscopy*, **56** (4), 241–252.

24 Unwin, N. (1995) Acetylcholine receptor channel imaged in the open state. *Nature*, **373** (6509), 37–43.

25 Baumeister, W. (2002) Electron tomography: towards visualizing the molecular organization of the cytoplasm. *Current Opinion in Structural Biology*, **12** (5), 679–684.

26 Plitzko, J., Frangakis, A.S., Nickell, S., Förster, F., Gross, A. and Baumeister, W. (2002) *In vivo* veritas: electron cryotomography of cells. *Trends in Biotechnology*, **20**, s40–s44.

27 Nickell, S., Forster, F., Linaroudis, A. *et al.* (2005) TOM software toolbox: acquisition and analysis for electron tomography. *Journal of Structural Biology*, **149** (3), 227–234.

28 Gruska, M., Medalia, O., Baumeister, W. and Leis, A. (2008) Electron tomography of vitreous sections from cultured mammalian cells. *Journal of Structural Biology*, **161** (3), 384–392.

29 Al-Amoudi, A., Norlen, L.P. and Dubochet, J. (2004) Cryo-electron microscopy of

vitreous sections of native biological cells and tissues. *Journal of Structural Biology*, **148** (1), 131–135.

30 Unser, M., Trus, B.L., Frank, J. and Steven, A.C. (1989) The spectral signal-to-noise ratio resolution criterion: computational efficiency and statistical precision. *Ultramicroscopy*, **30** (3), 429–433.

31 Frank, J. (2002) Single-particle imaging of macromolecules by cryo-electron microscopy. *Annual Review of Biophysics and Biomolecular Structure*, **31**, 303–319.

32 Frangakis, A.S. and Hegerl, R. (2001) Noise reduction in electron tomographic reconstructions using nonlinear anisotropic diffusion. *Journal of Structural Biology*, **135** (3), 239–250.

33 Kurner, J., Medalia, O., Linaroudis, A.A. and Baumeister, W. (2004) New insights into the structural organization of eukaryotic and prokaryotic cytoskeletons using cryo-electron tomography. *Experimental Cell Research*, **301** (1), 38–42.

34 Medalia, O., Beck, M., Ecke, M. *et al.* (2007) Organization of actin networks in intact filopodia. *Current Biology*, **17** (1), 79–84.

35 Beck, M., Forster, F., Ecke, M. *et al.* (2004) Nuclear pore complex structure and dynamics revealed by cryoelectron tomography. *Science*, **306** (5700), 1387–1390.

36 Beck, M., Lucic, V., Forster, F., Baumeister, W. and Medalia, O. (2007) Snapshots of nuclear pore complexes in action captured by cryo-electron tomography. *Nature*, **449** (7162), 611–615.

37 Grunewald, K., Medalia, O., Gross, A., Steven, A.C. and Baumeister, W. (2003) Prospects of electron cryotomography to visualize macromolecular complexes inside cellular compartments: implications of crowding. *Biophysical Chemistry*, **100** (1–3), 577–591.

38 Mercogliano, C.P. and DeRosier, D.J. (2006) Gold nanocluster formation using metallothionein: mass spectrometry and electron microscopy. *Journal of Molecular Biology*, **355** (2), 211–223.

39 Mercogliano, C.P. and DeRosier, D.J. (2007) Concatenated metallothionein as a clonable gold label for electron microscopy. *Journal of Structural Biology*, **160** (1), 70–82.

40 Gaietta, G., Deerinck, T.J., Adams, S.R. *et al.* (2002) Multicolor and electron microscopic imaging of connexin trafficking. *Science*, **296** (5567), 503–507.

41 Sartori, A., Gatz, R., Beck, F., Rigort, A., Baumeister, W. and Plitzko, J.M. (2007) Correlative microscopy: bridging the gap between fluorescence light microscopy and cryo-electron tomography. *Journal of Structural Biology*, **160** (2), 135–145.

42 Schwartz, C.L., Sarbash, V.I., Ataullakhanov, F.I., McIntosh, J.R. and Nicastro, D. (2007) Cryo-fluorescence microscopy facilitates correlations between light and cryo-electron microscopy and reduces the rate of photobleaching. *Journal of Microscopy*, **227** (Pt 2), 98–109.

43 Bohm, J., Frangakis, A.S., Hegerl, R., Nickell, S., Typke, D. and Baumeister, W. (2000) Toward detecting and identifying macromolecules in a cellular context: template matching applied to electron tomograms. *Proceedings of the National Academy of Sciences of the United States of America*, **97** (26), 14245–14250.

44 Frangakis, A.S., Bohm, J., Forster, F. *et al.* (2002) Identification of macromolecular complexes in cryoelectron tomograms of phantom cells. *Proceedings of the National Academy of Sciences of the United States of America*, **99** (22), 14153–14158.

45 Ortiz, J.O., Forster, F., Kurner, J., Linaroudis, A.A. and Baumeister, W. (2006) Mapping 70S ribosomes in intact cells by cryoelectron tomography and pattern recognition. *Journal of Structural Biology*, **156** (2), 334–341.

46 Kelly, D.F., Dukovski, D. and Walz, T. (2008) Monolayer purification: a rapid method for isolating protein complexes for single-particle electron microscopy. *Proceedings of the National Academy of Sciences of the United States of America*, **105** (12), 4703–4708.

9
Time-Resolved Two-Photon Photoemission on Surfaces and Nanoparticles

Martin Aeschlimann and Helmut Zacharias

9.1
Introduction

Occupied and unoccupied states of solid-state materials are traditionally investigated using photoemission and inverse photoemission spectroscopies. Due to the limited and short path length of electrons in the 20 to 100 eV range, an excellent surface specificity is achieved with these methods, enabling differentiation to be made between signals from the bulk of a material and from layers in the vicinity of the surface. The advent of ultra-short laser pulses among those laboratories investigating surface sciences has extended the field of research, as this technique provides the possibilities both to investigate occupied states with photon energies sufficiently high to surpass the work function of the material, and to study unoccupied states by using resonantly enhanced two-photon photoemission (2PPE). In this process, electrons are promoted by a first laser photon to an unoccupied state; this may be a volume, surface, an adsorbate state or an image potential state (IS) [1] of the material. Before the electron can re-equilibrate, a second photon from the same (or a different) laser beam is absorbed by the excited state. This promotes the electron of the excited state above the vacuum level, leading to an emission of photoelectrons which thus have been interacting with two laser fields. In this way, the spectral resolution for the intermediate unoccupied states is greatly improved compared to inverse photoemission [2].

The two-pulse scheme of photoemission bears another extremely important advantage, in that it allows the delay of one pulse against the other. This so-called time-resolved two-photon photoemission (TR-2PPE) allows investigation of the dynamics of the intermediate electronically excited state on a femtosecond time scale. Moreover, energy and momentum transfer processes in the excited state can also be studied, while scattering phenomena can also be addressed. These time-resolved photoemission experiments have been conducted for about 25 years. Initially, the thermalization dynamics of band edge carriers in semiconductors and electron–phonon (e–ph) dynamics in metals were studied using lasers of typically a

Nanotechnology. Volume 6: Nanoprobes. Edited by Harald Fuchs

ISBN: 978-3-527-31733-2

few tens of picoseconds pulse duration [3–6]. More recently, however, by using lasers with femtosecond time resolution, the lifetimes of image potential states on silver surfaces [7] and questions regarding electron–electron (e–e) scattering dynamics in bulk metals have been addressed [8–10].

During the past two decades, photoemission techniques have been coupled with electron imaging by means of photoemission electron microscopy (PEEM), and this has allowed material-specific microscopic images of nanostructures on surfaces to be obtained with sub-100 nm resolution [11]. Important information concerning various new and unexpected phenomena, including the nonlinear behavior of surface reaction dynamics [12] and of layered magnetic materials, has been obtained using this method. These imaging methods, when combined with time resolution in the femtosecond regime [11, 13], will become increasingly important in emerging areas such as molecular electronics, self-assembled and self-organized functional layers, organic solar energy converters and photocatalytic reaction centers.

In this chapter we will provide examples of the imaging methods, together with their recent application in the preparation of metallic nanostructures. Such nanostructures will become increasingly important as fields of molecular switching and electronics, plasmonics and functional molecular assemblies evolve. Moreover, as the dimensions of structures shrink to the submicrometer level, novel properties and functions will undoubtedly emerge. The spectral tuning of the emission of entities and controlled charge transport in organized nanostructured environments [14–16] represent just two aspects of functional modification in molecular electronics. It is further envisioned that the optical phase control of charge transport [17] may become very important in such systems. It follows that a precise knowledge of the basic electronic properties of these nanostructured systems is, therefore, of major importance.

9.2 Theoretical Background

The electron dynamics in metallic nanostructures are governed by a few elementary processes which will now briefly be described. Besides the fundamental charge screening, which takes place on an attosecond time scale, the decoherence of excited electrons and electron–electron scattering processes are primary processes leading to a rapid redistribution of electron energy and momentum. Thereby, a hot electron gas is created. On nanostructures a coherent excitation of the whole electronic system – the creation of particle plasmons – is of fundamental nature and a specific element to be considered in these systems. The decay processes of particle plasmons are presently under intense investigation. The hot electrons created by various processes then couple to the phonon system, eventually dissipating the energy to heat. A concise theoretical description of the basic processes involved, which is beyond the scope of this chapter, can be found in recent reports [18–20].

9.2.1
Electron–Electron Interaction

The simplest description of electronic interaction in metals can be derived from the Drude theory of metals [21]. Although based on currently outdated assumptions, this theory provides a good estimate of the magnitude of electron scattering for various classes of metals. The electron mean free path, divided by their velocity at the Fermi energy, yields the collision free time which denotes the lifetime of an electron at a certain energy. At room temperature, this lifetime ranges from a few femtoseconds in transition metals to a few tens of femtoseconds in noble metals.

A more quantitative description of the electron–electron scattering rate is obtained from the Fermi liquid theory [22, 23] for a free electron gas (FEG FLT). In brief, phase space arguments are invoked to describe the interaction of an electron with kinetic energy above the Fermi level with unexcited electrons. Due to the same mass of the interacting electrons, and the fact that all states below the Fermi level are occupied, the inelastic scattering of a hot electron with the cold Fermi gas yields two electrons, both with kinetic energies now above the Fermi level. For a single excited electron this inelastic process yields a new kinetic energy and momentum, thus limiting the mean free path of an electron at a given energy. It is clear that electrons far above the Fermi level have a larger phase space available for scattering than those close to the Fermi level. Therefore, the scattering rate strongly decreases as the energy of the hot electron relaxes towards the Fermi level.

These inelastic processes continue, and finally a hot electron gas is rapidly created. When an intense laser pulse creates the primary excitation, the excitation density is high and therefore also the density of the hot electron gas. Scattering events with adsorbed molecules then become probable, and processes such as desorption induced by electronic transitions (DIET) and desorption induced by multiple electronic transitions (DIMET) or electronic friction induced adsorbate excitation my become observable.

Within the relaxation time approximation, and invoking Boltzmann transport theory with the Fermi–Dirac distribution for the occupation, the Fermi liquid theory yields in three dimensions for the scattering rate [24]

$$\tau^{-1} = \tau_0^{-1}(E-E_F)^2 + b(k_B T)^2 \tag{9.1}$$

At excitation energies above 200 meV, the thermal contribution can usually be neglected at all practical temperatures. The prefactor τ_0 depends only on the electron density in the metal under consideration:

$$\tau_0 = \frac{64}{\sqrt{3}\pi^{5/2}} \sqrt{\frac{m_e}{ne^2}} \quad E_F^2 \tag{9.2}$$

The inelastic lifetime of an electron above the Fermi level, given by the inverse scattering rate, is therefore inversely proportional to the square of the energy difference to the Fermi level and, taking the dependence of E_F on the electron density into account, proportional to $n^{5/6}$. At excitation energies of about 3 eV this amounts to a few femtoseconds, whereas close to the Fermi level, $E < 0.2$ eV, it may approach the picosecond range, depending on the metal under consideration [25, 26].

Usually, the FEG FLT serves as a benchmark for comparison of the observed hot-electron inelastic lifetimes τ with the theory, and therefore electron relaxation dynamics has first been investigated for noble metals, where a reasonable agreement with the FEG FLT was expected [10, 27–30]. For all other metals, the key role in the low-energy electron relaxation dynamics is played by the electronic structure of the system close to the Fermi level. For example, in transition metals the high density of *d* bands in the proximity of the Fermi level leads to a very fast electron relaxation. In this context, the exact energy position and shape of the *d* bands must be considered in detail in order to achieve a complete understanding of the relaxation processes. During the past decade, several TR-2PPE experiments have been performed for the transition metals Ta [31], Ru [32], Mo and Rh [33], for ferromagnetic 3*d* metals Fe, Co and Ni [34], and for high-T_C superconductors [35] and 4*f*- rare earth metal Yb [36].

Theoretical calculations of excited electron lifetimes have been performed in the past within the *GW* approximation (GWA) for electron self-energy for bulk noble [37, 38] and 4*d* transition metals [39–42], as well as for the 5*d* transition metal Ta [31]. In contrast to noble metals, which show qualitatively similar band structure and density of states (DOS) [43], the electronic structure of 4*d* metals varies strongly on moving from the start of the 4*d* series to the end [43]. Calculations performed by Zhukov *et al.* [41] and Bacelar *et al.* [42] have shown that the evaluated lifetimes also vary widely along the 4*d* series, following trends in electronic structure. The extension, $GW + T$, of the *GW* approximation by inclusion of multiple electron-hole scattering within a *T*-matrix approximation [44–48] results in a decrease of the *GW* lifetime value that brings theory and experiment to better agreement [47, 48].

9.2.2 Plasmonic Processes

In noble metal nanoparticles collective electronic oscillations – so-called particle plasmons or localized surface plasmons (LSPs) – can be excited by electromagnetic waves. Therefore, they are detectable as pronounced resonances in the scattering and absorption cross-section, for the noble metals Ag and Au commonly located in the visible or ultraviolet (UV) region of the spectrum [49]. The resonance frequency of the plasma oscillation is determined by the dielectric properties of the metal and the surrounding medium, as well as by the particle size and shape [49–51]. The collective oscillation can be interpreted as a displacement of the electrons in the particle against the positively charged background of the atomic nuclei. Resonant excitation of this collective charge oscillation causes a large enhancement of the local field inside and near the particle [52] which dominates the linear and nonlinear responses of the particles to the light field. The field enhancement caused by the electron oscillation (see Figure 9.1) is thought to be responsible for the enhancement of nonlinear optical effects such as surface-enhanced Raman scattering (SERS) [53], surface second harmonic generation [54, 55] and multiphoton photoemission [56]. In recent years, the promising research field of plasmonics and ultrafast nano-optics has emerged, exploiting the high potential of plasmons to concentrate and channel light into subwavelength structures of nanoscopic circuits [57].

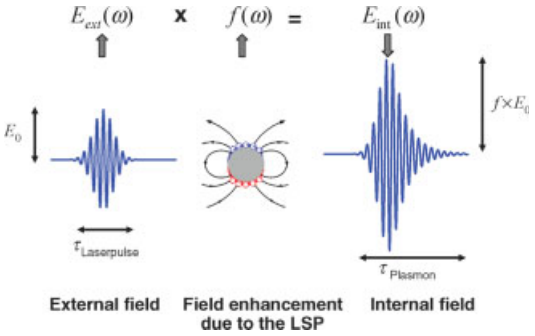


Figure 9.1 In metallic nanostructures collective electronic oscillations – local surface plasmons – can be excited by light; this results in a strong enhancement of the local field both inside and near the surface structure.

Although, the spectral positions of the resonances of particle plasmon excitations as a function of particle size, shape and dielectric properties are well understood [49–52], the ultrafast dynamics of these collective electronic excitations have remained a highly interesting topic to be studied in more detail. In order to understand the dynamics, it is essential to investigate those mechanisms relevant to the loss of phase coherence between the electrons contributing to the collective excitation – that is, the dephasing of the plasmonic state.

9.2.3 Two-Temperature Model

When a strong femtosecond laser pulse produces a large primary population of hot electrons, the electron–electron scattering leads to the formation of a hot electron gas. The formation and decay of this not-equilibrated hot electron distribution was first observed by Bokor and coworkers in thin gold films. The equilibration process of this electron gas can, in general, be described fairly well by the Fermi liquid theory, together with the Boltzmann transport equation in the relaxation time formulation [8, 9]. However, deviations from this description were noted at an early stage of these investigations. An important (and, at first sight, counterintuitive) finding was that the relaxation to a heated thermal distribution occurred faster if a more intense laser excitation was applied, and thus the density of hot electrons was higher.

The hot electron dynamics is usually modeled using a two-temperature model [58], where the electronic and phonon systems assume different temperatures; this is justified due to the wide differences in the heat capacities.

$$C_{e\ell}\frac{\partial T_{e\ell}}{\partial t} = \frac{\partial}{\partial z}\left(\kappa\frac{\partial T_{e\ell}}{\partial z}\right) - g_{\infty}(T_{e\ell} - T_{ph}) + S(z,t) \tag{9.3}$$

$$C_{ph}\frac{\partial T_{ph}}{\partial t} = g_{\infty}(T_{e\ell} - T_{ph}) \tag{9.4}$$

$S(z,t)$ represents the exciting optical intensity as it penetrates into the bulk of the material. The lateral dimensions are usually large, and thus uniform compared to the large gradients in z direction. $C_{el} = \gamma\, T_{el}$ represents the electronic heat capacity, and C_{ph} that of the phonon system. The coupling between the electronic system and the phonons is described by g_{∞}. With this system of equations, a good estimate of the magnitude and time dependence of electronic and phonon temperatures is usually achieved. This can easily be extended by frictional coupling to an adsorbate [59, 60], which thereby acquires internal energy and also serves as a cooling heat bath for the excited surface layers. Similarly, diffusive and ballistic electron transport from the surface layers represents an important sink for the deposited laser energy (see for example, Ref. [32]).

9.2.4 Electron–Phonon Coupling

The hot electron gas created by electron– electron collisions then couples to the phonons of the substrate, which eventually leads to a dissipation of the energy into a second heat bath. A theoretical description of this process has only in recent years become possible. Usually, the structure of a solid and the motion of its atomic constituents is very successfully described by the Debye model using quasi-particles, the phonons. For the electronic system, on the other hand, Bloch waves describe the motion and energetics of the quasi-particles. In both descriptions they are considered independent of each other, and this constitutes the usual Born–Oppenheimer approximation, which is well known from molecular physics. In this picture an electron–phonon interaction cannot take place, and therefore one must go beyond the Born–Oppenheimer approximation and formulate a description of these nonadiabatic processes.

As a formal derivation of the approach is beyond the scope of this chapter, the reader is advised to consult the appropriate literature, which is based on the Fröhlich description [61] of the electron–phonon (e–ph) interaction [62]. Physical insight into the problem is gained with the introduction of the so-called Eliashberg function, $\alpha^2 F(\omega)$, the product of the phonon density of states $F(\omega)$ and the electron–phonon interaction strength α^2 [63]. For the emission of a phonon it is given by

$$\alpha^2 F_{i,k_i}(\omega) = \int \mathrm{d}^2 q \sum_{f,\nu} \left| g^{i,f}_{q,\nu} \right|^2 \delta(E_{i,k_i} - E_{f,k_f} \pm \omega_{q,\nu}) \delta(\omega - \omega_{q,\nu}) \tag{9.5}$$

where $g^{i,f}_{q,\nu}$ denotes the electronic part of the interaction [62]. For the evaluation of $g^{i,f}_{q,\nu}$ one has to enter the screening potential. In order to illustrate this Eliashberg function, its dependence on phonon frequency is shown in Figure 9.2 for Be [64]. It transpires that the simple Thomas–Fermi screening yields results which are well compatible with experimental data. The electron–phonon coupling parameter $\lambda(E_i, k_i)$ is then obtained from

$$\lambda(E_i, k_i) = 2 \int_0^{\omega_{\max}} \frac{\alpha^2 F_{i,k_i}(\omega)}{\omega} \mathrm{d}\omega \tag{9.6}$$

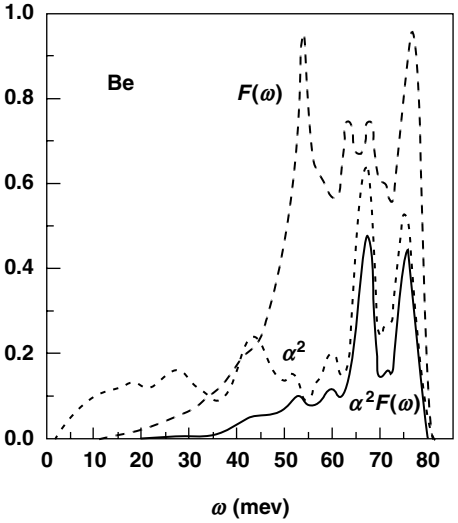


Figure 9.2 Phonon density of states $F(\omega)$ and Eliashberg function $\alpha^2F(\omega)$ for Be(0001). (From Ref. [64].)

Any phonon-induced interaction should be proportional to the occupation number $n_B(\omega)$ of phonons. The spectral broadening due to e–ph coupling is then

$$\Gamma(E_i, k_i) = 2\pi \int \alpha^2 F_{i,k_i}(\omega)[(2n_B(\omega) + 1) + f(E_{k_i} + \omega) - f(E_{k_i} - \omega)]\mathrm{d}\omega \qquad (9.7)$$

For increasing temperature this occupation number follows

$$n_B(\omega) \rightarrow k_B T/\omega, \qquad (9.8)$$

and thus shows at high temperatures a linear dependence on T [65]. The spectral broadening of a state due to e–ph coupling can then be written as

$$\Gamma(E_i, k_i) = 2\pi\lambda(E_i, k_i)k_B T. \qquad (9.9)$$

In order to calculate this width, it is necessary to know the phonon density of states $F(\omega)$ and the e–ph coupling strength $|g^{i,f}_{q,n}|^2$, given by [18]

$$g^{i,f}_{q,\nu} = \frac{1}{\sqrt{2M\omega_{q\nu}}} \langle \psi_{k_i} | \hat{\varepsilon}_{q\nu} \cdot \nabla_R V_{sc} | \psi_{k_f} \rangle \qquad (9.10)$$

where $\bigtriangledown_R V_{sc}$ is the gradient of the screened potential, $\hat{\varepsilon}_{q\nu}$ the polarization of the phonons, and M the atomic mass. ψ_{k_i,k_f} denotes the electronic wave functions of the initial and final states.

Experimentally, this e–ph coupling manifests itself via the spectral width Γ of states. An ideal situation to study this coupling arises therefore from surface states located in projected bulk bandgaps, as on the {111} surfaces of noble metals or of quantum-well states [66]. In this case, e–e scattering and electron–defect scattering may also contribute to the total linewidth Γ, and hence it is necessary to

seek alternative ways to separate these contributions. The e–ph coupling is also important when describing the cooling of a hot, nonequilibrium electron gas created by an intense laser pulse, as described above. Here, a connection between the microscopic constant λ and the phenomenologically used value g_∞ is given by [67]

$$g_\infty = \frac{3\hbar\gamma}{\pi k_B}\lambda\langle\omega^2\rangle, \tag{9.11}$$

where $\langle\omega^2\rangle$ denotes the second moment of the phonon spectrum.

9.3 Experimental

The investigation of electronic dynamics on nanostructures requires first an *in situ* preparation of the nanostructured system. For this purpose, a surface science ultra-high-vacuum machine with standard techniques for cleaning and preparation, like low-energy electron diffraction (LEED) and Auger electron spectroscopy, must also be equipped with instruments for either surface structuring or controlled growth techniques. Further, an *in situ* control and manipulation of the sample is suggested, which can be achieved by using scanning electron microscopy (SEM), scanning tunneling microscopy (STM) [68] or atomic force microscopy (AFM) [69] or by photoelectron emission spectroscopy [70] and microscopy [11]. Each of these techniques has in the past been described extensively, and the reader is referred to respective review articles.

In order to study the ultrafast electron dynamics, an appropriate laser source must be added, as well as a means of detecting the laser-generated photoelectrons. This can be achieved by a conventional dispersive electron spectrometer, equipped with two-dimensional (2-D) signal detection, or by a time-of-flight (ToF) detector, which makes use of the multiplexing advantage, especially when a multi-anode assembly is added. Figure 9.3 shows, in schematic form, a typical experimental set-up. The standard laser system consists of a Ti : sapphire laser oscillator operating in the spectral vicinity of 800 nm and with pulse durations of typically 12–25 fs. As the work functions of typical metals range from about 4 to 6 eV, frequency doubling and tripling in optically nonlinear crystals, such as β-BaB_2O_4 (BBO) and $LiBO_3$ (LBO), is often employed. For an optimal time resolution the frequency doubling crystal should be thin, with about 100–200 μm thickness depending on the spectral width that is to be converted. Efficiencies to produce 400 nm radiation in the range of 15–20% can be achieved. The third harmonic of the 800 nm fundamental radiation is produced by sum frequency mixing the fundamental with its second harmonic in a second BBO crystal. The frequency-converted pulses should be recompressed to compensate for the material dispersion in both the doubling and mixing crystals as well as the chamber windows, or any other optics in the beam path from the laser to the sample (see Figure 9.3). Besides simply reducing the number of photons required to reach the vacuum level of a system, the use of two colors – one from the fundamental and one from a

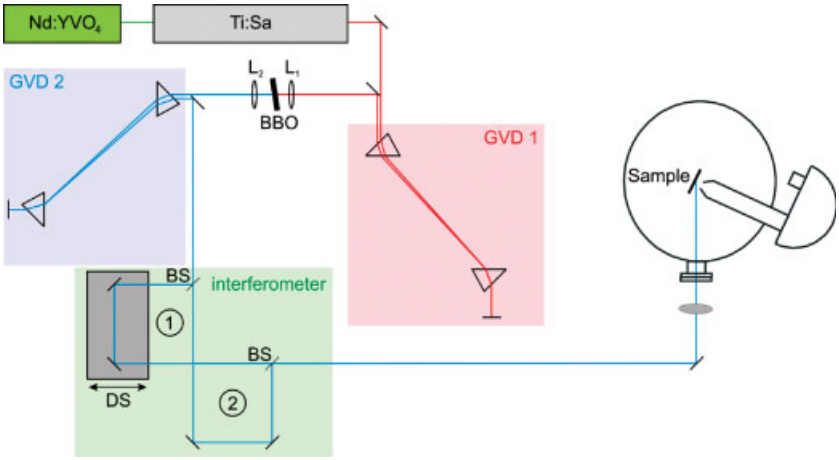


Figure 9.3 Schematic set-up for space and time-resolved two-photon photoemission experiments. The sketched dispersive hemispherical electron energy analyzer can be replaced by a time-of-flight detector or an imaging PEEM. Pulse compressors (GVD 1 and 2) (pre-)compensate a pulse stretching in optical elements in the path of both the fundamental and an harmonic beam.

frequency-doubled or -tripled beam – leads to a significant increase in the analytical power of the experiment.

Amplified Ti : sapphire radiation is required when an optical parametric oscillator is to be pumped in order to obtain radiation that is tunable across the visible spectrum. This allows one specifically to tune to and excite unoccupied resonance states of the sample. The dispersion of a state can also be followed. Amplified Ti : sapphire radiation with pulse energies in the 1 mJ regime and durations of 20–35 fs can also be used to produce extreme ultraviolet (XUV) radiation by high harmonic generation, with tunable photon energies up to about 100 eV and beyond [71–74]. Besides accessing more strongly bound states, or even shallow core levels, this radiation also allows the momentum space of a state to be addressed in the whole Brillouin zone. In addition, high-lying plasmonic states of all metals can, in principle, be reached directly. By shortening the pump pulse duration in the sub- 10 fs regime, using nonlinear optical techniques, it may be possible to produce not only high harmonics in the water window spectral range ($\lambda \sim 4.4$ to 2.3 nm) [71, 72]. By employing carrier envelope phase-stabilized pulses and correctly selecting the XUV spectral range, pulse trains of approximately 100 as [75–78] or even single pulses with durations as short as 80 as [79] can be produced. Such an approach will undoubtedly open a future window for investigating the electron dynamics of screening and dephasing processes.

For time-resolved experiments, care must be taken to properly delay and recombine both pulses on the sample. For this, either a Mach–Zehnder or a Michelson interferometer set-up may be employed. A temporal delay of both pulses is achieved

with different path lengths for the two interferometer arms; this difference must be controlled at better than 150 nm when a time step resolution of 1 fs is to be achieved, as the pulse will travel twice through the delayed arm. When the two pulses are collinearly overlapped, interferometric correlation signals with a periodicity of the optical wavelength are obtained. Due to a cycle duration of about 1.3 fs for 400 nm light, a much smaller step size (20 nm) and higher stability and reproducibility is required. Then, instead of using a conventional motorized translation stage, a piezoelectrically driven stage and an interferometer that is actively controlled with a frequency-stabilized HeNe laser beam are employed [80]. It is also important to control the polarization (s, p or circular) of both beams applied to the sample. The performance of both phase-resolved and phase-averaged 2PPE has been tested on a polycrystalline tantalum film (see Figure 9.4; photon energy 3.1 eV) [81]. In this way, the oscillation fringes due to the interference between pump- and probe- pulse are clearly resolved; that is, the accurate reproduction and periodicity of these measurements over the entire temporal delay proves the position stability of the set-up employed.

Due to increasing interest in the specific hot-electron dynamics of spatially heterogeneous systems such as metallic nanostructures, an extension to a space- and time-resolved 2PPE set-up, using time-resolved photoemission electron microscopy (TR-PEEM) for 2-D-electron detection has been established [13]. This method is capable of high spatial resolution in the 20 nm regime, which enables the focus to be set on the details of an individual nanoparticle. A typical TR-PEEM experiment is shown schematically in Figure 9.3 which is, except for the detector, identical to that of a TR-2PPE set-up. The TR-PEEM method has been well reviewed [11]. For a full pump-probe scan, the delay between the two pulses is varied in small steps (typically $\Delta t = 1$ fs), and for each step a PEEM image is taken

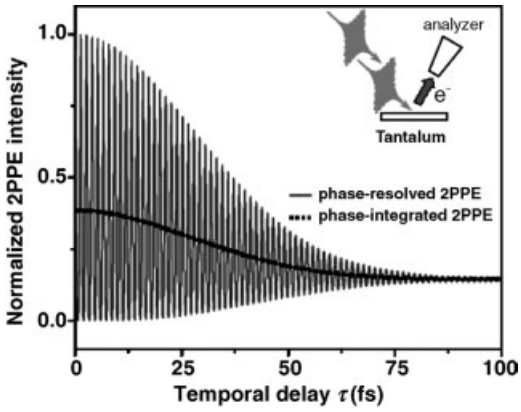


Figure 9.4 Time- and phase-resolved 2PPE from polycrystalline tantalum measured at an electron kinetic energy of 6 eV. The gray line is the 2PPE interferogram; the black dotted line shows the data for a conventional phase-averaged 2PPE measurement. (From Ref. [81].)

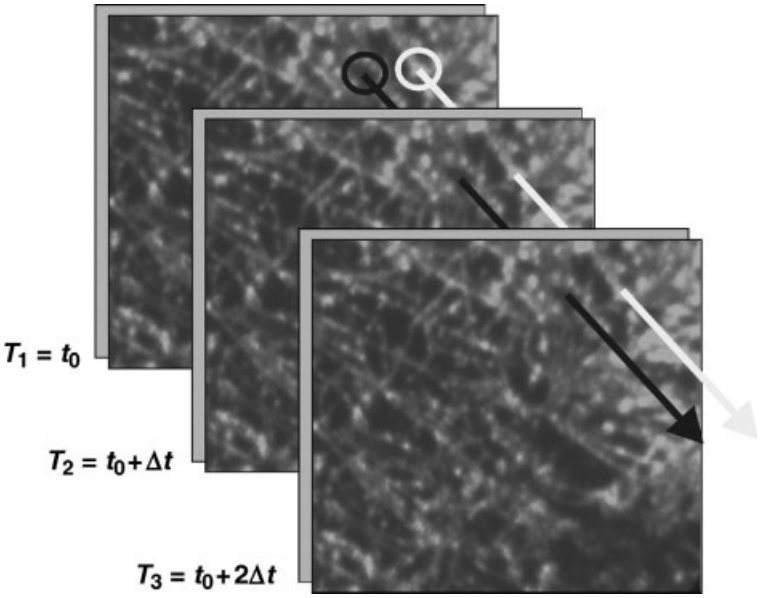


Figure 9.5 Schematics of measuring time-resolved 2PPE pump–probe images obtained with a PEEM.

(see Figure 9.5). This results in a series of images that contains a correlation trace for each pixel (Figure 9.6); these can then be plotted in a lifetime map deduced from a pixel-wise analysis of the cross-correlation traces of a TR-PEEM scan, as shown in Figure 9.7. The lifetime map contains information on the dynamic behavior of the electron system at the sample surface (decay time T_1 of the intermediate state) with the spatial resolution of the PEEM.

Nanostructured samples are, in general, prepared using either electron beam lithography (EBL) for larger structures (<40 nm), or cluster deposition for nanostructures as small as <1 nm. EBL, as a lift-off process, allows the controlled and

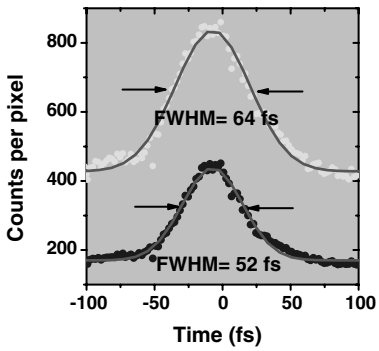


Figure 9.6 Typical autocorrelation traces of individual pixels of the PEEM images of Figure 9.5.

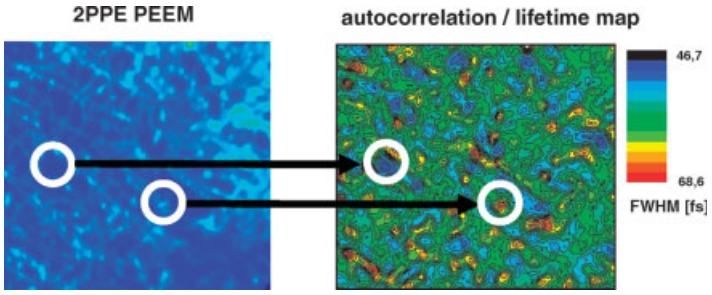


Figure 9.7 2-D autocorrelation map (right) obtained from time-resolved PEEM images taken according to Figure 9.5.

flexible design of metallic nanoparticles with regards to their shape and size. The optical properties – and especially the position and width of the LSP-resonance – depend critically on the shape and size of the particles, and this allows the characteristic LSP resonance frequencies to be tuned to the wavelength regime accessible by the available femtosecond laser system, e.g., the fundamental and second harmonic of a Ti : sapphire laser. Figure 9.8 shows SEM images of different silver nanostructures deposited on indium tin oxide (ITO) -covered glass substrates [82]. The dimensions of the elliptically shaped silver nanoparticles in Figure 9.8a are 140 nm (long axis), 60 nm (short axis) and 50 nm (height). These constitute versatile samples for investigating variations in the LSP decay in respect of resonant or off-resonant excitation. The silver nanodot array (Figure 9.8b; diameter 200 nm, height 50 nm) and the silver nanowire array (Figure 9.8c; length 1.6 µm, width 60 nm, height 50 nm) can be used to illustrate the potential of the time-resolved PEEM technique to map retardation effects associated with a plasmon excitation at nanometer resolution. Studies of the plasmon-induced coupling between neighboring nanoparticles are possible with nanodot pairs of varying center-to-center spacing. Figure 9.8d shows an example of 50 nm dimers (height 40 nm) at an interparticle spacing of 130 nm (grating constant 740 nm).

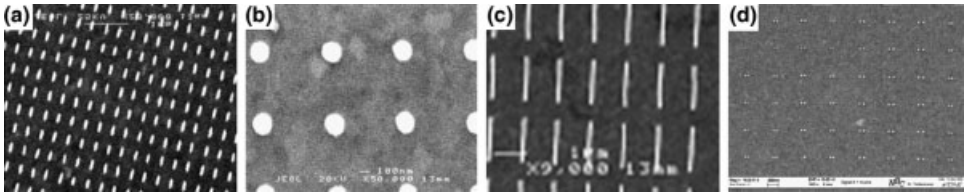


Figure 9.8 SEM images of different Ag nanostructures on an indium tin oxide (ITO) substrate. (a) 140 nm × 60 nm × 50 nm height; (b) Ø 200 nm × 50 nm height; (c) l = 1400 nm, w = 60 nm, h = 50 nm; (d) pairs with Ø 50 nm × 40 nm height and a separation of 130 nm.

For small cluster deposition onto surfaces, an UHV gas-aggregation cluster source is required including a quadrupole mass selector [83]. This allows the flexible *in situ* preparation of monodisperse cluster distributions over a broad size range. An alternative possibility would be to produce a defined density of atomic defects in a graphite surface, known as the Hövel-method [84]. This involves using a focused ion beam (FIB) technique that allows for the etching of well-defined nanopits into highly ordered pyrolytic graphite (HOPG) substrates and a subsequent oxidation procedure. The pits have a statistical variation in their depth between one and three monolayers (MLs). Further details on the FIB technique and oxidation procedure are available in Ref. [85]. The evaporation of about four MLs of silver at room temperature results in the condensation of near-monodisperse silver clusters in the native and artificially created defects in the HOPG surface.

9.4 Relaxation of Excited Carriers

During the past two decades, the study of image potential states has evolved as a paradigm for the investigation of electron dynamics at surfaces. This field has recently extensively been reviewed [86], and therefore only selected aspects will be discussed here. The main decay channel of these states isolated in a directional bandgap is the overlap of their wave function with bulk electronic wave functions. This overlap increases as the energy of the states approach the band edges. Then also the phonon contribution to the decay rate increases. For well-isolated image states, such as the $n = 1$ state on Cu(100), this contribution to the linewidth is expectedly very low, with a coupling parameter of $\lambda \sim 0.01$ and a contribution to the spectral width of $\Gamma < 1$ meV [87, 88], while for Cu(111) λ increases to 0.06 [89].

On the other hand, interband and intraband electron scattering also constitute important scattering mechanisms for image potential states. Steps and adatoms on the surface play an important role in these scattering phenomena. For example, on stepped surfaces scattering within the $n = 1$ image state results in a broadening of the level and thus in a shortening of the lifetime [90]. On vicinal Cu surfaces with (111) terraces, Roth *et al.* [91] found that for electrons with a downwards momentum the steps show a greater decay rate than in the opposite direction, and this results in lifetime differences within $n = 1$ of up to 4fs. Copper and cobalt adatoms induce intraband and interband scattering between the image states [92, 93]. An inelastic interband scattering of $n = 2$ electrons with bulk electrons populates the bottom of the $n = 1$ band. When probing levels within the $n = 1$ band above the bottom of the $n = 2$ level, a resonant (quasi-)elastic interband scattering occurs from $n = 2$ to $n = 1$, with strongly increasing probability as the energy increases further. On the other hand, the adatoms do not have any significant influence on the intraband scattering rate. It can be envisioned that a lateral confinement of nanostructures might also lead to an increased coupling between image states and bulk bands due to a shift of the band edges.

An especially illustrative example for a hole decay has been provided by Berndt and coworkers, who studied the occupied surface states of noble metals by using

Table 9.1 Experimental and theoretical hole lifetimes of occupied surface states on noble metals. (From Ref. [94].)

				Experiment		Theory	
Metal	**Δ (meV)**	**β**	**τ (fs)**	**Γ_{STM} (meV)**	**Γ_{PES} (meV)**	**Γ_{old} (meV)**	**Γ_{new} (meV)**
Ag	8	0.89	120	6	20[a]	5.3	7.2
Au	23	0.82	35	18	60[b]	8.6	18.9
Cu	30	0.80	27	24	21[c]	10.2	21.7

[a] Ref. [70].
[b] Extrapolated to $T = 0$ K from Ref. [95].
[c] Ref. [96].

STM [94]. The width Δ of the onset of dI/dV spectra can be transformed into lifetimes τ via

$$\tau = \beta h/(4\Delta), \tag{9.12}$$

where h denotes Planck's constant and β is a scale factor close to unity. Using this method, it can be ensured that the surface area studied is indeed clean, and therefore defect scattering can be excluded. The results obtained, together with a comparison of values derived from photoelectron spectroscopy [70, 95, 96], and a comparison with theoretical calculations, is shown in Table 9.1. The agreement between these measurements and the latest theoretical calculations is excellent. The results show that the holes created in the surface state are filled by electron–electron scattering from the 2-D electron gas of the still-occupied surface state band, rather than from the underlying 3-D bulk electron gas.

The global relaxation dynamics of a hot, nonequilibrium electron gas can be assessed by measuring the energy distribution function as a function of delay time after creating the hot distribution. This may be exemplified for hot electrons in Ru as studied by Wolf and coworkers [32]. The group investigated the relaxation dynamics at different pump pulse fluences. The relaxation dynamic is then modelled using a modified two-temperature model, where the electron distribution was parametrized by splitting it into a thermalized component at low temperature, while a second component was also thermalized, albeit at a higher temperature and with a lower population. This procedure describes the nonthermal electron energy distribution quite well. Both electron distributions couple to the phonon heat bath (for details, see Ref. [32]). Besides the expected localization of energy at the surface by electron–phonon coupling (as discussed above), it has also been found that ballistic transport out of the surface region has a significant influence on the temperatures and their time dependencies in the surface region. This leads in turn to notably lower temperatures (as shown impressively in Figure 9.9), with corresponding consequences for electronic coupling to adsorbates and laser-induced desorption dynamics. Such transport effects have also been observed for copper surfaces [97].

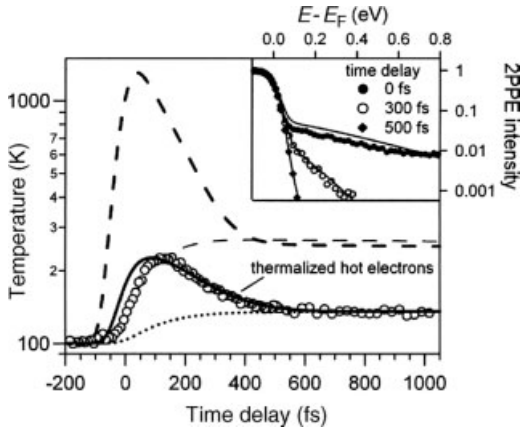


Figure 9.9 Time evolution of the electron and phonon temperatures according to the two-temperature model (thick and thin dashed lines). The extended heat bath model yields significantly lower values (shown by the thick and dotted lines) for the electrons and phonons, respectively. The open dots are experimental results. The inset shows measured electron kinetic energy distributions at different delay times. Data and calculations are for a Ru surface with a pump fluence of about 500 μJ cm^{-2}. (From Ref. [32].)

As mentioned above, on a microscopic level the electron–phonon coupling phenomena are best studied via isolated states in the band gap of materials. By using high-resolution photoelectron spectroscopy, the occupied surface states of noble metals have been studied [96, 98]. Figure 9.10 shows the experimental linewidth of the surface states, together with theoretical calculations [99], whereby an excellent agreement is obtained. This figure also shows the theoretically determined contribution of e–e scattering (Γ_{ee}) to the total width of the states. At low binding energies of the hole, the contribution of the Rayleigh phonon mode is dominant, while the overall spectral width becomes small. A similarly good agreement between theory and experiment was obtained previously for the very isolated surface state at Γ on the Be (0001) surface [100]. Here, the main contribution comes again from intraband scattering.

When defects are present on the surface, an additional broadening occurs. Such an effect has recently been reported for the Au (111) [101, 102] and Al(111) occupied surface states [102]. By monitoring the deviation of broadening from the expected linear temperature dependence, an activation energy for the creation of surface defects can be derived

$$\Gamma_{ed} = C \exp\{-[E_a/k_B T]\}. \tag{9.13}$$

On aluminum, the experimentally determined value for the activation energy of about $E_a = 170$ meV compares well with theoretical expectations for the creation of defects by kinks on a step edge, while the creation of adatoms requires about 300–600 meV. The corresponding value for Au(111) amounts to $E_a = 81$ meV, which suggests that kinks at step edges might well be responsible for the observed broadening.

In image states such a defect scattering can be isolated from the e–ph scattering, because these states are located in front of the surface and therefore are detached

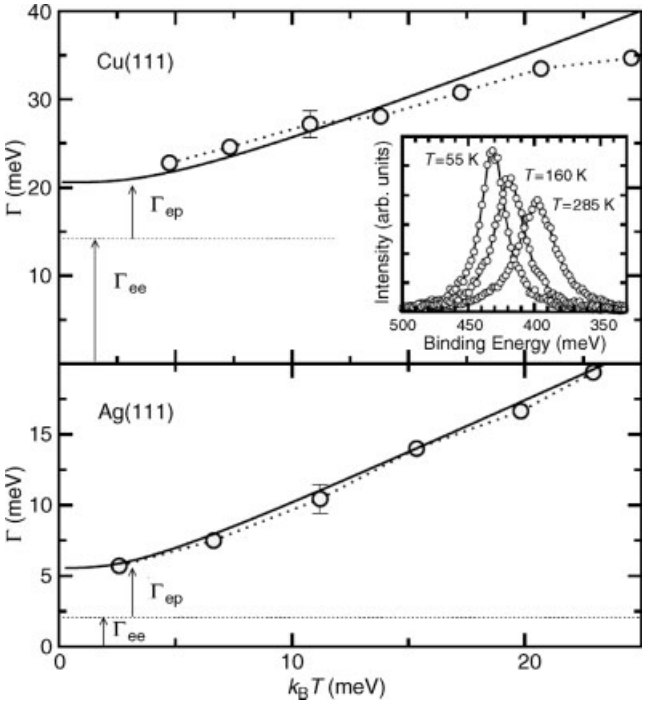


Figure 9.10 Temperature dependence of the experimental and theoretical line width of the occupied surface state on Cu(111) and Ag(111). Γ_{ee} denotes the e–e scattering contribution to the line width, independent of temperature. The inset shows experimental photoemission spectra at the Γ point at selected temperatures. (From Ref. [99].)

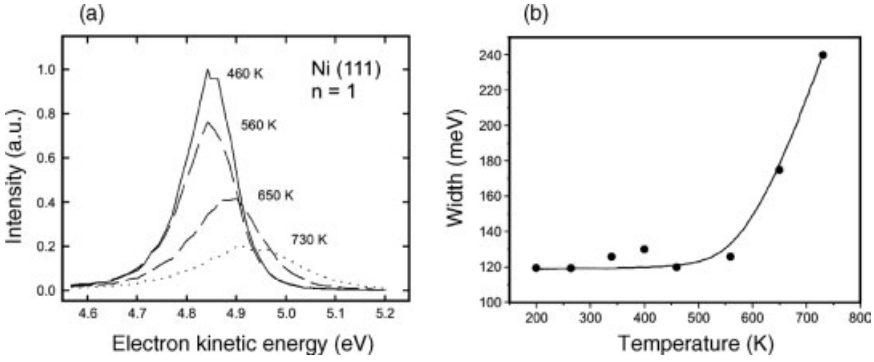


Figure 9.11 (a) Spectra and (b) line width of the ($n = 1$) image state on Ni(111) as a function of temperature. The state is resonantly excited from the occupied surface state at $E_b = 0.2$ eV.

from the motion of the atomic surface constituents. A similar strong temperature dependence of the width of an isolated state has been observed for the ($n = 1$) image potential state on Ni(111) (H. Zacharias and R. Paucksch, unpublished results). In this experiment, one photon at $\lambda = 263$ nm ($h\nu = 4.7$ eV) resonantly excites the image state from the occupied surface state at $E_b = 0.2$ eV [103], while a second photon of the same energy liberates the excited electron. Figure 9.11 shows the resonant spectra of $n = 1$ and the observed widths as a function of temperature. Above 450 K, an exponential growth of the spectral width is observed. Based on this temperature dependence, an activation energy of about $E_a = 130$ meV was derived, this again being in agreement with the creation of kinks at step edges.

9.5
Volume Excitation in Metallic Nanostructures Investigated by TR-PEEM

The potential of TR-PEEM as a versatile tool for mapping the electron dynamics of metallic nanoparticles was introduced by Schmidt *et al.* [13]. Figure 9.12a shows the investigated lifetime map of a patterned silver film (hexagon-shaped patches) on a silicon substrate. The mapped color variations correspond to a change in the full-width half-maximum (FWHM) of the correlation trace for each pixel, which becomes more evident from a statistical analysis (histogram) of the investigated regions, as shown in Figure 9.12b.

Figure 9.13a shows a PEEM image of a Ag nanoparticle array (as shown in Figure 9.8b) at resonance excitations. The 400 nm laser light used for the TR-2PPE experiment couples almost resonantly to the in-plane mode of the particle. The 2PPE image shows distinct interparticle brightness variations which are dominated by defect-induced indirect transitions rather than by differences in the collective electron response [104]. Although the properties of the particles' plasmon resonances are not affected by this, they must of course be included in the analysis of the time-resolved data. The lifetime map, as shown in Figure 9.13b, visualizes the lateral

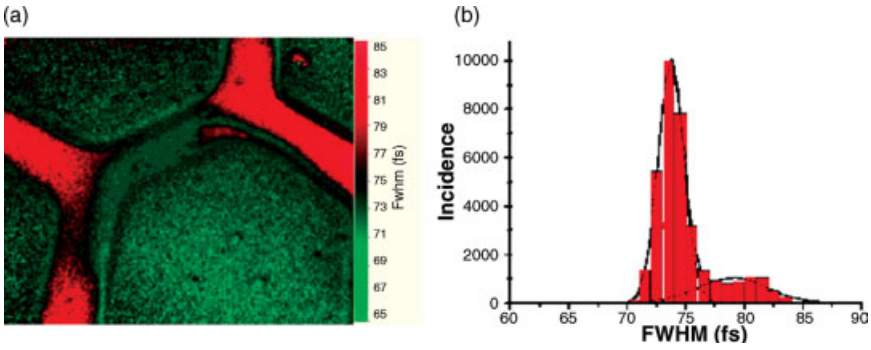


Figure 9.12 (a) Lifetime map of a hexagonal silver nanoparticle; (b) Lifetime map distribution showing a most probably value of 73.5 fs with a distribution width of ± 1.5 fs.

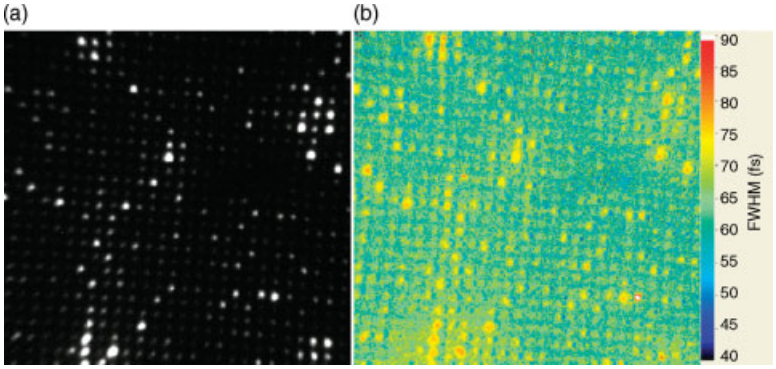


Figure 9.13 (a) PEEM image and (b) corresponding lifetime map of the nanostructure from Figure 9.8b. The FWHM of the autocorrelation curves ranges from 60 to 90 fs.

variations in the electron dynamics in color coding. Here, the red tones correspond to high FWHM values, indicating long decay times, while blue tones are associated with a faster decay. There is certain correlation between brightness in the 2PPE image and the FWHM values of the lifetime map: those particles which appear bright in the 2PPE image tend to exhibit longer average decay times in the lifetime image. This effect must obviously be related to the defect induced transitions. Whilst on the one hand the involved intermediate single-electron states give rise to a higher overall transition probability and a higher photoemission yield, on the other hand they show longer decay times.

The dynamic processes associated with a plasmon excitation in a single particle can be studied in further detail when the phase-resolved set-up is employed. The image

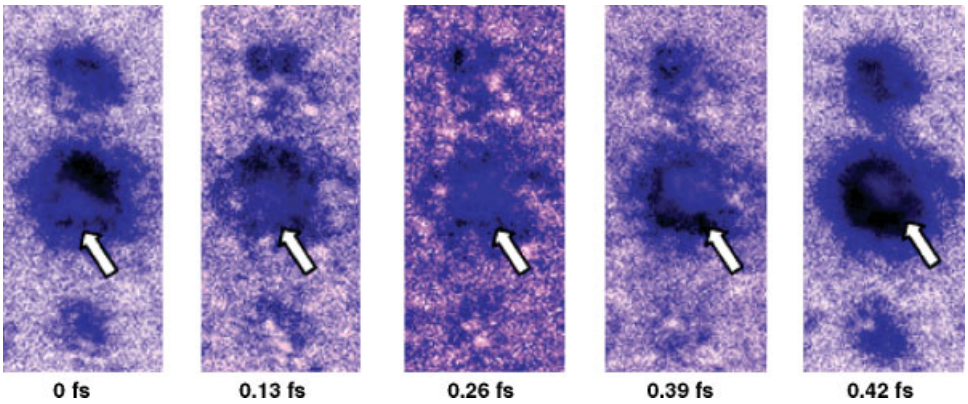


Figure 9.14 Spatiotemporal femtosecond dynamics of a silver nanoparticle with a diameter of 200 nm. A modulation of the lateral photoemission distribution as a function of phase delay between two identical exciting femtosecond laser pulses for 2PPE is observed. This is assigned to a phase propagation of a plasmon through the nanoparticle.

sequence in Figure 9.14 shows the plasmon dynamics in a single particle, where the time interval between the two images is 0.13 fs. A clear variation in the contrast within the area of the nanoparticle in the sub-femtosecond time scale is detectable. The result can be explained in the following way: The electric field amplitude is determined by the phase delay $\Delta\varphi(\tau)$ between the pump and the probe laser pulse, as adjusted by the Mach–Zehnder interferometer, as well as the polarization field of the particle plasmon which is oscillating at its resonance frequency. Due to oblique incidence from the right, the laser light would be expected to couple first to the LSP-mode at the right edge of the particle. Here, the external (laser) field and internal (plasmon) field attain a fixed phase relation. As the propagation velocity of the external and internal fields vary, a position-dependent phase lag between the two field components is acquired as the plasmon excitation travels through the particle. The particle internal structure visible in a single PEEM image of Figure 9.14 is a residual of the varying interference between the external light field and the particle internal LSP-field, directly connected to the plasmon phase. The parallel data acquisition by the PEEM allows any systematic errors to be excluded.

These results are in good agreement with the findings by Kubo *et al.* [105, 106], who also combined ultrafast laser spectroscopy and electron microscopy in order to image the quantum interference of localized surface plasmon polariton (SPP) waves with sub-wavelength spatial resolution and sub-femtosecond temporal precision. The sample used was based on a 400 nm-thick silver film perforated by an array of 100 nm-wide slits with a period of 780 nm. This approach resulted in a silver grating, which had the properties of an optical band-pass transmission filter. The polycrystalline grating creates nanoscale roughness, in which localized plasmon modes can be excited. So, by scanning the time delay between identical, phase-correlated pump and probe pulses in 174 optical cycle steps, and recording the resulting change in the polarization interference pattern, a movie of the SPP propagation wave packet at the Ag–vacuum interface can be created, as shown in Figure 9.15. As the driving pulse wanes, the coherent polarization excited at each dot shifts to its own resonant frequency. For instance, as shown in Figure 9.16, the phase of dots A, B and D (dot C) is retarded (advanced), causing the intensity maxima to rise later (sooner) with respect to the phase of the driving field. The circled hotspots in Figure 9.15 indicate the change in the intensity maxima (constructive interference) in five cycle intervals due to the phase slip of the surface plasmon modes with respect to the driving field.

The SPP wave packet propagation length – and hence coherent control studies – can be improved by using single-crystal nanostructures, as demonstrated by L.I. Chelaru *et al.* [107]. Figure 9.17 shows an example of the appearance of SPP waves in Ag single-crystal particles that were formed by self-assembly [108]. In Figure 9.17a, a SPP wave is started at the marked edge of the triangular island, and travels across the island during the time of observation. The striped pattern on the otherwise dark islands is a representation of the SPP wave by means of a beat pattern formed between the propagating plasmon wave and the laser pulse used to probe the structure [109]. Modulation of the local near-field in the surrounding of the island is caused by diffraction of the laser pulse. In Figure 9.17b, the two independent SPP waves are created at the edges of a hexagonal Ag island. The SPP waves superpose,

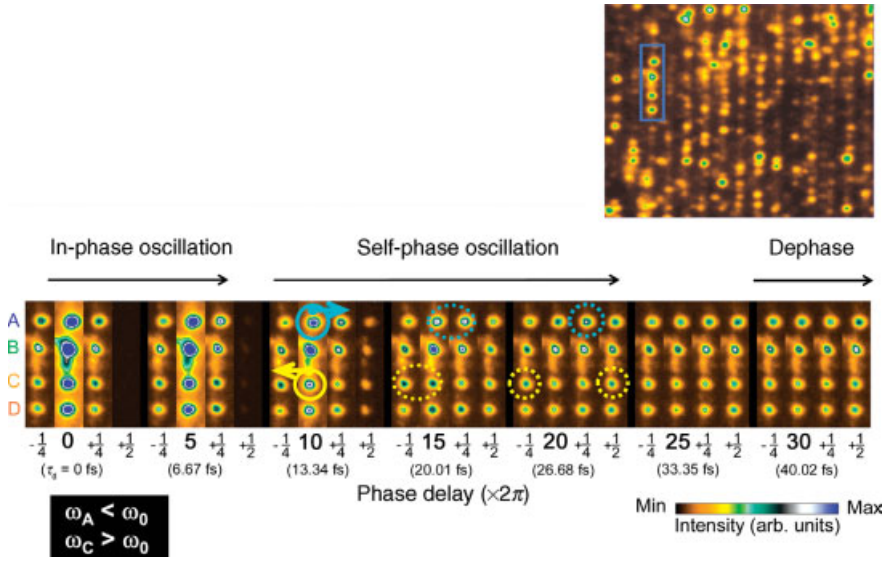


Figure 9.15 Interferometric TR-PEEM image of four localized surface plasmons on a Ag grating. The delay time between pump and probe pulses is advanced from −0.33 fs to + 40.69 fs in π/2 steps (0.33 fs) of the carrier wavelength of 400 nm. During excitation (up to 5½ × 2π) all dots oscillate in phase; later the phase in dots A, B and D is retarded, but in dot C it is advanced compared to the exciting laser field. (Reproduced from Ref. [105].)

modulate the overall local electric field, and are reflected at the end of the structure. This is different in panel Figure 9.17c, where the angle between the two overlapping SPPs is smaller and the modulation of the electric field strength is much more pronounced. The field strength at the edge of the particle is strongly modulated and

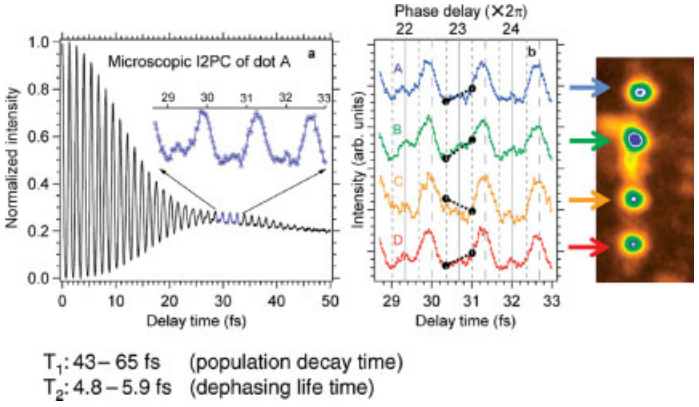


Figure 9.16 Phase slip between the LSP and the advancing light field of dots A to D in Figure 9.15. (From Ref. [105].)

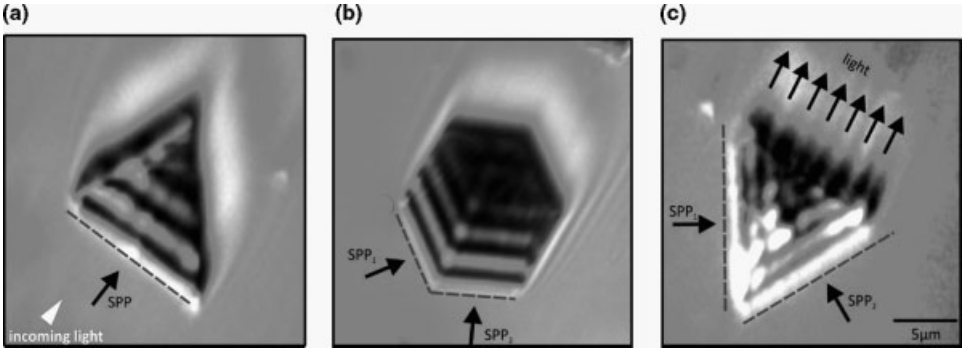


Figure 9.17 SPP waves on silber islands observed by PEEM illuminated under $\theta = 75^{\circ}$. (a) Triangular shape; (b) Hexagonal islands with SPP interferences; (c) Interference between two SPP waves where the island acts as beam splitter for the SSP wave. The scale bar represents 5 µm. (From Ref. [108].)

the SPP is converted back into light only at those positions where the field strength is particularly high. Ultimately, the island in panel Figure 9.17c acts as a beam-splitter. For particles such as those shown in Figure 9.17, propagation of the plasmon with ~60% of the speed of light is observed as a systematic shift of the beat pattern as a function of the delay time between pump pulse and probe pulse [106].

9.6 Long-Lived Resonances in Adsorbate/Substrate Systems

Adsorbate atoms may not only serve as general scattering centers on well-prepared surfaces but also shorten the lifetime of well-defined surface states. When the electrons do scatter resonantly into unoccupied states of adsorbates, or when such states are directly optically excited, the lifetimes may be significantly prolonged compared to those expected from a simple Drude or Fermi liquid picture for electrons at the same excitation energy. A prominent example is the adsorption of alkali atoms on noble metal surfaces. In the low coverage limit an adsorbate-induced antibonding (A) state around 2.5–3 eV above Fermi is found on noble metals. The experimentally observed relatively long lifetimes of excited adsorbate states of up to 50 fs at low temperatures for Cu(111) [110, 111] and Ag(111) [112, 113] have initiated a number of theoretical investigations of this effect.

Although the decay of such states could, in principle, be viewed as a one-electron resonant charge transfer, which would yield very short lifetimes in the sub-femtosecond regime, the directional band gap at these surfaces hinders a fast decay, because the overlap of the state with bulk wave functions is small. With one-electron wave packet propagation calculations this reasoning could be supported [114–117]. Figure 9.18 shows the differences of the wave packet propagation for the Cs 6s state adsorbed on a jellium and a Cu(111) surface. Initially, the wave packet propagates in

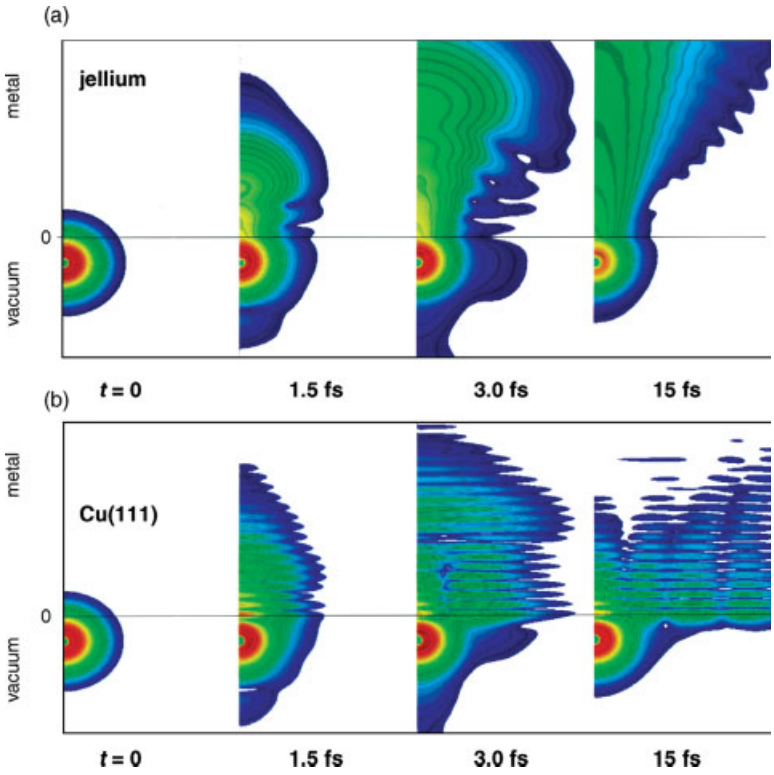


Figure 9.18 Dynamics of the Cs (6s) wave packet excited in front of (a) a jellium and (b) a Cu(111) surface. The Cs atom is placed at z = 10 a.u. from the image plane. A propagation along the surface normal into the bulk initially observed for both substrates continues only for the jellium surface. For Cu(111), propagation stops after only 3 fs, followed by a propagation in a high-$\vec{k}_{||}$ direction along the surface. A high intensity of the 6s wave packet remains in the vicinity of the surface. (After Ref. [114].)

both cases into the bulk. On jellium (Figure 9.18a), this propagation continues with minimal lateral spreading of the wave packet, but on Cu(111) (Figure 9.18b) the movement of the wave packet into the bulk has come to a halt after only 3 fs, followed by a propagation in high-$\vec{k}_{||}$ directions parallel to the surface, with the highest intensity remaining in the vicinity of the surface. Such behavior causes a relatively long resonance lifetime. Whilst these long-lived states were unexpected in the adsorbate/substrate systems, an understanding of the origin of long-lived states has opened some extremely interesting new possibilities towards the control of reactions at surfaces. Then, other processes may also come into play, including inelastic e–e scattering with bulk electrons, where the adsorbate atom begins to move on the new electronically excited potential energy surface. The excited electron then couples to the nuclear motion, and processes as in DIET or DIMET become important [118]. This type of motion has been observed directly and spectroscopically using TR-2PPE [119].

By taking into account the bandwidth of the ultrashort laser pulse and the nuclear motion of the adsorbate after excitation, an excellent agreement between experimental and theoretical lifetimes of the excited state for all alkalis studied has been achieved [114–117]. In differing from many other antibonding states of adsorbate/substrate systems, for the alkalis on noble metal surfaces this state can be populated by a direct dipole-allowed transition from the occupied surface state, also located in the same directional bandgap. This provides the opportunity to initiate the excitation with temporally and spectrally shaped pulses in order to control the desorption process [120]. For excitation laser pulses shorter than the lifetime of the excited A state, a complete population transfer to the A state can be achieved when dissipation is neglected. Including dissipation, the populations transfer falls to about 40% for a Gaussian pulse of 20 fs duration and a peak intensity of 0.5 TW cm^{-2}, while for an optimal pulse shape a transfer of 95% can still be achieved. In the case of longer and less intense pulses, the optimally controlled pulses are even more successful. Using again 20 fs-long Gaussian pulses, but now at only 10^{10} W cm^{-2} (which represents an experimentally feasible intensity at metal surfaces), only 3% population transfer is to be expected (in theory). For controlled pulses of 60 fs duration – that is, much longer than the lifetime of the A state – a transfer of 75% may still be achieved. In the control cases pulses are used which show the highest intensities at the extreme end of the pulse, although it will be difficult to produce such pulses experimentally. Nevertheless, even when taking the nuclear motion of the Cs atom on the excited potential into account, the use of shaped pulses might lead to an increased yield also for dissipative surface reactions.

With organic adsorbates, modified and new properties of nanostructures are obtained, the aim being to create functional materials in a controlled manner [121]. The organic constituents of such adsorbate layers provide a wide variability in the desired properties, which may range from sensor and molecular recognition to molecular electronics, photo-switches and molecular magnets. In most of these envisioned applications a nanostructured assembly of the functional materials is necessary. Many of the systems' properties rely on an heterogeneous electron transfer between the adsorbate and substrate and, in the case of sensor or recognition devices, between the active layer and the incoming molecules. The dynamic properties of these organic adsorbate films are therefore of fundamental interest for such functional materials.

Basic investigations of the action of alkane layers on the electronic properties of the underlying metals, as well as the electron localization in organic overlayers, have been carried out by Harris and coworkers [122, 123]. In the context of this chapter, the photo-induced electron transfer dynamics from (self-) organized and specifically bound organic molecules are of particular interest, because here the dynamics can be probed directly in the time domain. These processes also relate to light-harvesting applications, either for reactions or for the direct production of an electrical current.

Willig and coworkers investigated the influence of various anchor groups between a perylene-derived chromophore and the (110) surface of rutile TiO_2 on electron transfer [124]. These anchor groups were seen to serve two purposes: (i) to separate the electronic states of the chromophore and the solid surface; and (ii) to enable a

control of the electronic coupling strength between both systems. The anchor groups also stabilize the complex and provide the option of definite binding angles and sites. The electron transfer is initiated by a direct femtosecond laser excitation of the singlet state of the chromophore. In the case of one conjugate anchor group – acrylic acid, which may be viewed as a molecular wire – the electron transfer time is found to be about fourfold faster ($\tau \sim 13.5$ fs) than for an insulating group of the same length, propionic acid, which acts as a tunneling barrier. For an even wider barrier which consists of a three-ring structure, where the central ring is aliphatic and rigid, the transfer times are prolonged to tenfold that of the first type [124].

When the electron transfer time is short – which can be achieved by directly binding the organic molecule to the substrate – the electron migration inside the substrate can be monitored using TR-2PPE. The experimentally observed times for the electron to escape from the surface layers (or, more specifically, from the detection depth of 2PPE) agrees well with rates obtained from density functional theory (DFT) calculations for alizarin adsorbates [125]. For the system investigated, it turned out that in rutile the energy does not relax on a 200 fs time scale, which is tentatively assigned to the population of new interface states created as the bonds with the organic adsorbates are formed. Such electron escape to the bulk states of a substrate is important in terms of the total energy transfer from an organic adsorbate to unoccupied states of a solid. When this process is rapid, charge conduction across the interface is favored; however, when it is slow an accumulation of charge in excited states of the organic adsorbate, with the possibility of a radiative deactivation, may occur.

A somewhat unexpected feature of organic adsorbates derives from the existence of very long-lived triplet states with lifetimes in the microsecond range. As yet, such states have not been identified for small inorganic adsorbates or for bare surfaces. Here, the dynamics in unoccupied states of ordered C_{60} on metal surfaces will be discussed as a specific example of an organic adsorbate. Due to the large work function of C_{60} of 6.8 eV above the highest occupied molecular orbital (HOMO), and depending on the work function of the metal, the lowest unoccupied molecular orbital (LUMO) state of the first C_{60} ML may be either below or above the metal Fermi level. Therefore, the character of the C_{60} film may be either metallic or semiconducting [126]. For thicker layers above about 5 ML the properties of C_{60} crystals are generally assumed, in which the interaction energies of the individual C_{60} constituents are comparatively weak. Hence, the properties of the film resemble those of the C_{60} clusters. Such layer dependency already provides the possibility of tuning the properties of a functional film, whilst in addition a substitution at a C–C bond by functional groups enables further variability [127, 128].

The symmetry properties of C_{60} allow optical transitions from the HOMO to the LUMO + 1, and from the HOMO-1 to the LUMO states – which are actually comparatively broad bands rather than sharply defined single states. With the third harmonic of a Nd-laser at photon energies of about $h\nu = 3.5$ eV ($\lambda \sim 355$ nm), a direct excitation via both transitions is possible. The population both in these states and in the energetically lower lying excitonic states, which are populated via internal energy transfer, can be probed with the fifth harmonic of a Nd-laser at 210 nm ($h\nu \sim 5.88$ eV).

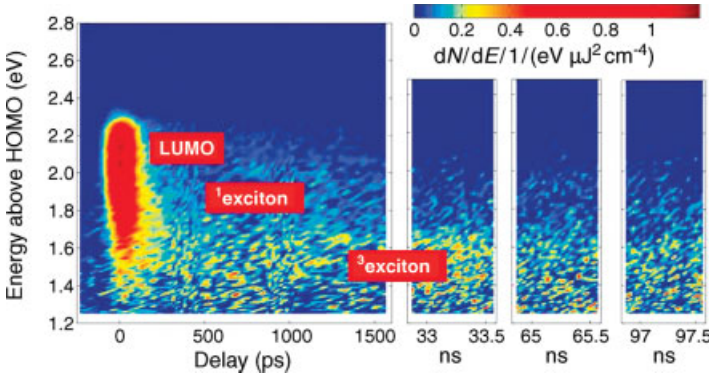


Figure 9.19 2-D lifetime plot for the LUMO, ^{1}exciton, and ^{3}exciton states on C_{60} up to a delay time of 100 ns. On this time scale the intensity of the ^{3}exciton decreases only marginally.

This second photon has insufficient energy to directly emit photoelectrons from the HOMO state. Figure 9.19 shows, in a 2-D plot, the kinetic energy distribution of emitted photoelectrons as a function of time delay between a pump and a probe laser pulse. In this case, the LUMO state has been directly populated in a dipole-forbidden transition, and it is evident that, besides the LUMO, lower-lying states are populated. Even for delay times exceeding 100 ns after pumping a signal intensity is observed for these energetically lower states. This long-lived state is assigned as a signature of the triplet exciton. In energetic terms, the singlet exciton is identified somewhat higher and below the directly excited LUMO.

An analysis of the intensity dependence at certain kinetic energies revealed for the LUMO a lifetime of about 84 ps – shorter than a previously reported value of 134 ps [129]. For the singlet exciton a lifetime of about 1050 ps was found, and was in good agreement with earlier reports [129]. In order to measure the lifetime of triplet excitons a second pump laser was used such that the probe laser could be electronically delayed with respect to the pumping laser. For these states, lifetimes between 22 μs for free and up to 200 μs for bound excitons were observed, depending also on the thickness of the C_{60} film (A. Rosenfeldt, B. Göhler and H. Zacharias, unpublished results). This was significantly longer than had been observed for a photo-polymerized C_{60} film [130], and in good agreement with lifetimes of these states in solution [131]. When the density of triplet excitons becomes high, the comparatively fast (spin-allowed) process of triplet–triplet annihilation can be observed (Figure 9.20).

It should be noted that these longlived states may be chemically active. Following the adsorption of NO molecules onto a thick, ordered C_{60} film, UV laser excitation of the system led to a desorption of the NO, with delay times of up to 200–400 μs after pumping. An analysis which yielded a chemically active state with a decay time of about 160 μs [132] was recently confirmed by directly measuring the velocity of the late-arriving molecules. Their velocity was found to be much greater than the corresponding arrival time at the detecting laser, which meant that these molecules

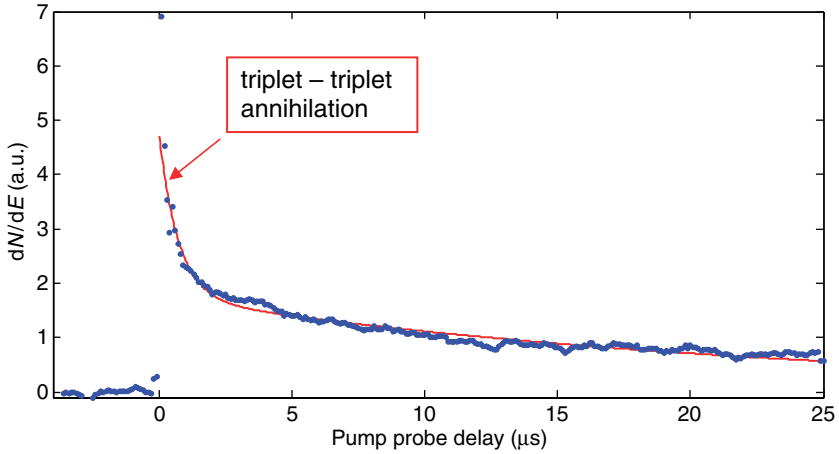


Figure 9.20 Lifetime of a C_{60} triplet excitonic state with $\tau = 22\,\mu s$ in a 14 ML film. Also indicated is an exciton–exciton annihilation occurring on a time scale of 720 ns.

were not desorbed promptly but rather at a later time (T. Hoger *et al.* unpublished results). Hence, these excitonic states in organic thin films may serve as efficient energy reservoirs for photochemical activity and reactions [133].

9.7 Outlook: Spatial and Temporal Control of Nano-Optical Fields

The optical response of nanostructures creates a variety of fascinating properties, including sub-wavelength variation of the field, local field enhancement and local fields with vector components perpendicular to those of the incident field. Moreover, the combination of ultrafast laser spectroscopy (i.e. illumination with broadband coherent light sources) and near-field optics continue to open new realms for nonlinear optics on the nanoscale. As an example, the spectral phase of the incident light will influence the peak intensity of the local field distribution [134], thus providing a means to manipulate the nonlinear response of a nanostructure. More recently, it has been shown theoretically that the interaction of polarization-shaped laser pulses with a nanostructure allows the simultaneous control of spatial and temporal evolution of the optical near-field distribution [135].

A first step towards the experimental demonstration of simultaneous spatial and temporal field control used adaptive control, combining multi-parameter pulse shaping with a learning algorithm, and demonstrated the generation of user-specified optical near-field distributions in an optimal and flexible fashion. Shaping of the polarization of the laser pulse provides a particularly efficient and versatile nano-optical manipulation method [136]. Additionally, pump-probe sequences can be generated, in which excitations occur not only at different times but also at different

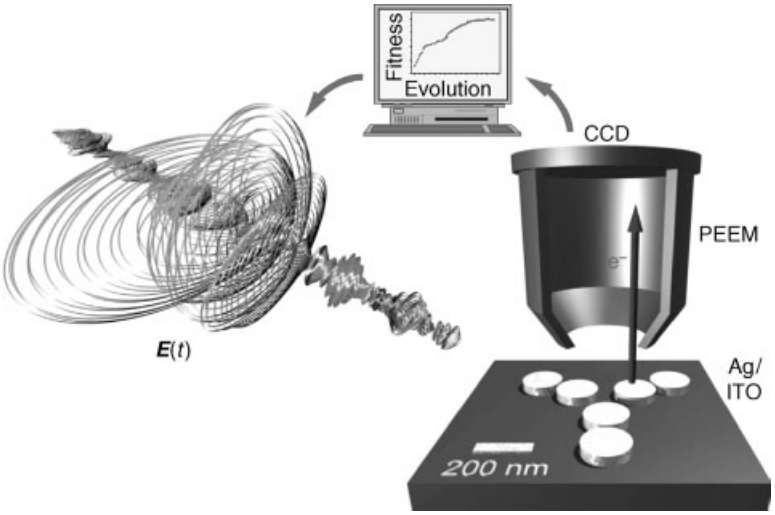


Figure 9.21 Experimental set-up for applying polarization-shaped femtosecond pulses to a nanostructure. A PEEM was used to measure the 2PPE signal.

positions that are separated by less than the diffraction limit. This in turn opens a route towards space–time-resolved spectroscopy with potential for the direct observation of nanoscopic energy or electron transport [135]. In a first experiment [137], femtosecond polarization shaping, adaptive optimization and two-photon PEEM were combined (see Figure 9.21). The pulse shaper contains a two-layer, 128-element liquid crystal display (LCD) spatial light modulator in the Fourier plane of a zero-dispersion compressor in folded 4f configuration [136]. Each LCD layer modulates the spectral phase of one of the two transverse polarization components, leading to a spectral variation of polarization state and phase. Hence, in the time domain the intensity, instantaneous oscillation frequency and polarization state (elliptical eccentricity and orientation) can be controlled to vary within a single laser pulse. Pulse-shape optimization occurs for both LCD layers independently. The data in Figure 9.22 show that adaptive polarization pulse shaping allows the optimization of a particular emission pattern. In Figure 9.22b and c, the ratio A/B between the emission yields integrated over the two rectangles A and B was optimized using an evolutionary algorithm. The experimental results showed that the local interference of the optical near-fields generated by the two orthogonal incident polarization components could be utilized to manipulate the local field distribution.

Currently, in these experiments two-photon PEEM is used to qualitatively monitor the local field distribution, with the interpretation of the acquired images being based on the assumption that the highest local field intensity will produce the highest photoemission yield in a 2PPE process. However, this yield is influenced by additional parameters such as the intermediate state lifetime in the metal, or the field component perpendicular to the surface. A detailed modeling of the emission process, together with a comparison with the experimentally observed emission

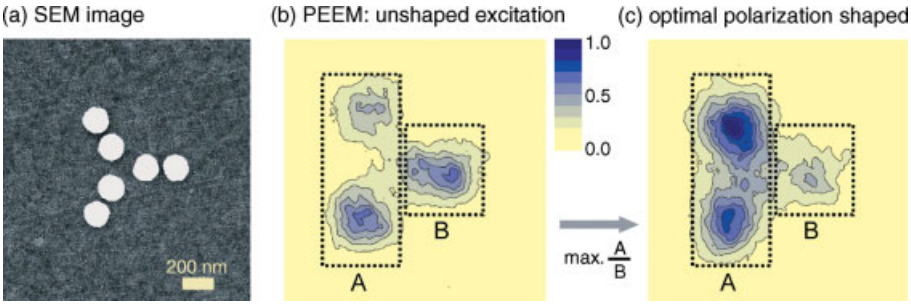


Figure 9.22 Control of nanoscopic photoelectron emission. (a) Scanning electron microscopy image of a single nanostructure; (b) An unshaped laser pulse is used as a reference; (c) Maximization of the integrated emission ratio A/B by optimal polarization shaping leads to emission patterns that exhibit a high contrast between region A versus region B. Hence, the photoelectron emission can be controlled experimentally on a nanoscopic length scale.

pattern, should provide an improved understanding of two-photon photoemission from nanostructured objects.

Acknowledgments

The authors enjoyed fruitful and stimulating discussions and support with and by many colleagues, notably E.V. Chulkov and P.M. Echenique (San Sebastian), J. P. Gauyacq (Paris), F.-J. Meyer zu Heringdorf (Duisburg), W. Pfeiffer (Bielefeld) and M. Bauer (Kiel). They also wish to thank their coworkers and graduate students. Financial support from the Deutsche Forschungsgemeinschaft via the priority programs SPP 1093 'Dynamics of electron transfer processes at interfaces', through projects Za 110/21 and AE 19/3, and SPP 1153 'Cluster in contact with surfaces' through project AE 19/12, is gratefully acknowledged.

References

1 Echenique, P.M. and Pendry, J.B. (1978) *The Journal of Physics C*, **11**, 2065.
2 Giesen, K., Hagen, F., Himpsel, F.J., Riess, H.J. and Steinmann, W. (1985) *Physical Review Letters*, **55**, 300.
3 Yen, R., Liu, J.M., Bloembergen, N., Yee, T.K., Fujimoto, J.G. and Salour, M.M. (1982) *Applied Physics Letters*, **40**, 185.
4 Williams, R.T., Boyt, R.R., Rife, J.C., Long, J.P. and Kabler, M.N. (1982) *Journal of Vacuum Science and Technology*, **21**, 509.
5 Bokor, J. (1989) *Science*, **246**, 1130.
6 Haight, R. (1995) *Surface Science Reports*, **21**, 275.
7 Schoenlein, R.W., Fujimoto, J.G., Eesley, G.L. and Capehart, T.W. (1988) *Physical Review Letters*, **61**, 2596.
8 Fann, W.S., Storz, R., Tom, H.W.K. and Bokor, J. (1992) *Physical Review Letters*, **68**, 2834.
9 Fann, W.S., Storz, R., Tom, H.W.K. and Bokor, J. (1992) *Physical Review B - Condensed Matter*, **46**, 13592.

10 Schmuttenmaer, C.A., Aeschlimann, M., Elsayed-Ali, H.E., Miller, R.J.D., Mantel, D.A., Cao, J. and Gao, Y. (1994) *Physical Review B - Condensed Matter*, **50**, 8957.

11 Schönhense, G., Elmers, H.J., Nepijko, S.A. and Schneider, C.M. (2006) *Advances in Imaging and Electron Physics*, **142**, 159.

12 Rotermund, H.H. (1997) *Surface Science Reports*, **29**, 267.

13 Schmidt, O., Bauer, M., Wiemann, C., Porath, R., Scharte, M., Andreyev, O., Schönhense, G. and Aeschlimann, M. (2002) *Applied Physics B: Lasers and Optics*, **74**, 223.

14 Clark, T.D., Tien, J., Duffy, D.C., Paul, K.E. and Whitesides, G.M. (2001) *Journal of the American Chemical Society*, **123**, 7677.

15 Hurst, S.J., Payne, E.K., Qin, L.D. and Mirkin, C.A. (2006) *Angewandte Chemie - International Edition*, **45**, 2672.

16 Popovic, Z., Otter, M., Calzaferri, G. and De Cola, L. (2007) *Angewandte Chemie - International Edition*, **46**, 6301.

17 Güdde, J., Rohleder, M., Meier, T., Koch, S.W. and Höfer, U. (2007) *Science*, **318**, 1287.

18 Chulkov, E.V., Borisov, A.G., Gauyacq, J.P., Sanchez-Portal, D., Silkin, V.M., Zhukov, V.P. and Echenique, P.M. (2006) *Chemical Reviews*, **106**, 4160.

19 Schöne, W.-D. (2007) *Progress in Surface Science*, **82**, 161.

20 Ueba, H. and Gumhalter, B. (2007) *Progress in Surface Science*, **82**, 193.

21 Ashcroft, N.W. and Mermin, N.D. (1976) *Solid State Physics*, Holt, Rinehart & Winston, New York.

22 (a) Landau, L.D. (1956) *Journal of Experimental and Theoretical Physics (USSR)*, **30**, 1058; (b) Landau, L.D. (1957) *Journal of Experimental and Theoretical Physics (USSR)*, **32**, 59; (c) Landau, L.D. (1958) *Journal of Experimental and Theoretical Physics (USSR)*, **35**, 97.

23 Pines, D. and Nozieres, P. (1989) *The Theory of Quantum Liquids*, Addison-Wesley, Reading.

24 Quinn, J.J. and Ferrell, R.A. (1958) *Physical Review*, **112**, 812.

25 Aeschlimann, M., Bauer, M. and Pawlik, S. (1996) *Chemical Physics*, **205**, 127.

26 Hertel, T., Knoesel, E., Wolf, M. and Ertl, G. (1996) *Physical Review Letters*, **76**, 535.

27 Ogawa, S. and Petek, H. (1996) *Surface Science*, **357**, 585.

28 Knoesel, E., Hotzel, A. and Wolf, M. (1998) *Physical Review B - Condensed Matter*, **57**, 12812.

29 (a)Petek, H., Nagano, H. and Ogawa, S. (1999) *Physical Review Letters*, **83**, 832; (b) Petek, H., Nagano, H. and Ogawa, S. (1999) *Applied Physics B: Lasers and Optics*, **68**, 369.

30 Aeschlimann, M., Bauer, M., Pawlik, S., Knorren, R., Bouzerar, G. and Bennemann, K.H. (2000) *Applied Physics A: Materials Science and Processing*, **71**, 485.

31 Zhukov, V.P., Andreyev, O., Hoffmann, D., Bauer, M., Aeschlimann, M., Chulkov, E.V. and Echenique, P.M. (2004) *Physical Review B*, **70**, 233106.

32 Lisowski, M., Loukakos, P.A., Bovensiepen, U., Stähler, J., Gahl, C. and Wolf, M. (2004) *Applied Physics A: Materials Science and Processing*, **78**, 165.

33 Mönnich, A., Lange, J., Bauer, M., Aeschlimann, M., Nechaev, I.A., Zhukov, V.P., Echenique, P.M. and Chulkov, E.V. (2006) *Physical Review B - Condensed Matter*, **74**, 035102.

34 Knorren, R., Bennemann, K.H., Burgermeister, R. and Aeschlimann, M. (2000) *Physical Review B - Condensed Matter*, **61**, 9427.

35 Nessler, W., Ogawa, S., Nagano, H., Petek, H., Shimoyama, J., Nakayama, Y. and Kishio, K. (1998) *Physical Review Letters*, **81**, 4480.

36 Marienfeld, A., Cinchetti, M., Bauer, M., Aeschlimann, M., Zhukov, V.P., Chulkov, E.V. and Echenique, P.M. (2007) *Journal of Physics - Condensed Matter*, **19**, 496213.

37 Echenique, P.M., Pitarke, J.M., Chulkov, E.V. and Rubio, A. (2000) *Chemical Physics*, **251**, 1.

38 Zhukov, V.P., Aryasetiawan, F., Chulkov, E.V., de Gurtubay, I.G. and Echenique,

P.M. (2001) *Physical Review B - Condensed Matter*, **64**, 195122.
39 Ladstädter, F., de Pablos, P.F., Hohenester, U., Puschnig, P., Ambrosch-Draxl, C., de Andres, P.L., García-Vidal, F.J. and Flores, F. (2003) *Physical Review B - Condensed Matter*, **68**, 085107.
40 Ladstädter, F., Hohenester, U., Puschnig, P. and Ambrosch-Draxl, C. (2004) *Physical Review B - Condensed Matter*, **70**, 235125.
41 Zhukov, V.P., Aryasetiawan, F., Chulkov, E.V. and Echenique, P.M. (2002) *Physical Review B - Condensed Matter*, **65**, 115116.
42 Bacelar, M.R., Schöne, W.-D., Keyling, R. and Ekardt, W. (2002) *Physical Review B - Condensed Matter*, **66**, 153101.
43 Papaconstantopoulos, D.A. (1986) *Handbook of the Band Structure of Elemental Solids*, Plenum, New York.
44 Springer, M., Aryasetiawan, F. and Karlsson, K. (1998) *Physical Review Letters*, **80**, 2389.
45 Karlsson, K. and Aryasetiawan, F. (2000) *Physical Review B - Condensed Matter*, **62**, 3006.
46 Nechaev, I.A. and Chulkov, E.V. (2005) *Physical Review B - Condensed Matter*, **71**, 115104.
47 Nechaev, I.A. and Chulkov, E.V. (2006) *Physical Review B - Condensed Matter*, **73**, 165112.
48 Zhukov, V.P., Chulkov, E.V. and Echenique, P.M. (2005) *Physical Review B - Condensed Matter*, **72**, 155109.
49 Mie, G. (1908) *Annalen der Physik*, **25**, 377.
50 Kreibig, U. and Vollmer, M. (1995) *Optical Properties of Metal Clusters, Springer Series in Materials Science, Vol. 25*, Springer, Berlin.
51 Bohren, C.F. and Huffmann, D.R. (1983) *Absorption and Scattering of Light by Small Particles*, Wiley, New York.
52 Kottmann, J.P. and Martin, O.J.F. (2001) *Optics Letters*, **26**, 1096.
53 Moskovits, M. (1985) *Reviews of Modern Physics*, **57**, 783.
54 Simon, H.J. and Chen, Z. (1989) *Physical Review B - Condensed Matter*, **39**, 3077.
55 Bouhelier, A., Beversluis, M., Hartschuh, A. and Novotny, L. (2003) *Physical Review Letters*, **90**, 013903.
56 Scharte, M., Porath, R., Ohms, T., Aeschlimann, M., Krenn, J.R., Dittelbacher, H., Aussenegg, F.R. and Liebsch, A. (2001) *Applied Physics B: Lasers and Optics*, **73**, 305.
57 Salerno, M., Krenn, J.R., Lamprecht, B., Schider, G., Ditlbacher, H., Félidj, N., Leitner, A. and Aussenegg, F.R. (2002) *Opto-Electronics Review*, **10**, 217.
58 Anisimov, S.I., Kapeliovich, B.L. and Perel'man, T.L. (1974) *Soviet Physics – JETP*, **39**, 375.
59 Budde, F., Heinz, T.F., Kalamarides, A., Loy, M.M.T. and Misewich, J.A. (1993) *Surface Science*, **283**, 143.
60 Brandbyge, M., Hedegard, P., Heinz, T.F., Misewich, J.A. and Newns, D.M. (1995) *Physical Review B - Condensed Matter*, **52**, 6042.
61 Fröhlich, H. (1950) *Physical Review*, **79**, 845.
62 (a) Eiguren, A., Hellsing, B., Chulkov, E.V. and Echenique, P.M. (2003) *Physical Review B - Condensed Matter*, **67**, 235432; (b) Eiguren, A., de Gironcoli, S., Chulkov, E.V., Echenique, P.M. and Tosatti, E. (2003) *Physical Review Letters*, **91**, 166803.
63 Grimvall, G. (1981) *The Electron–Phonon Interaction in Metals*, North-Holland, Amsterdam.
64 Sklyadneva, I.Yu., Chulkov, E.V., Schöne, W.-D., Silkin, V.M., Keyling, R. and Echenique, P.M. (2005) *Physical Review B - Condensed Matter*, **71**, 174302.
65 McDougall, B.A., Balasubramanian, T. and Jensen, E. (1995) *Physical Review B - Condensed Matter*, **51**, 13891.
66 Mathias, S., Wiesenmayer, M., Aeschlimann, M. and Bauer, M. (2006) *Physical Review Letters*, **97**, 236809.
67 Brorson, S.D., Kareroonian, A., Moodera, J.S., Face, D.W., Cheng, T.K., Ippen, E.P., Dresselhaus, M.S. and Dresselhaus, G. (1990) *Physical Review Letters*, **64**, 2172.

68 Fuchs, H., Hölscher, H. and Schirmeisen, A. (2005) *Encyclopedia of Materials: Science and Technology*, Elsevier, pp. 1–12.

69 Schirmeisen, A., Anczykowski, B. and Fuchs, H. (2007) *Handbook of Nanotechnology II* (ed. B. Bushan), Wiley, pp. 737–765.

70 Matzdorf, R. (1998) *Surface Science Reports*, **30**, 153.

71 Chang, Z., Rundquist, A., Wang, H., Murnane, M.M. and Kapteyn, H.C. (1997) *Physical Review Letters*, **79**, 2967.

72 Spielmann, Ch., Burnett, N.H., Sartania, S., Koppitsch, R., Schnürer, M., Kan, C., Lenzner, M., Wobrauschek, P. and Krausz, F. (1997) *Science*, **278**, 661.

73 Siffalovic, P., Drescher, M., Spieweck, M., Wiesenthal, T., Lim, Y.C., Weidner, R., Elizarov, A., Heinzmann, U. (2001) *Review of Scientific Instruments*, **72**, 30.

74 Tsilimis, G., Benesch, C., Kutzner, J. and Zacharias, H. (2003) *Journal of the Optical Society of America B - Optical Physics*, **20**, 246.

75 Goulielmakis, E., Uiberacker, M., Kienberger, R. *et al.* (2004) *Science*, **305**, 1267.

76 Sansone, G., Benedetti, E., Calegari, F. *et al.* (2006) *Science*, **314**, 443.

77 Paul, P.M., Toma, E.S., Breger, P., Mullot, G., Augé, F., Balcou, Ph., Muller, H.G. and Agostini, P. (2001) *Science*, **292**, 1689.

78 Mauritsson, J., Johnsson, P., Gustafsson, E., L'Huillier, A., Schafer, K.J. and Gaarde, M.B. (2006) *Physical Review Letters*, **97**, 013001.

79 Goulielmakis, E., Schultze, M., Hofstetter, M. *et al.* (2008) *Science*, **320**, 1640.

80 Ogawa, S., Nagano, H., Petek, H. and Heberly, A. (1997) *Physical Review Letters*, **78**, 1339.

81 Lange, J., Bayer, D., Rohmer, M., Wiemann, C., Gaier, O., Aeschlimann, M. and Bauer, M. (2006) *Proceedings of SPIE*, **6195**, 61950.

82 Bayer, D., Wiemann, C., Gaier, O., Bauer, M. and Aeschlimann, M. (2008) *Journal of Nanomaterials*, **00**, 249514.

83 Schlipper, R., Kusche, R., von Issendorff, B. and Haberland, H. (1998) *Physical Review Letters*, **80**, 1194.

84 Becker, Th., Hövel, H., Bettac, A., Reihl, B., Tschudy, M. and Williams, E.J. (1997) *Journal of Applied Physics*, **81**, 154.

85 Rohmer, M., Galeh, F., Aeschlimann, M., Bauer, M. and Hövel, H. (2007) *European Journal of Physics D*, **45**, 491.

86 Echenique, P.M., Berndt, R., Chulkov, E.V., Fauster, Th., Goldmann, A. and Höfer, U. (2004) *Surface Science Reports*, **52**, 219.

87 Weinelt, M. (2002) *Journal of Physics - Condensed Matter*, **14**, R1099.

88 Eiguren, A., Hellsing, B., Chulkov, E.V. and Echenique, P.M. (2003) *Journal of Electron Spectroscopy and Related Phenomena*, **129**, 111.

89 Knoesel, E., Hotzel, A. and Wolf, M. (1998) *Journal of Electron Spectroscopy and Related Phenomena*, **88–91**, 577.

90 Berthold, W., Höfer, U., Feulner, P., Chulkov, E.V., Silkin, V.M. and Echenique, P.M. (2002) *Physical Review Letters*, **88**, 056805.

91 Roth, M., Weinelt, M., Fauster, T., Wahl, P., Schneider, M.A., Diekhöner, L. and Kern, K. (2004) *Applied Physics A: Materials Science and Processing*, **78**, 155.

92 Boger, K., Weinelt, M., Wang, J. and Fauster, T. (2004) *Applied Physics A: Materials Science and Processing*, **78**, 161.

93 Hirschmann, M. and Fauster, T. (2007) *Applied Physics A: Materials Science and Processing*, **88**, 547.

94 Kliewer, J., Berndt, R., Chulkov, E.V., Silkin, V.M., Echenique, P.M. and Crampin, S. (2000) *Science*, **288**, 1399.

95 LaShell, S., McDougall, B.A. and Jensen, E. (1996) *Physical Review Letters*, **77**, 3419.

96 Theilmann, F., Matzdorf, R., Meister, G. and Goldmann, A. (1997) *Physical Review B - Condensed Matter*, **56**, 3632.

97 Lisowski, M., Loukakos, P.A., Bovensiepen, U. and Wolf, M. (2004) *Applied Physics A: Materials Science and Processing*, **79**, 739.

98 Reinert, F., Nicolay, G., Schmidt, S., Ehm, D. and Hüfner, S. (2001) *Physical Review B - Condensed Matter*, **63**, 115415.

99 Eiguren, A., Hellsing, B., Reinert, F., Nicolay, G., Chulkov, E.V., Silkin, V.M., Hüfner, S. and Echenique, P.M. (2002) *Physical Review Letters*, **88**, 066805.

100 Silkin, V.M., Balasubramanian, T., Chulkov, E.V., Rubio, A. and Echenique, P.M. (2001) *Physical Review B - Condensed Matter*, **64**, 085334.

101 LaShell, S., McDougall, B.A. and Jensen, E. (2006) *Physical Review B - Condensed Matter*, **74**, 033410.

102 Fuglsang Jensen, M., Kim, T.K., Bengio, S., Sklyadneva, I.Yu., Leonardo, A., Eremeev, S.V., Chulkov, E.V. and Hofmann, Ph. (2007) *Physical Review B - Condensed Matter*, **75**, 153404.

103 Kutzner, J., Paucksch, R., Jabs, C., Zacharias, H. and Braun, J. (1997) *Physical Review B - Condensed Matter*, **56**, 16003.

104 Wiemann, C., Bayer, D., Rohmer, M., Aeschlimann, M. and Bauer, M. (2007) *Surface Science*, **601**, 4714.

105 Kubo, A., Onda, K., Petek, H., Sun, Z., Jung, Y.S. and Kim, H.K. (2005) *Nano Letters*, **5**, 1123.

106 Kubo, A., Pontius, N. and Petek, H. (2007) *Nano Letters*, **7**, 470.

107 Chelaru, L.I., Horn von Hoegen, M., Thien, D. and Meyer zu Heringdorf, F.-J. (2006) *Physical Review B - Condensed Matter*, **73**, 115416.

108 Meyer zu Heringdorf, F.-J., Buckanie, N.M., Chelaru, L.I. and Raß, N. (2008) in *EMC 2008*, vol. 1 (eds. M. Luysberg, K. Tillmann and T. Weirich), Springer, Berlin, pp. 737.

109 Chelaru, L.I. and Meyer zu Heringdorf, F.-J. (2007) *Surface Science*, **601**, 4541.

110 Bauer, M., Pawlik, S. and Aeschlimann, M. (1997) *Physical Review B - Condensed Matter*, **55**, 10040.

111 Ogawa, S., Nagano, H. and Petek, H. (1999) *Physical Review Letters*, **82**, 1931.

112 Bauer, M., Pawlik, S. and Aeschlimann, M. (1999) *Physical Review B - Condensed Matter*, **60**, 5016.

113 Petek, H., Nagano, H., Weida, M.J. and Ogawa, S. (2001) *Journal of Physical Chemistry B*, **105**, 6767.

114 Borisov, A.G., Kazansky, A.K. and Gauyacq, J.P. (2001) *Physical Review B - Condensed Matter*, **64**, 201105.

115 Borisov, A.G., Gauyacq, J.P., Kazansky, A.K., Chulkov, E.V., Silkin, V.M. and Echenique, P.M. (2001) *Physical Review Letters*, **86**, 488.

116 Gauyacq, J.P. and Kazansky, A. (2005) *Physical Review B - Condensed Matter*, **72**, 045418.

117 Gauyacq, J.P., Borisov, A.G. and Bauer, M. (2007) *Progress in Surface Science*, **82**, 244.

118 Frischkorn, C. and Wolf, M. (2006) *Chemical Reviews*, **106**, 4207.

119 Petek, H., Weida, M.J., Nagano, H. and Ogawa, S. (2000) *Science*, **288**, 1402.

120 Kröner, D., Klamroth, T., Nest, M. and Saalfrank, P. (2007) *Applied Physics A: Materials Science and Processing*, **88**, 535.

121 Tans, S.J., Verschueren, A.R.M. and Dekker, C. (1998) *Nature*, **393**, 49.

122 Harris, C.B., Ge, N.-H., Lingle, G.R., Jr, McNeill, J.D. and Wong, C.M. (1997) *Annual Review of Physical Chemistry*, **48**, 711.

123 Szymanski, P., Garrett-Roe, S. and Harris, C.B. (2005) *Progress in Surface Science*, **78**, 1.

124 (a) Gundlach, L., Ernstorfer, R. and Willig, F. (2007) *Applied Physics A: Materials Science and Processing*, **88**, 481; (b) Gundlach, L., Ernstorfer, R. and Willig, F. (2007) *Progress in Surface Science*, **82**, 355.

125 Duncan, W.R., Stier, W.M. and Prezhdo, O.V. (2005) *Journal of the American Chemical Society*, **127**, 7941.

126 see e.g. Tzeng, C.-T., Lo, W.-S., Yuh, J.-Y., Chu, R.-Y. and Tsuei, K.-D. (2000) *Physical Review B - Condensed Matter*, **61**, 2263, and references therein.

127 Maggini, M., Scorrano, G. and Prato, M. (1993) *Journal of the American Chemical Society*, **115**, 9798.

128 Maggini, M. and Prato, M. (1998) *Accounts of Chemical Research*, **31**, 519.

129 Jacquemin, R., Kraus, S. and Eberhardt, W. (1995) *Solid State Communications*, **105**, 449.

130 Long, J.P., Chase, S.J. and Kabler, M.N. (2001) *Physical Review B - Condensed Matter*, **64**, 205415.

131 See e.g. Wasielewski, M.R., O'Neil, M.P., Lykke, K.R., Pellin, M.J. and Gruen, D.M. (1991) *Journal of the American Chemical Society*, **113**, 2774.

132 Hoger, T., Grimmer, D. and Zacharias, H. (2007) *Applied Physics A: Materials Science and Processing*, **88**, 449.

133 O'Regan, B. and Grätzel, M. (1991) *Nature*, **353**, 737.

134 Stockman, M.I., Faleev, S.V. and Bergman, D.J. (2002) *Physical Review Letters*, **88**, 067402.

135 Brixner, T., García de Abajo, F.J., Schneider, J. and Pfeiffer, W. (2005) *Physical Review Letters*, **95**, 093901.

136 Brixner, T., Garcia Abajo, S.J., Schneider, J., Spindler, C. and Pfeiffer, W. (2006) *Physical Review B - Condensed Matter*, **73**, 125437.

137 Aeschlimann, M., Bauer, M., Bayer, D., Brixner,T., Garcia de Abajo, F.J., Pfeiffer, W., Rohmer, M., Spindler, C. and Steeb, F. (2007) *Nature*, **446**, 301.

10
Nanoplasmonics

Gerald Steiner

10.1
Introduction

The basic foundation of nanoplasmonic involves interactions between light and metal particles. Such particles, which are often referred to as *clusters*, have been the subject of a large number of investigations. Gustav Mie's theory about the study of optical properties of small gold particles, published in 1908, was the starting point for the new field of plasmonics. During the past half-century the nature of interactions between photons and electrons in metal clusters has been the focus of many research fields, including not only physics but also chemistry, materials science, medicine, biology and the environmental sciences. Clearly, metal cluster photonics is important in many fields, even in nanotechnology. Although the term 'cluster' is often used for a number of unspecified particles, there is no precise definition of a cluster. Hence, in this chapter clusters are defined as particles composed of a certain number of atoms that form a spherical or elliptical particle with the dimension in the range between 5 and 100 nm, corresponding to $500 \ldots 10^7$ atoms. The range from single atoms to bulk material is illustrated in Figure 10.1. It should be noted, that some publications use the term cluster just for very small clusters consisting of less than 100 atoms. The defined cluster region covers a rather wide range in relation to their optical properties. Metal clusters are subjected to variations not only in size but also in shape (some frequently identified forms are shown in Figure 10.2), but they are rarely spherical. In particular, clusters adsorbed onto a surface are ellipsoid in shape. Nanorods are tiny rod-shaped particles, less than 100 nm in diameter, but often with a length in excess of a few micrometers. The advantage of nanorods lies in their high aspect ratio (the ratio of length to diameter), as this permits them to show certain properties not seen with spherical or elliptical clusters. The optical properties of metal clusters depend heavily on the shape, the environment of the cluster, and the type of metal. Nonetheless, the two general features of all clusters are their ability to interact with light and to produce a localized electromagnetic field, the details of which are discussed in the following sections [1–3].

Nanotechnology. Volume 6: Nanoprobes. Edited by Harald Fuchs

ISBN: 978-3-527-31733-2

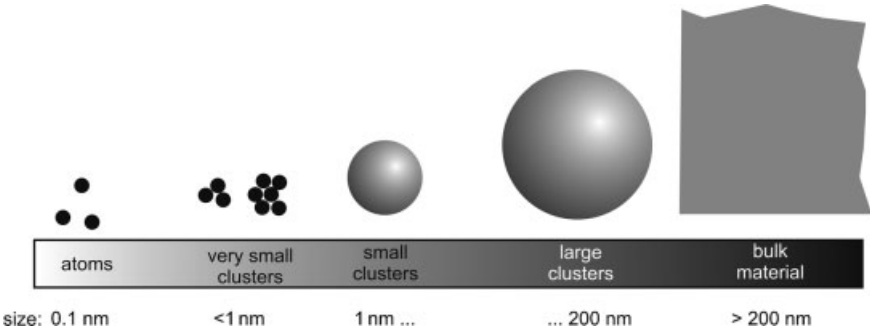


Figure 10.1 Dimensions of metal clusters on the scale from single atoms to bulk material.

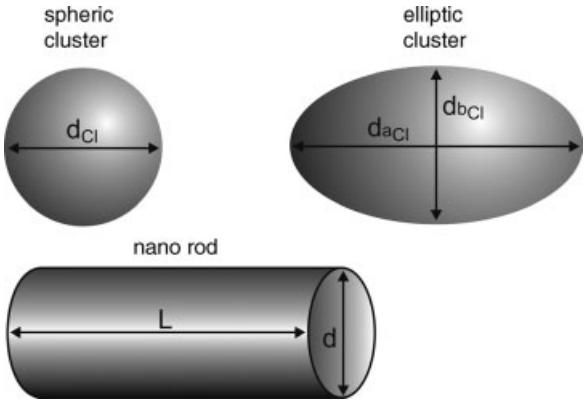


Figure 10.2 Different forms of metal clusters.

10.2 Single Clusters

In general, the description of the interaction between light and metal clusters is obtained by the solution of Maxwell's equation and imposing the boundary conditions. A simple approach to understand the optical response of metal clusters is to consider the free electrons. The positive charges are assumed to be immobile; then, if the cluster is illuminated by light the electric vector of the light wave will displace the free electrons. Under consideration of the boundary condition the cluster exhibits polarization; this effect is illustrated in Figure 10.3.

The internal field E_i of a spherical cluster is given by

$$E_i = E_0 \frac{3\varepsilon_1}{(\varepsilon r_{Cl} + \varepsilon i_{Cl}) + 2\varepsilon_1} \tag{10.1}$$

where E_0 is the electric field of the incident light. The dielectric function of the cluster material is given by the real part εr_{Cl} and the imaginary part $\varepsilon i_{Cl.}$ The surrounding

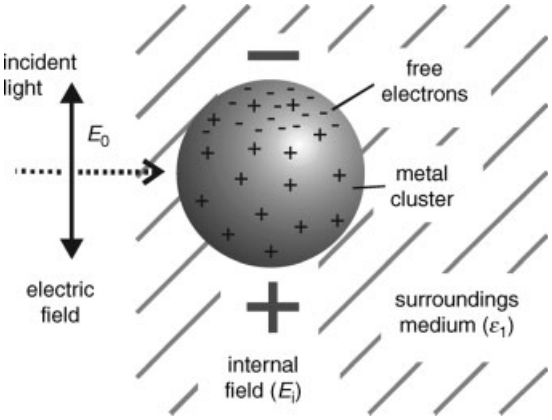


Figure 10.3 Polarization of a metal cluster caused by the electric vector of the incident light wave.

medium is characterized by the dielectric function ε_1. An important parameter is the static polarizability α of the cluster

$$\alpha = 4\pi\varepsilon_0 d^3 \frac{(\varepsilon r_{Cl} + \varepsilon i_{Cl}) - \varepsilon_1}{(\varepsilon r_{Cl} + \varepsilon i_{Cl}) + 2\varepsilon_1} \tag{10.2}$$

with the cluster diameter d. Unfortunately, this solution describes only static conditions. If the cluster is placed in an electromagnetic field, the free electrons will move inside the metal cluster with frequency of the external field. The excitation of the free electrons leads to an internal field. In this case, the dielectric constants must be replaced by their frequency-dependent values. As the magnetic fields do not occur, the complex dielectric function (ε_{Cl}) must be replaced with the frequency-dependent values:

$$\bar{\varepsilon}_{Cl} = \bar{\varepsilon}_{Cl}(\omega) \tag{10.3}$$

where ω is the angular frequency of the incident light wave.

The internal electric field shows a resonance when

$$|\varepsilon r_{Cl}(\omega) + 2\varepsilon_1|^2 + |\varepsilon i_{Cl}(\omega)|^2 = \min \tag{10.4}$$

This means that, in a spherical cluster, the resonance frequency is found by the relationship

$$\varepsilon r_{Cl}(\omega) = -2\varepsilon_1 \tag{10.5}$$

The real part of the dielectric function can be also expressed by the relationship

$$\varepsilon r_{Cl} \approx 1 - \frac{\omega_P^2}{\omega^2} \tag{10.6}$$

where is ω_P the Drude plasma frequency. Consequently, the resonance frequency ω_R of a spherical cluster is given by

$$\omega_R = \frac{\omega_P}{\sqrt{2\varepsilon_1 + 1}} \tag{10.7}$$

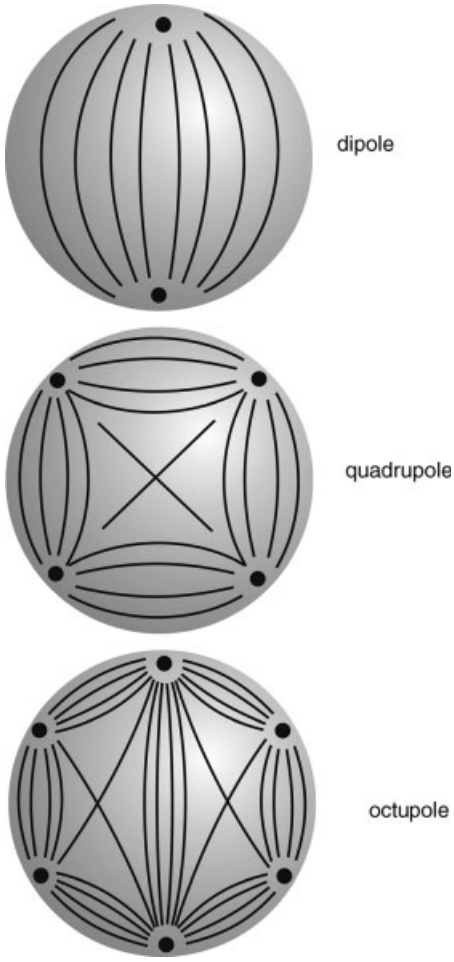


Figure 10.4 Excitation of multipoles within a metal cluster.

A metallic cluster exhibits not only a distinct dipole but also a quadrupole and higher multipole, as depicted in Figure 10.4. The excitation of these dipoles is dependent on the size, the shape and the type of metal, as well as on the wavelength of the incident light. The internal field at the resonance frequency, where the free electrons exhibit a strong collective oscillation, is known as the *surface plasmon*. This term is normally used to describe the excitations at a metal surface, whereas the plasmons in a metal cluster are localized and are thus referred to *localized surface plasmons*. One consequence of the internal field is that there is a strong electromagnetic field in the proximity of the cluster, as shown in Figure 10.5. It should be noted that the term 'localized surface plasmons' also describes the resulting external field of the cluster. As mentioned above, the spectral position of the localized surface

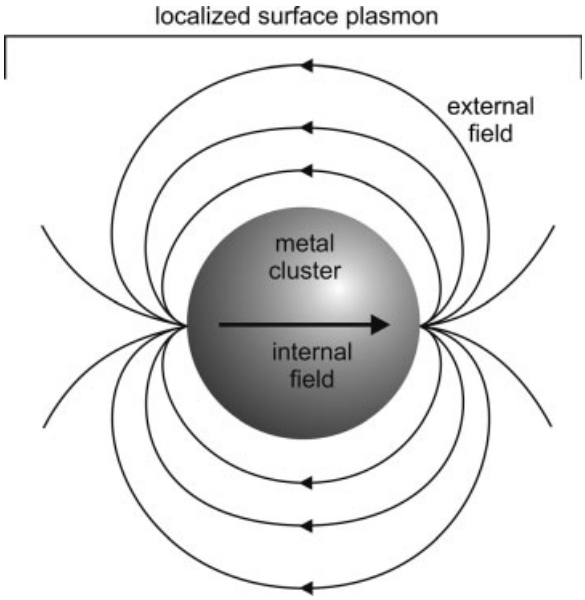


Figure 10.5 The excitation of an internal field causes a strong electromagnetic field around the cluster.

plasmons is dependent on the size, shape and metal type of the cluster, and the dielectric function of the surrounding medium. As an example, Figure 10.6 illustrates the effect of cluster size on the spectral position of the resonance for a gold cluster placed on a glass slide. For larger clusters, higher-order multipoles become important, and this results in a more pronounced shift of the localized surface plasmon resonance (SPR) as the particle size increases. In addition, the position and shape of the resonance are also dependent on dielectric function of the surrounding medium [3–6].

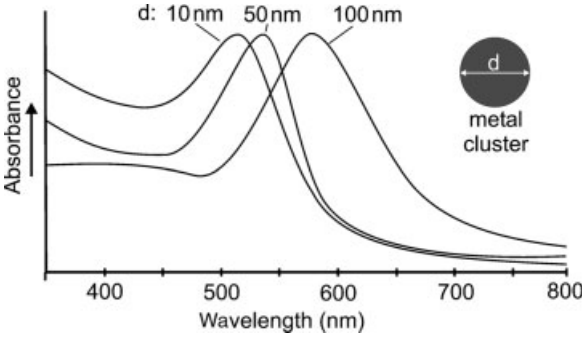


Figure 10.6 Calculated absorbance spectra of a single spherical gold cluster with different sizes.

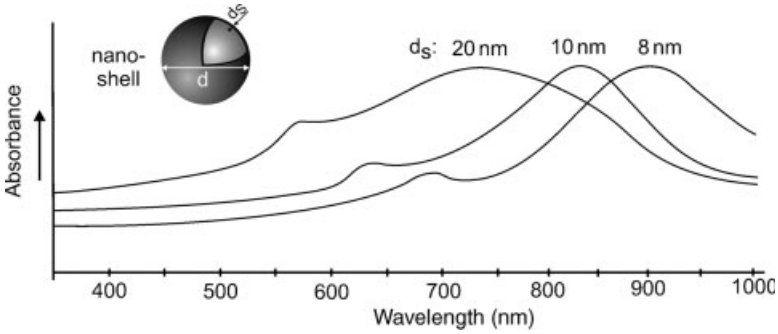


Figure 10.7 Absorbance spectra of nanoshells consisting of a glass core and gold shell.

10.3 Nanoshells

A nanoshell contains different materials in the core and shell. Often, the core material is metallic while the shell forms a thin dielectric layer, and in this case the localized SPR is affected by the shell. When the dielectric constant or the thickness of the shell is changed, however, the resonance conditions are also changed. For example, when the shell material is metallic and the core is a dielectric material, the change in SPR can be dramatic, with the spectral position of the resonance being shifted to longer wavelengths than those in the corresponding solid metal cluster. The same effect occurs when the metal shell thickness is decreased. The calculated absorbance spectra for nanoshells consisting of a glass core and a thin gold shell are shown in Figure 10.7. These nanoshells serve as the main component of many applications in biosensing and semiconducting [2, 7].

10.4 Layer of Clusters

When clusters become closer to each other, the electromagnetic field causes an electromagnetic coupling between the individual units [8]. One simple way to describe this coupling is with a system of two interacting dipolar oscillators, when two principal situations are possible: (i) the orientation of the cluster dipole is in the *same direction*; or (ii) the orientation is in the *opposite direction*. These two configurations are illustrated in Figure 10.8. In the case of perpendicular orientation (Figure 10.8a), the internal dipole vectors have a different polarization, whereas for parallel orientation (Figure 10.8b) the internal dipole vectors exhibit the same polarization. If both dipoles are excited together with the same polarization, then the resulting field is enhanced. In the case of perpendicular orientation when both clusters are excited together, but with different polarization, the resulting field will be weaker – a fact which has a major influence on the spectral position of the SPR. Although the interaction between clusters is also determined by the cluster size and

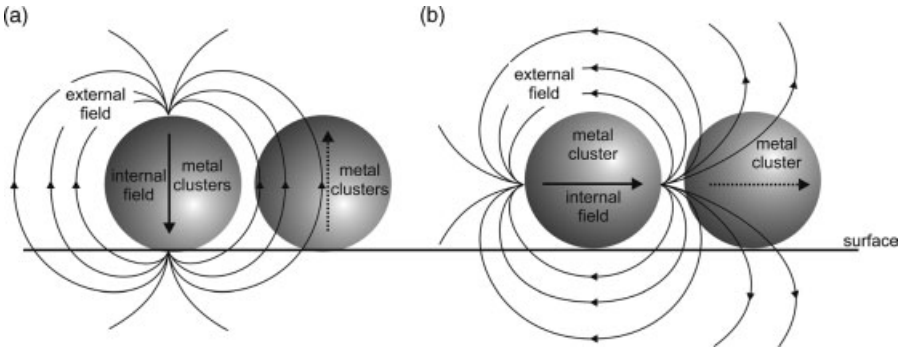


Figure 10.8 Electromagnetic coupling between two clusters.
(a) Different polarization; (b) Similar polarization.

distance, as well as by the optical properties of the medium between the clusters, the most critical parameter is polarization of the localized surface plasmons. To illustrate this fact, Figure 10.9 shows the absorbance spectra of two identical gold clusters with the same and different polarizations of the localized surface plasmons. When the

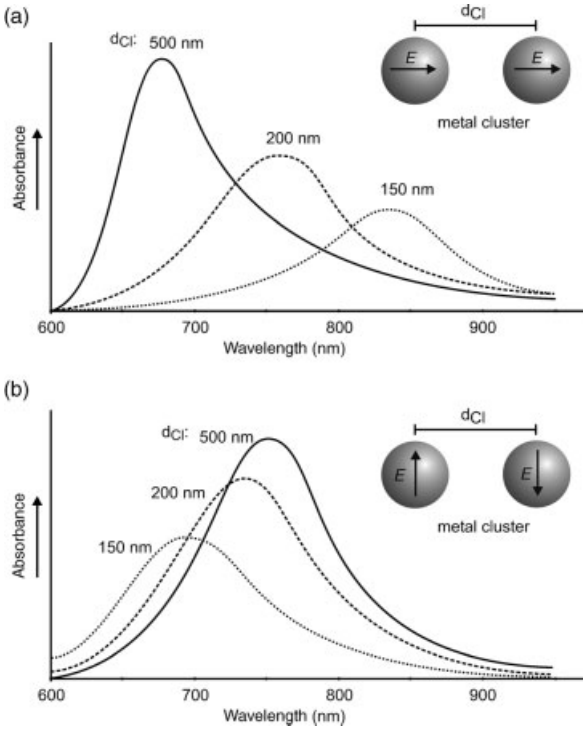


Figure 10.9 Calculated absorbance spectra for two gold clusters at a certain distance. (a) With the same polarization; (b) With different polarization.

internal field of the clusters exhibits the same polarization, the position of the SPR shifts towards longer wavelengths as the inter-cluster distance is increased. In the case of a different polarization a weak blue-shift occurs when the inter-cluster distance is increased. This different optical response – which is also evident for more than two clusters – has major consequences on the sensitivity when a cluster film is used to enhance a weak optical signal. As the red-shift is much stronger than the blue-shift, the cluster that exhibits localized surface plasmons with the same polarization provides a higher sensitivity. A similar behavior can be observed when a cluster film is excited by light with a different polarization of the wave. The substantial difference in electromagnetic fields around clusters caused by an excitation with parallel and perpendicular polarized light is illustrated in Figure 10.10. The direction of the electric vector (E) of the exciting field is the critical parameter. The induced dipoles in the clusters have the same direction as the exciting field. In the case of a perpendicular orientation, E is oriented perpendicular to the surface. The induced dipoles in the clusters must then point in the same direction, and the fields around the clusters develop accordingly. Due to the parallel direction of the dipoles, almost no electric coupling can occur between clusters. However, the

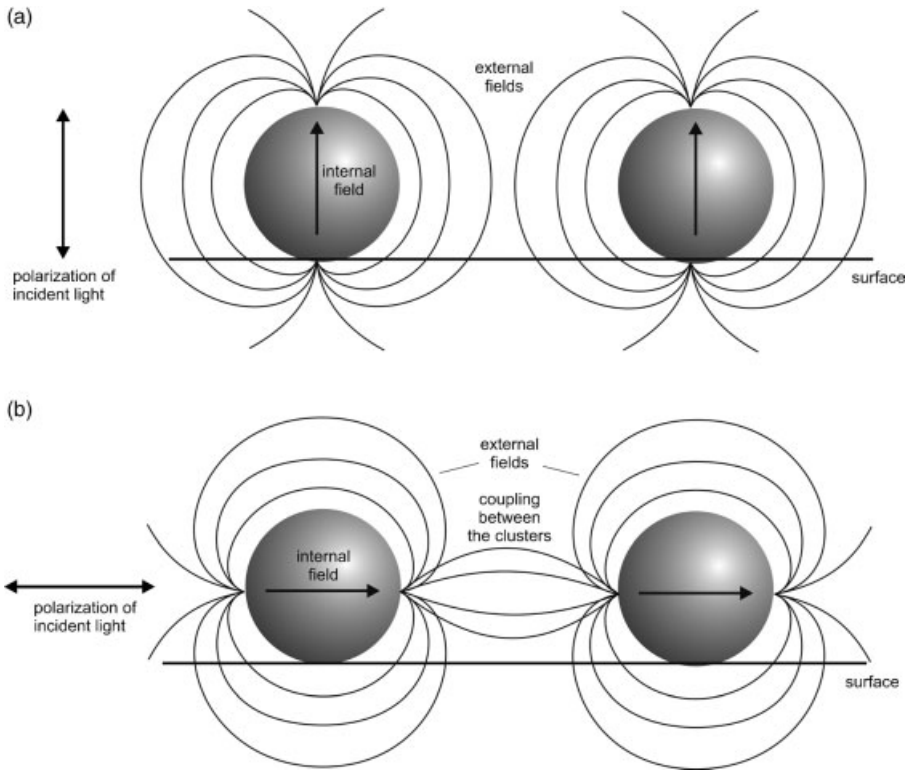


Figure 10.10 Excitation of localized surface plasmons in case of (a) perpendicular orientation and (b) parallel polarization of the incident light wave.

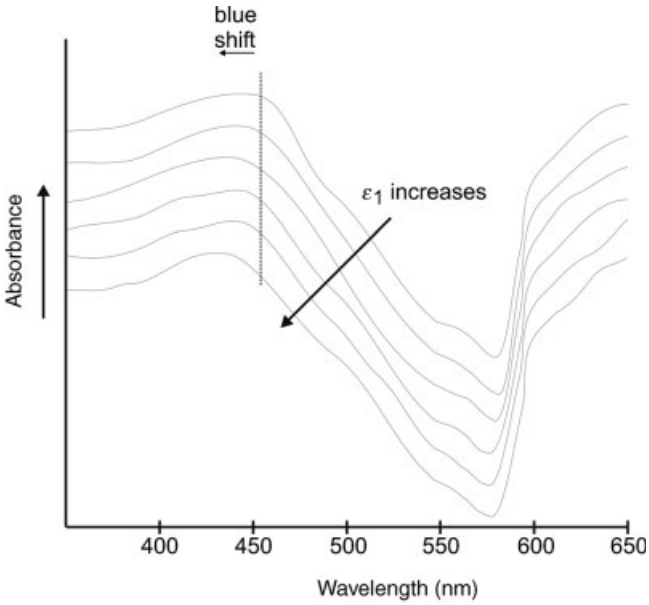


Figure 10.11 Absorbance spectra of a cluster layer for different refractive indices of the surrounding medium (ε_1), measured with unpolarized light.

situation changes dramatically when the orientation of E is parallel to the surface. Now, because of the induced dipoles, the field of the clusters will stretch both towards the substrates and towards the surrounding medium. The direction of the internal dipoles permits a powerful coupling between neighboring clusters, and only in this case will the position of the SPR shift when the refractive index of medium between the clusters changes. The absorbance spectra of cluster layers with different refractive indices of the surrounding medium are shown in Figure 10.11. Since coupling between the clusters occurs only at a parallel polarization of the electric field vector, a spectral shift of the SPR is observed only for parallel polarized light. Upon increasing the refractive index of the medium, however, the field coupling between metal clusters becomes stronger and the SPR shifts towards shorter wavelengths.

Although the optical analysis of cluster arrangements plays an important role, it is impossible to describe the optical properties by extending Equations (10.1) to 10.7 to an arrangement of thousands of clusters. In such a case, an effective medium theory must be applied. Such theories describe the connection and interaction *between* clusters, as well as with their surrounding medium, and provide the 'effective' optical behavior of a cluster arrangement (Figure 10.12).

Several different effective medium theories have been devised, each of which is more or less accurate depending on the cluster material, the size form and the distance between clusters. However, they all assume that the macroscopic material is homogeneous, and generally fail to predict the properties of a composite material.

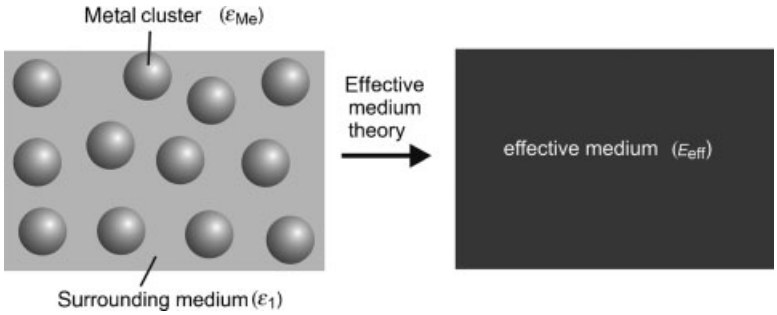


Figure 10.12 The transition from a macroscopically inhomogeneous medium to a homogeneous, optically effective medium.

Details of the three most frequently used effective medium models are summarized in Table 10.1 [9, 10].

The first application of an effective medium theory was made by Maxwell-Garnett to explain the color of a discontinuous metal film. The model was based on clusters which were assumed to be spheres in a host medium; this resulted in an effective dielectric function of the composite material. The *Maxwell-Garnett model* is valid at low volume fractions as it is assumed that the metal clusters are spatially separated. In contrast, the *Bruggeman model* does not distinguish between embedded metal clusters and matrix; rather, the two materials appear in a completely symmetric manner, and consequently the Bruggeman model can easily be extended to more than two components. Free structures of the embedded metal clusters and the matrix and a static treatment of the geometry are also possible using the Bergmann model. The spectral density $g(n,f)$ function (see Table 10.1) carries all of the geometric information and depends only on topology. As a result, the Bergmann model can be used to distinguish between geometric quantities and dielectric properties [11–13].

10.5 Surface-Enhanced Spectroscopy

10.5.1 Surface-Enhanced Raman Scattering

The effect of surface-enhanced Raman scattering (SERS) on small metal clusters has been recognized for more than 30 years. The Raman scattering of a molecule located in close proximity to the surface of a gold or silver cluster is 'enhanced' up to one million-fold. Such an enhancement effect is based on two principal mechanisms:

- The strong electric field of the localized surface plasmons interacts with the electron orbitals of the molecule, which leads to an enhancement of the Raman cross-section by up to four orders of magnitude.

Table 10.1 Most common effective medium model.

Model	Formula	Properties
Maxwell-Garnett	$\frac{\varepsilon_{\text{eff}}-\varepsilon_E}{\varepsilon_{\text{eff}}+2\varepsilon_E}=f_{\text{Me}}\frac{\varepsilon_{\text{Me}}-\varepsilon_E}{\varepsilon_{\text{Me}}+2\varepsilon_E}$ (f_{Me}: volume fraction of the metal clusters)	Suitable for uniform and spherical clusters with a relatively low volume fraction; computations are very efficient; describes poorly composite materials with high volume fraction of the metal clusters.
Bruggeman	$f_{\text{Me}}\frac{\varepsilon_{\text{Me}}-\varepsilon_{\text{eff}}}{\varepsilon_{\text{Me}}+2\varepsilon_{\text{eff}}}+(1-f_{\text{Me}})\frac{\varepsilon_E-\varepsilon_{\text{eff}}}{\varepsilon_E+2\varepsilon_{\text{eff}}}=0$	Self-consistent; provides a quite good correlation between theory and experiment also for higher volume fractions; the result can be a dielectric function of a higher polynomial.
Bergmann	$\varepsilon_{\text{eff}}=\varepsilon_E\left(1-f_{\text{Me}}\int_0^1\left(\frac{g(n,f_{\text{Me}})(\varepsilon_E-\varepsilon_{\text{Me}})}{\varepsilon_E-n(\varepsilon_E-\varepsilon_{\text{Me}})}\right)\mathrm{d}n\right)$ the zeroth and first moment of $g(n,f)$ are $\int_0^1 g(n,f_{\text{Me}})\mathrm{d}n=1\int_0^1 ng(n,f_{\text{Me}})\mathrm{d}n=\frac{1-f_{\text{Me}}}{3}$	Best approximation of composite materials; percolations of the cluster are considered; computations are more complex.

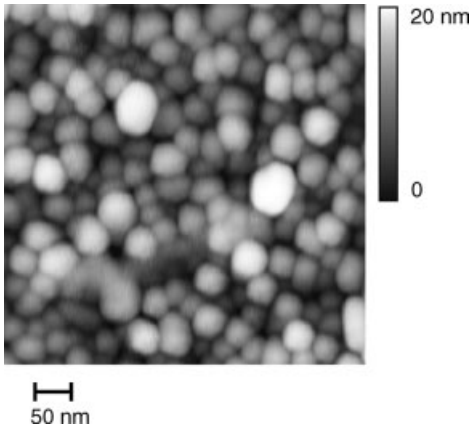


Figure 10.13 Atomic force microscopy image of a silver cluster layer, prepared by evaporation onto a smooth silicon wafer.

- The second effect is related to charge transfer processes; such a chemical interaction may cause an additional enhancement by typically two orders of magnitude.

SERS has been observed for many molecules when adsorbed to gold or silver clusters, or even to nanoscopic rough metal surfaces [14]. The intensity of the Raman scattering is increased as the wavelength is increased. For example, the strongest enhancement for gold clusters may be obtained with a laser excitation wavelength beyond 1000 nm. The SERS spectra obtained are almost completely depolarized. Figure 10.13 shows an atomic force microscopy image of a silver cluster layer used as a substrate for SERS. An example of the SERS effect is represented in Figure 10.14,

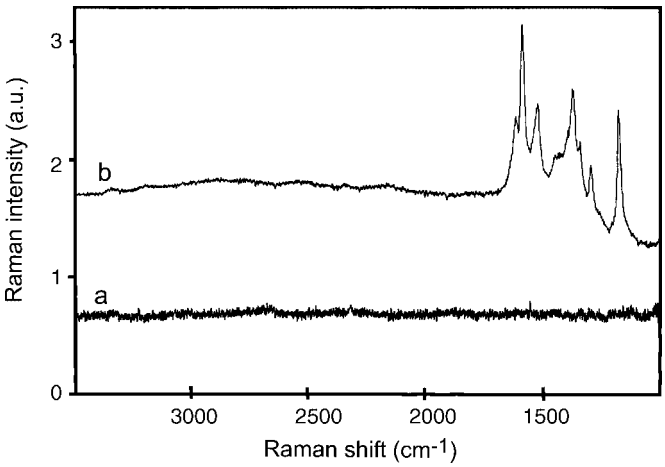


Figure 10.14 Raman spectrum (a) and surface-enhanced Raman spectrum (b) of a very thin film basic fucsine. The Raman spectrum is enhanced 25-fold.

which shows the SER and Raman spectra of a very thin-film basic fucsine on the SERS substrate and a pure silicon wafer.

Clearly, the cluster layer yields an enhancement which is manifested on the most prominent Raman bands. In contrast to the conventional Raman spectrum (Figure 10.14, spectrum a), where no bands can be observed, prominent Raman modes – such as the ring vibration at 1614 cm^{-1}, a deformation vibration of the NH_2 groups at 1586 cm^{-1}, and the valence vibration of the C–NH_2 groups at 1340 cm^{-1}, appear in the SERS spectrum.

Unfortunately, in recent years the SERS effect has not been used as a routine method in Raman spectroscopy. Although the strong enhancement provides many advantages, such as high sensitivity and the opportunity to characterize thin layers of molecules, SERS does not permit quantitative measurements to be made, despite these very often being required in a routine analysis. The reason for this is that the cluster layers themselves are poorly reproducible. Nonetheless, recent developments have been devoted to overcoming this limitation by using sol–gel clusters embedded in porous glass, or by using photonic crystals with a uniform metal-covered nanostructure.

10.5.2
Surface-Enhanced Fluorescence

Surface-enhanced fluorescence (SEF) is comparable to the SERS effect. The enhancement of the fluorescence signal is also caused by strong interactions between the localized surface plasmons of a metal cluster and the electrons of molecule next to the cluster surface [15]. However, at least two important factors must be considered here:

- First, it is well known that a direct contact between a molecule and a metal leads to a quenching of the fluorescence – this is also known as the 'first layer effect'. Therefore, the optimum enhancement of fluorescence does not derive from molecules that are adsorbed to the metal cluster surface; rather, maximum fluorescence is obtained at a certain distance of few nanometers between the molecule and the cluster surface.
- The second factor includes the size and form of the metal clusters. When the clusters become large, the damping of the fluorescence will be increases, whereas small clusters may not be stable and exhibit only a weak electric field.

The enhancement factors for SEF are comparable to those for SERS. The spectra in Figure 10.14, which were measured at 544 nm excitation, also exhibit a fluorescence which is seen as a very broad signal across the whole spectral range. Although the fluorescence of the first (or more) layer of adsorbed molecules is quenched, the enhancement factor is approximately 50. Experiments with silver clusters covered with a dielectric film of SiO_2 a few nanometers thick, produced a much greater fluorescence than did pure silver clusters. As with SERS, SEF has not become a routine method, due not only to the poor reproducibility of the metal cluster but also to the 'first layer effect'.

10.5.3
Surface-Enhanced Infrared Absorption Spectroscopy

Surface-enhanced infrared absorption (SEIRA) spectroscopy also originates from the enhanced electromagnetic field around a metal cluster. The enhancement of infrared absorption is generally on the order of one to two magnitudes [16, 17]. As such enhancement is remarkable for adsorbed molecules on a metal cluster surface, an increase in absorption coefficients and a selection rule for the absorption bands provide additional enhancement and information regarding the orientation of the adsorbed molecule [18, 19]. Figure 10.15 shows a Fourier-transformed infrared (FTIR) spectrum and a SEIRA spectrum of basic fucsine. Both spectra were measured using an attenuated total reflection (ATR) method, with the same measurement parameters. Enhancement of the absorption bands in spectrum b in Figure 10.15 can be clearly seen. Yet, in comparison to the SERS spectrum in Figure 10.14, the strongest band (at 1586 cm^{-1}) exhibited only a threefold enhancement.

Due to the surface selection rule, some absorption bands do not appear, or appear only weakly [21]. Shifts of the absorption bands may seem due to the influence of chemisorption of the complex. Although the optimum thickness for SEIRA-active metal cluster films shows some variation from study to study, it is consistently in the range of 5 to 12 nm. The preparation of the SEIRA active surface is straightforward, usually by the evaporation of an ATR crystal surface with gold or silver. A fresh gold surface shows a strong affinity towards many organic substances. In particular, sulfur complexes are chemisorbed immediately and very strongly onto gold cluster surfaces, and this may lead to an unwanted SEIRA spectrum.

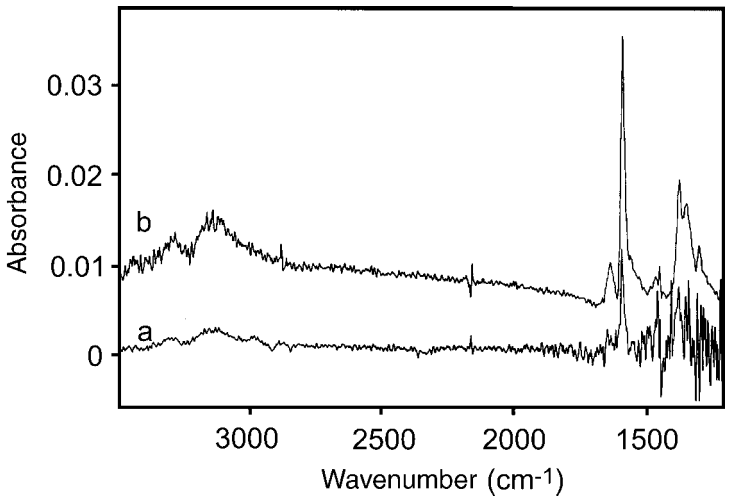


Figure 10.15 FTIR spectrum (a) and SEIRA spectrum (b) of a very thin layer of basic fucsine.

10.6 Biosensing

Metal clusters of gold or silver are stable, inexpensive and simple to prepare, easy to modify, and have optical properties that make them attractive for biosensing and biochips. Today, metal clusters with a functionalized surface are used to detect biomolecules, with an extremely high sensitivity [22, 23]. As mentioned above, the SPR of a metal cluster is also affected by other metal clusters that are in its immediate vicinity. When two or more clusters are brought into proximity, the electric fields will couple, and this will result in a shift of the resonance wavelength. This effect can be easily used in biosensing as an intrinsic enhancement of the detection signal. An example of this is the study of the dynamics of DNA hybridization at the single-molecule level [24]. The principal mechanism of detection is illustrated schematically in Figure 10.16. Here, surface-functionalized metal clusters are used as an anchor for single-stranded DNA (ssDNA). The ssDNA molecules have a biotin molecule attached at one end, and this allows them to bind to the streptavidin-coated 'anchor cluster'. When the surface-bound ssDNA are introduced into a solution with biotin-functionalized clusters, a spectral shift of the SPR occurs, such that gold clusters with a diameter of approximately 40 nm turn from green to orange, while the color of silver metal clusters changes from blue to yellow-green. The limit of detection is in the zeptomolar range. The binding of proteins, interactions between antibodies and antigens and receptor–ligand interactions, can also be detected in this way. The coupling of localized surface plasmons between individual clusters also represents

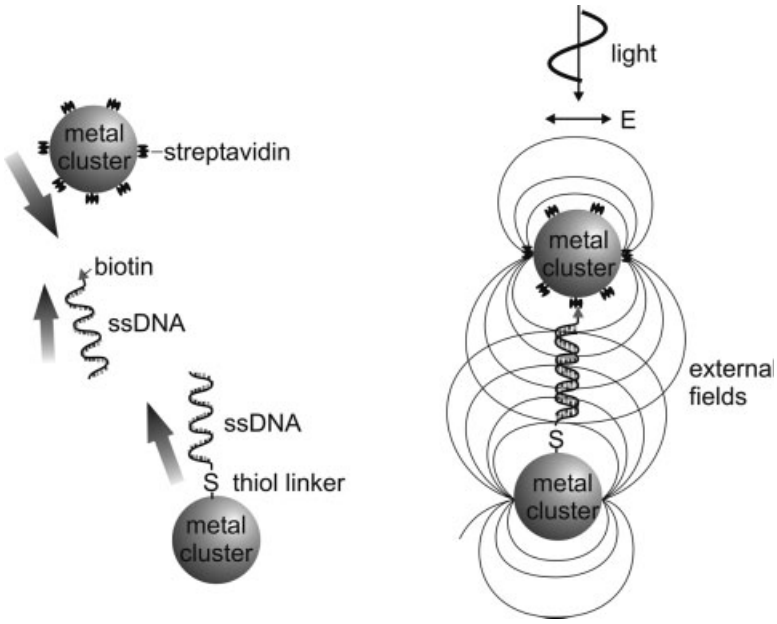


Figure 10.16 High-sensitivity detection of a biomolecular interaction using metal clusters.

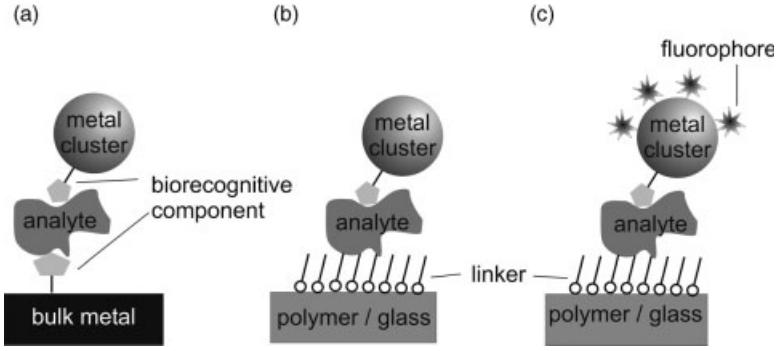


Figure 10.17 Different strategies for metal cluster-enhanced biosensing.

an alternative to the Förster resonance energy transfer (FRET) for monitoring nanometer-scale distances. The plasmon coupling allows measurements to be made over longer time periods and larger distances compared to FRET [20].

Other strategies for biosensing are represented in Figure 10.17. Metal clusters at a defined distance from the surface of a bulk metal (Figure 10.17a) interact with the free electrons of the metal. At a certain distance between the metal cluster and surface, a feedback mechanism enhances the absorption of light, which is then reflected onto the metal surface. The intensity of the absorption is directly proportional to the number of clusters interacting with the metal surface. Metal clusters can be also coupled to a polymer surface (Figure 10.17b); here, the linker molecules not only keep the clusters at a defined distance from the surface but also maintain their spatial distance to other clusters. An optically detectable signal occurs when a certain number of clusters are coupled due to biochemical recognition. Finally, Figure 10.17c shows a similar arrangement where the metal clusters are labeled with fluorophores, and where the fluorescence signal is directly proportional to the bound clusters. The advantage here is an extremely high sensitivity (in the lower femtomolar range) with, under optimal conditions, even single clusters being visible as a color change. The response of the SPR systems is proportional to the product of the adsorbed molecules and the refractive index of the surrounding medium [25]. In addition, metal cluster-based detection provides real-time information on the course of binding over a broad range of binding affinities. Recently, several groups have shown that stable metal cluster layers can be prepared on optical substrates, which makes SPR sensors potentially applicable for continuous-flow detection, as well as for *in vivo* applications in cells and biological fluids [26–29].

More recently, the trend has been towards using nanotechnology to develop sensing surfaces embedded in metal clusters, with such devices improving the sensitivity of a common SPR prism coupler system by a factor of 10. Another novel approach – the use of hydrogels is shown in Figure 10.18. These represent a novel class of polymer that can swell and shrink, depending on their chemical and/or physical environment. As such swelling and shrinkage can be controlled by various molecules, hydrogels may also serve as a host medium for metal clusters. The swelling of a hydrogel containing embedded metal clusters will also cause an increase

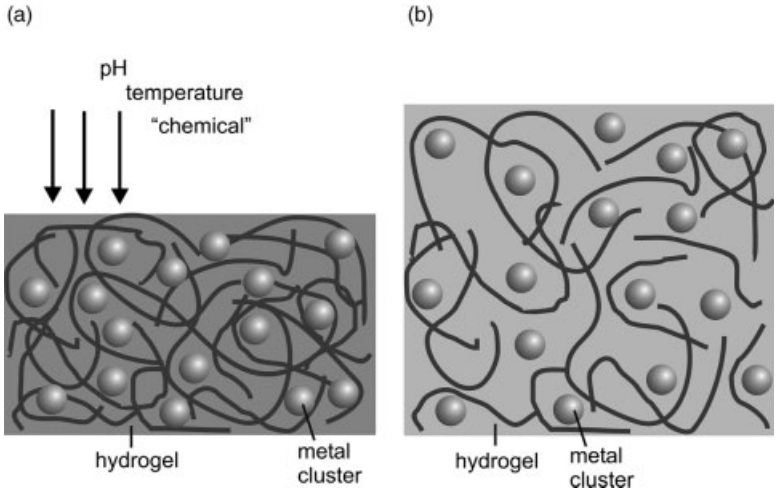


Figure 10.18 Metal clusters embedded in a swelling hydrogel allow sensitive, multiple and reproducible detection of chemical and biochemical parameters.

in the inter-cluster distance, with a spectral shift in color occurring that can easily be measured. An added attraction is that this type of sensor is not only reversible but is also stable over a large number of measurement cycles [30].

References

1 Prasad, P.N. (2004) *Nanophotonics,* John Wiley & Sons Inc, Hokboken, New Jersey.

2 Brongersma, M.L. and Kik, P.G. (eds) (2007) *Surface Plasmon Nanophotonics, Springer Series in Optical Science,* Vol. **131**, Springer, Dordrecht.

3 Ohtsu, M. and Hori, H. (1999) *Near-Field Nano-Optics,* Kluwer Academic/Plenum Publishers, New York.

4 Gluodenis, M., Manley, C. and Foss, C.A. Jr (1999) In situ monitoring of the change in extinction of stabilized nanoscopic gold particles in contact with aqueous phenol solutions. *Analytical Chemistry,* **71**, 4554.

5 Steiner, G., Sablinskas, V., Hübner, A., Kuhne, Ch. and Salzer, R. (1999) Surface plasmon resonance imaging of microstructured monolayers. *Journal of Molecular Structure,* **509**, 265.

6 Kreibig, U. and Vollmer, M. (1995) *Optical Properties of Metal Clusters, Springer Series in Material Science*, Vol. 25, Springer, Berlin, Heidelberg, New York.

7 Endo, T., Kerman, K., Nagatani, N., Hiepa, H.M., Kim, D.K., Yonezawa, Y., Nakano, K. and Tamiya, E. (2006) Multiple label-free detection of antigen-antibody reaction using localized surface plasmon resonance-based core-shell structured nanoparticle layer nanochip. *Analytical Chemistry,* **78**, 6465.

8 Thaxton, C.S. and Mirkin, C.A. (2005) Plasmon coupling measures up. *Nature Biotechnology,* **23**, 681.

9 Kuhne, C. (1999) Investigations to an application of surface-enhanced infrared absorption spectroscopy, (in German), *PhD Thesis,* Dresden University of Technology.

10 Ross, D. and Aroca, R. (2002) Effective medium theories in surface enhanced

infrared spectroscopy: The pentacene example. *Journal of Chemical Physics*, **117**, 8095.

11 Sturm, J., Grosse, P. and Theiß, W. (1991) Effective dielectric functions of samples obtained by evaporation of alkali halides. *Zeitschrift für Physik D - Atoms Molecules and Clusters*, **20**, 341.

12 Stroud, D. (1998) The effective medium approximations: Some recent developments. *Superlattices and Microstructures*, **23**, 567.

13 Theiß, W. (1994) *Advances in Solid State Physics*, Vol. 33 (ed. R. Helbig), Vieweg Braunschweig, Wiesbaden.

14 Bordo, V.G. and Ruban, H.-G. (2005) *Optics and Spectroscopy at Surfaces and Interfaces*, Wiley-VCH, Weinheim.

15 Tarcha, P.J., Desaja-Gonzalez, J., Rodriguez-Llorente, S. and Aroca, R. (1999) Surface-enhanced fluorescence on SiO_2-coated silver island films. *Applied Spectroscopy*, **53**, 43.

16 Jensen, T.R., Van Duyne, R.P., Johnson, S.A. and Maroni, V.A. (2000) Surface-enhanced infrared spectroscopy: a comparison of metal island films with discrete and nondiscrete surface plasmons. *Applied Spectroscopy*, **54**, 371.

17 Wanzenböck, H.D., Mizaikoff, B., Weissenbacher, N. and Kellner, R. (1998) Surface enhanced infrared absorption spectroscopy (SEIRA) using external reflection on low-cost substrates. *Fresenius Journal of Analytical Chemistry*, **362**, 15.

18 Nishikawa, Y., Fujiwara, K., Ataka, K.I. and Osawa, M. (1993) Surface-enhanced infrared external reflection spectroscopy at low reflective surfaces and its application to surface analysis of semiconductors, glasses and polymers. *Analytical Chemistry*, **65**, 556.

19 Kellner, R., Mizaikoff, B., Jakusch, M., Wanzenböck, H.D. and Weissenbacher, N. (1997) Surface-enhanced vibrational spectroscopy: a new tool in chemical sensing? *Applied Spectroscopy*, **51**, 495.

20 Sönnichsen, C., Reinhard, B.M., Liphardt, J. and Alivisatos, A.P. (2005) A molecular ruler based on plasmon coupling of single gold and silver nanoparticles. *Nature Biotechnology*, **23**, 741.

21 Griffiths, P.R., de Haseth, J.A. (2007) *Fourier, Transform Infrared Spectrometry*, John Wiley & Sons, Inc, Hoboken, New Jersey.

22 Bauer, G., Pittner, F. and Schalkhammer, Th. (1999) Metal nano-cluster biosensors. *Microchimica Acta*, **131**, 107.

23 Yonzon, C.R., Stuart, D.A., Zhang, X., McFarland, A.D., Haynes, C.L. and Van Duyne, R.P. (2005) Towards advanced chemical and biological nanosensors – An overview. *Talanta*, **67**, 438.

24 Wang, Q., Yang, X. and Wang, K. (2007) Enhanced surface plasmon resonance for detection of DNA hybridization based on layer-by-layer assembly films. *Sensors and Actuators B*, **123**, 227.

25 Steiner, G., Pham, M.T., Kuhne, C. and Salzer, R. (1998) Surface plasmon resonance within ion implanted silver cluster. *Fresenius Journal of Analytical Chemistry*, **362**, 9.

26 Hoa, X.D., Kirk, A.G. and Tabrizian, M. (2007) Towards integrated and sensitive surface plasmon resonance biosensors: A review of recent progress. *Biosensors and Bioelectronics*, **23**, 151.

27 Nath, N. and Chilkoti, A. (2006) Label-free biosensing by surface plasmon resonance of nanoparticles on glass: optimization of nanoparticles size. *Analytical Chemistry*, **76**, 5370.

28 Chau, K.L., Lin, Y.F., Cheng, S.F. and Lin, T.J. (2006) Fiber-optic chemical and biochemical probes based on localized surface plasmon resonance. *Sensors and Actuators B*, **113**, 100.

29 Steiner, G. (2004) Surface plasmon resonance imaging. *Analytical and Bioanalytical Chemistry*, **379**, 328.

30 Hashimoto, N., Hashimoto, T., Teranishi, T., Nasu, H. and Kamiya, K. (2006) Cycle performance of sol-gel optical sensor based on localized surface plasmon resonance of silver particles. *Sensors and Actuators B*, **113**, 382.

11
Impedance Analysis of Cell Junctions

Joachim Wegener

11.1
A Short Introduction to Cell Junctions of Animal Cells

Within the human body there are more than 200 different, highly specialized cell types, each of which has its own individual functions and the corresponding molecular equipment. However, in order to achieve a specific physiological functionality, it often requires the concerted action of a population of cells, whether of the same kind or of a mixed but well-defined population. Within these organized assemblies of cells – known as *tissues* – the cells must interact with each other mechanically by direct cell-to-cell contacts, chemically by secreting chemicals on the one side and responding to these signals on the other, or electrically by means of cell junctions that transmit electrical signals. In an *in vivo* environment the cells also have to interact with their surrounding *extracellular* material for orientation, migration and signaling. This extracellular material – which is known as the extracellular matrix (ECM) – is most often a complex mixture of proteins and carbohydrates embedded in a more or less aqueous environment, depending on the precise location inside the body. The interactions of cells with other cells of the same or a different type, and the interactions of the cells with the ECM, are summarized by the term cell junctions, including both, cell–cell junctions and cell–matrix junctions [1]. The different cell junctions, as found in epithelial cells, are shown schematically in Figure 11.1. Epithelial cells serve as a good example here as they express most of the important cell junctions very prominently, and in a highly organized fashion. It should be noted that other cell types may lack the more specific junctions, such as tight junctions or gap junctions.

From a functional viewpoint, cell junctions can be grouped in three categories which:

- Provide mechanical contacts and stability of the tissue
- Chemically seal extracellular pathways between cells
- Allow direct molecular or ionic exchange between adjacent cells [2].

Nanotechnology. Volume 6: Nanoprobes. Edited by Harald Fuchs

ISBN: 978-3-527-31733-2

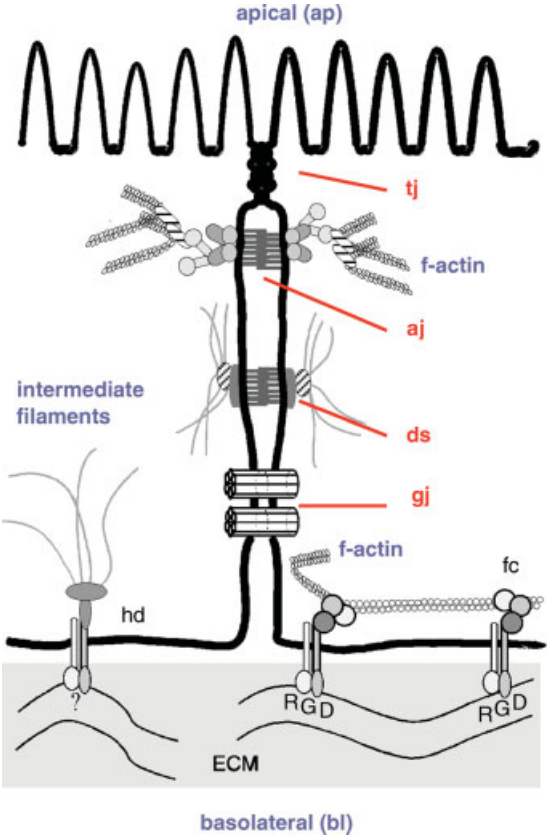


Figure 11.1 Cell–cell and cell–matrix junctions in epithelial cells. The following cell–cell junctions are found in the intercellular cleft between two adjacent epithelial cells: tight junctions (tj), adherens junctions (aj), desmosomes (ds) and gap junctions (gj). The cells are anchored to the extracellular matrix (ECM) by focal contacts (fc) or hemidesmosomes (hd). Cells that provide mechanical stability to cells and tissue are connected to the intracellular filament system of the cytoskeleton, such as the intermediate filaments or the actin cytoskeleton.

However, only specific interactions of the receptor–ligand type mediated by molecular recognition will be addressed in this chapter. Unspecific (electrostatic, electrodynamic, entropic) interactions between two adjacent cells, or between a cell and the surrounding ECM, will not be considered.

11.1.1 Cell Junctions for Mechanical Stability of the Tissue

The mechanical stability of a tissue is provided by cell–cell as well as cell–matrix junctions, which also show a remarkable similarity with respect to their molecular architecture. For both types of junction transmembrane receptors make contact via

their extracellular domains to a corresponding protein on the surface of an adjacent cell, or to a binding site within the ECM. On the intracellular site these receptor proteins are connected to the cytoskeleton mediated by highly specialized adaptor proteins. It is this connection of the receptor proteins to the cytoskeleton (an intracellular network of protein filaments) which provides the molecular basis for the mechanical stability of the individual junction that distributes any punctual mechanical load into the entire tissue. With regard to the individual proteins or protein families involved in junction formation, the mechanical junctions can be subdivided as follows.

11.1.1.1 **Adherens Junctions**

Adherens junctions (aj; see Figure 11.1) are cell–cell junctions that are formed by transmembrane proteins of the *cadherin* family [3]. The name cadherin is derived from the fact that these proteins only bind to their cadherin counterparts on the surface of adjacent cells in the presence of Ca^{2+} (calcium + adhesion). As cadherins on one cell interact with the same cadherins on the opposing cell, the interaction is termed *homophilic*. On the intracellular site, the cadherins are linked to the actin cytoskeleton (also called the *microfilament system*) by a distinct set of linker proteins such as α-, β- or γ-catenin, actinin or paxilin. As shown in Figure 11.1, the adherens junctions with their underlying microfilaments form a very localized structure located close to the upper (apical) pole of the cells, and circumscribe the cell bodies like a belt with bundles of filaments running along on the intracellular site.

11.1.1.2 **Desmosomes**

Desmosomes (ds in Figure 11.1) have a very similar molecular architecture compared to adherens junctions. The transmembrane proteins are also of the cadherin type, although the adaptor proteins on the intracellular side are different and are connected to the *intermediate filament* instead of the microfilament system [4, 5]. As the intermediate filaments are very different from the microfilaments with respect to their dynamic and mechanical properties, both junctions serve the individual needs of the cells that will not be addressed in detail in this chapter (the reader is referred elsewhere for further details [2, 6]). Both types of junction are synergistically responsible for the mechanical properties of cell–cell adhesion. The desmosomes are located further down the intercellular cleft, just beneath the adherens junctions. However, they do not form a belt-like structure around the cells but rather more punctuate, bullet-like contact sites. The connection between cadherins and intermediate filaments is provided by the adaptor proteins *desmoglein* and *desmocolin*. Although both the adherens junctions and desmosomes tie the intercellular cleft together mechanically, they do not operate as barriers for diffusion along the intercellular cleft, other than confining the cleft width.

11.1.1.3 **Focal Contacts**

Focal contacts (fc in Figure 11.1) are the most prominent sites of cell–matrix adhesion [7]. They are clusters of individual molecular connections between the cell interior and the ECM. In general, cell–matrix junctions that eventually develop

into mature focal contacts are composed of a transmembrane protein that makes contact with the extracellular binding partner. The major family of transmembrane proteins involved in cell–matrix adhesion is the *integrin* family [8]. Integrins are α,β-heterodimeric proteins that extend out of the membrane by approximately 20 nm, and are capable of binding specifically to the ECM proteins. Integrins present in focal contacts are connected to the actin cytoskeleton via linker proteins such as vinculin, paxilin and talin. By means of this molecular construction, an intracellular macromolecular network is connected mechanically to an extracellular macromolecular network, such that the cells and the extracellular environment form a unit that is remarkably resistant to mechanical challenges [8, 9]. Moreover, the binding affinity of integrins in focal adhesion sites can be regulated by the cells either to form or to loosen, and in this way they play a major role in transmembrane signaling in both directions, inside-out and outside-in [2].

11.1.1.4 **Hemi-Desmosomes**

The hemi-desmosomes (hd in Figure 11.1), as the second class of cell–matrix junctions, differ from the focal contacts in essentially the same way as do desmosomes from adherens junctions. The transmembrane component is provided by a special type of integrin ($\alpha_6\beta_4$) that is exclusively localized in the hemi-desmosomes. Moreover, hemi-desmosomes are connected to the *intermediate filament* system rather than to the microfilaments. As mentioned above, the different properties of these intracellular cytoskeletal networks provide individual functionalities to the two different classes of cell–matrix adhesion sites.

11.1.1.5 **Less-Prominent Types of Mechanical Junctions**

In addition to the above-mentioned cell junctions there are other, less-prominent molecular assemblies that provide mechanical stability. These include mainly the Ca^{2+}-independent cell adhesion molecules (CAM) as important mediators of cell–cell adhesion, or transmembrane proteoglycans for cell–matrix adhesion. Further details regarding these assemblies are provided elsewhere [2].

11.1.2
Cell Junctions Sealing Extracellular Pathways: Tight Junctions

There is only one type of junction responsible for sealing the extracellular pathway between adjacent cells, and these are known as *tight junctions* (tj in Figure 11.1). To the author's present knowledge, tight junctions are expressed only in epithelial and endothelial cell layers which form the interfacial layers along all inner and outer surfaces of the human body, such as the skin, the gut lining, the bladder or blood vessels. By virtue of their location, it is the predominant physiological task of these tissues to serve as an interface between the two separated compartments, and to control the flux of metabolites or xenobiotics from one compartment to another [10]. Flux control and the exclusion of selected substances is, however, only effective as long as any uncontrolled paracellular diffusion through the intercellular cleft between adjacent cells is limited. Tight junctions provide this seal or *occlusion* of the intercellular

cleft and are, thus, also referred to as *occluding junctions* or *zonula occludens*. The tight junctions are located at the apical pole of two apposing epithelial cells (cf. Figure 11.1). Similar to adherens junctions, they are also very focused structures that span no more than 200 nm along the intercellular cleft. Nonetheless, they can be extremely efficient in maintaining chemical or electrochemical gradients in highly specialized tissues such as those which form the blood–brain or the blood–CSF barriers. In other epithelia or endothelia, the barrier function is significantly lower according to the physiological task at the particular location inside the body [11].

Structurally, the tight junctions are not yet fully understood. Many transmembrane or peripheral proteins have been localized almost exclusively to functional tight junctions, including *Occludin*, members of the *Claudin* or *Jam* family, as well as the peripheral proteins ZO-1, ZO-2 and ZO-3, to mention a few. Although a large number of proteins have been identified as constituents of functional tight junctions, various experimental indications exist which suggest that lipids must be involved in junction formation and thus provide some of the junctions' unique properties [12]. (For additional information the reader is referred to [13–17].) Yet, no matter how they are molecularly composed, the formation of tight junctions requires close cell-to-cell-apposition. Tight junctions will only form when a mechanically stable cell–cell adhesion site has been established through adherens junctions and desmosomes before.

11.1.3 Communicating Junctions: Gap Junctions and Synapses

Both, *chemical synapses* and *gap junctions* (gj in Figure 11.1) are cell junctions that are involved in intercellular communication, despite their entirely different structure and molecular architecture. Moreover, it is well known that synapses are only formed at the contact sites between two neurons or a neuron and a muscle cell, whereas gap junctions are expressed by most cell types within the human body, although to very different degrees.

11.1.3.1 Chemical Synapses

In chemical synapses [18] the opposing membranes of two cells approach each other very closely, leaving a water-filled cleft of only 20 nm between them (the synaptic cleft). When a membrane depolarization wave passing along the presynaptic membrane reaches the synapse, transmitter molecules are released from the sender cell, diffuse through the synaptic cleft, and open ligand-gated ion channels in the plasma membrane of the receiver cell (postsynaptic membrane). This incident triggers a depolarization wave in the receiver cell, such that the electrical signal is transmitted. Depending on the length of the axon, the distance between two communicating nerve cells can be up to 1 m [19].

11.1.3.2 Gap Junctions

In contrast to the chemical synapses, gap junctions are simply water-filled channels which locate between two adjacent cells and allow the sharing of molecules with a molar

mass of less than 1000 g mol^{-1} (this is termed metabolic coupling) [20, 21]. From a structural viewpoint gap junctions are composed of two hemi-channels, one provided by each of the opposing cells. Each hemi-channel is known as a *connexon*, which in turn is composed of six *connexins* (each of which is a single span transmembrane protein) arranged in a hexagonal pattern. The central opening in this protein cluster ($d = 1.5$ nm) provides an aqueous channel between the two connected cells. These channels are not static but can be precisely regulated in their permeability by both cells, the sender and the receiver. As membrane depolarization waves can also be transmitted from one cell to the neighboring cells via gap junctions (e.g. in heart muscle cells), they are also referred to as *electrical synapses*. Notably, the transfer of electrical signals occurs significantly faster via gap junctions than via chemical synapses.

11.2 Established Physical Techniques to Study Cell Junctions

In this section we will provide a rather brief overview of the techniques which have been applied in the past to study cell–cell or cell–matrix junctions from a structural or functional viewpoint. It should be noted, however, that within the chapter we cannot provide a complete survey of all available techniques, nor details of the selected methods. However, further information is available via the references provided in the appropriate passages of the text.

11.2.1 Cell–Matrix Junctions

When studying cell–matrix junctions *in vitro*, the ECM proteins are commonly predeposited on a technical surface such as a Petri dish or a microscope slide, with the cells adhering to this protein-decorated surface. Thus, analyzing *cell–matrix junctions* also means studying *cell–surface junctions* or *cell–substrate junctions* [22]. Today, these three terms are generally used synonymously, depending on the background of the study under discussion.

From the structural viewpoint our current knowledge of cell–matrix or cell–surface junctions is largely based on light microscopy or transmission electron microscopy (TEM) of chemically stabilized (fixed) samples, in combination with the powerful tools of modern molecular biology and immunology. Often, the target molecule suspected of being involved in cell–matrix adhesion is specifically tagged by an antibody – which itself is labeled with a fluorochrome or an electron-dense marker – and then imaged microscopically under various experimental conditions. Studies like this, together with the biochemical analysis of interaction partners, have identified the molecular architecture of cell–matrix junctions as known today and described above.

11.2.1.1 Scanning Probe Techniques

Until now, scanning probe techniques have not contributed a lot to the structural analysis of cell–matrix adhesion sites, as the latter are protected from one side by the

cell body and from the other side by the growth substrate. Nanoprobes simply cannot gain access to these structures in living cells. Most recently, scanning electron microscopy (SEM) has been used in combination with controlled erosion of the sample by a focused ion beam (FIB). Thereby, the organic material of the cell bodies was locally removed in order to obtain a side view of the cell by SEM. These images provided a detailed view on the profile of the cell–substrate interface [23]. Recently, Wrobel *et al.* studied the cell–surface interface using TEM after preparing of thin sections of the sample [24].

11.2.1.2 **Nonscanning Microscopic Techniques**

Several nonscanning light microscopy techniques have been developed that utilize certain optical effects to provide images from this internal interface between the cell body and the growth substrate. Reflection interference contrast microscopy (RICM) – also referred to as interference reflection microscopy (IRM) – has contributed the most to the existing literature about cell–substrate interactions [25, 26]. This technique allows imaging of the 'footprints' of cells on a substrate rather than only the projections of the cell body. RICM in its basic form is applied to living cells grown on ordinary coverslips, and does not require any staining or fixation. The sample is illuminated from below with an inverted microscope, using monochromatic light. In a first approximation the image is generated from the light reflected either from the glass/medium interface or the lower plasma membrane. Interference of the reflected light then provides an image of the cell–substrate contact area in which the brightness of the pixels code for the optical path difference between the two interfaces. Thus, RICM images map the distance between the lower cell membrane and the glass surface, while time-lapse RICM studies provide a microscopic view of the dynamics of cell–surface junctions with video rate time resolution.

11.2.1.3 **Fluorescence Interference Contrast Microscopy**

A major improvement with respect to absolute cell–substrate distance measurements was introduced by Braun, Lambacher and Fromherz in 1997 [27–29]. For this technique, which is referred to as fluorescence interference contrast microscopy (FLIC), the cells are grown on silicon substrates that have well-defined steps made from silicon oxide on their surface. The step height ranges between 20 and 200 nm and is, thus, only a fraction of the wavelength of visible light. For the measurement, the adherent cells are stained with a fluorescent dye that integrates into the plasma membranes. As the silicon/silicon oxide interface acts as a mirror, the intensity of the fluorescence light emitted by the dye in the lower membrane is dependent on the distance between the membrane and the silicon surface. The steps on the surface serve as well-defined spacers and make the intensity–distance relationship unique. FLIC microscopy provides the cell–substrate separation distance with an accuracy better than 1 nm.

11.2.1.4 **Total Internal Reflection (Aqueous) Fluorescence Microscopy**

Total internal reflection fluorescence microscopy (TIRF) and total internal reflection aqueous fluorescence microscopy (TIRAF) are variants of the same microscopic

principle, and are based on fluorophore excitation by an evanescent electric field. Here, the cells are grown on a transparent substrate that is illuminated from below with a laser beam at an angle θ relative to the surface normal that is bigger than – or equal to – the critical angle of total reflection, θ_{crit}. Under these conditions diffraction phenomena at the interface generate an evanescent wave at the surface. The penetration depth of the associated electric field is rather short, and the field decays within 100 nm of the surface, or slightly beyond. Thus, fluorescence is only excited in those molecules that are close enough to the surface. In TIRF, membrane proteins or other membrane constituents are fluorescently labeled and can be imaged with improved resolution, as fluorescence light from further inside the sample is not excited. For TIRAF measurements a water-soluble fluorescent dye is added to the extracellular fluid. If a cell adheres to the surface it displaces the aqueous phase and thereby the fluorophore from the interface. Thus, cell-covered areas appear dark in TIRAF images. Both techniques have contributed significantly to our understanding about cell–matrix adhesion *in vitro* [30, 31].

11.2.1.5 **Quartz Crystal Microbalance**

Another emerging tool to study cell–matrix junctions *in vitro* is based on using thickness shear-mode piezo resonators as growth substrates for adherent cells. The interactions of cells with the protein-decorated crystal surface is monitored by reading the resonance frequency and the energy dissipation of the shear oscillation [32]. In principle, this approach has evolved from the so-called quartz crystal microbalance (QCM) technique that is a widely accepted technique for following adsorption reactions at the solid–liquid interface. However, in combination with cells it is not the mass of the cells that determines the signal but rather their anchorage to the oscillating quartz surface, together with the viscoelasticity of the cell bodies. Time-resolved measurements of the resonance frequency can be used to follow the attachment and spreading of cells to the quartz surface that may be precoated with a matrix component of interest [33]. Even though the growth substrate of the cells oscillates mechanically, the technique is considered as being noninvasive as the maximum shear displacement in the center of the resonator is in the order of 1 nm, with a frequency in the MHz regime.

11.2.1.6 **Other Techniques**

Besides the various techniques mentioned above, several other experimental methods can be used to provide information on cell–matrix adhesion *in vitro*; some details of these are listed very briefly here. Many assays study the forces necessary to remove a cell from a protein-coated surface after it has been allowed to attach and anchor for a predefined time. These approaches can be applied either in an integral manner to a population of cells, or on the single-cell level. The former method is, for instance, realized by exposing the cells to centrifugal forces and counting those that are capable of remaining attached to the matrix proteins on the surface [34]. On the single-cell level, a variant of classical scanning force microscopy – then referred to as cell adhesion force microscopy – is used to measure the forces required to remove a cell from a particular ECM-coated substrate, either by pulling or pushing [35]. Moreover

much has been learned from extracellular recordings of neuronal action potentials by means of field effect transistors (FET) regarding the electrochemical properties of the thin cleft between cell and surface as the sealing properties of the junctional area determine the sensitivity of the measurement [36].

11.2.2 Cell–Cell Junctions

The molecular architecture of cell–cell junctions, as shown in Figure 11.1, has been unraveled significantly by using both TEM and SEM after freeze-fracture preparation or other contrasting protocols, as well as with fluorescence microscopy after immunolabeling of the molecular target. When used in conjunction with the molecular biology technique to knock out or knock in a gene of interest, to silence the expression of a given gene or to identify interactions partners, these techniques have provided a comprehensive understanding of the molecular composition of cell–cell junctions. Whilst understanding the molecular arrangements of a given cell junction under stationary conditions is one thing, to follow and identify the changes that occur during regulation or development is another. Thus, a detailed understanding of cell–cell junctions, their interplay and regulation remains an area of intense research worldwide, notably because of their extraordinary biomedical and pharmaceutical relevance. Among the four different types of cell–cell junction (tight junctions, adherens junctions, desmosomes, gap junctions), tight and gap junctions have received most attention with respect to their functional properties, and hence only these will be addressed here. Both of these junctions provide control over the flux of chemicals (metabolites, xenobiotics), either through the interspaces between two adjacent cells or between adjacent cytoplasms.

11.2.2.1 Tight Junctions

Tight junctions, as expressed by endothelial and epithelial cells, form the structural basis for the barrier function of epithelial and endothelial cell layers. A straightforward and popular approach to probe the efficiency of this barrier function – and thus the tightness of the junctions – is a simple *permeation/diffusion assay* [37]. Here, the cells are grown on highly porous filter membranes that support the cell layer mechanically without acting as a significant diffusion barrier themselves. The cell-covered membrane is then placed between two fluid compartments such that any flux of solutes from one compartment to the other must pass the interfacial cell layer. In a typical experiment, a tracer compound is added to one compartment (donor), while samples are taken from the other compartment (acceptor) after well-defined time intervals. From the concentration increase of the tracer in the acceptor compartment, it is possible to calculate the permeation rate P_E that reports on the barrier properties of the cell layer under study. As long as the probe cannot migrate across the cell membranes and is dependent on extracellular diffusion, the experimentally measured flux is predominantly determined by the functional properties of the tight junctions. By using probes of different molecular mass or shape, the size-exclusion properties of the junctions can be inferred from such measurements [38].

The probes can either be radio- or fluorescence-labeled, otherwise their concentration must be determined using chromatography. The readout of this widely established assay is, however, easily compromised by defects in the cell layer (even single cell defects) that may serve as short-cuts for the substrate flux leading to a serious underestimation of the barrier function. Moreover, such a permeation assay is not an *in situ* method but rather requires time for the probe to accumulate in the acceptor compartment [39].

Another method of probing the functional properties of tight junctions relies on measuring the *electrical resistance* of the cell layer, when it is placed on a porous filter membrane between two fluid compartments, as described above. Either compartment is equipped with a pair of Ag/AgCl electrodes, that are used for current injection and voltage reading in either compartment. The resistance value, which is then referred to as the transepithelial or transendothelial electrical resistance (TER), strictly reports on the ionic permeability of the entire cell layer. However, as the extracellular ionic current pathway around the cell bodies through the cell–cell junctions is significantly less resistive in most cases than the ionic current pathways across the dielectric plasma membranes, TER reports on the tightness of the junctions as a first approximation. Unfortunately, the TER approach suffers from the same shortcomings as do readings of the P_E rate (see above), namely that it is very prone to artifacts arising from defects within the cell monolayer. On the other hand, TER readings can provide a snapshot of the barrier function of the junctions, and the readings can be performed in a time-resolved manner. Accordingly, changes in junctional tightness can be monitored more or less in real time with a time resolution that can be reduced to the order of minutes [40].

TER measurements with lateral resolution have been performed by using microelectrodes for local potential measurements while a uniform and homogeneous current density was established across the entire cell layer [41, 42]. By scanning the electrode across the cell layer, it was possible to establish resistance maps that could identify local shortcuts in the barrier function of the cell layers [43]. Moreover, it was possible to quantify the resistance of the *paracellular* current pathway across the tight junctions in contrast to the *transcellular* current pathway across the membranes [42]. In particular, this discrimination between paracellular and transcellular resistance is not possible with the integral TER approach, and may leave data interpretation with some ambiguity as the transcellular current pathway has a high but finite resistance after all. Several modifications of the basal conductance scanning technique have been described and, in scanning ion conductance microscopy (SICM), the lateral resolution has been improved to the submicrometer regime such that the junctional tightness along the periphery of a single cell can now be recorded and analyzed. Korchev and coworkers have even studied the conductance of a single ion channel on a living cell by using SICM [44].

11.2.2.2 **Gap Junctions**

As for the other junctions, the general structure and composition of gap junctions has been revealed to a large extent by using electron and fluorescence microscopy after labeling individual molecules suspected of being involved. From a functional

perspective, junctional performance is routinely monitored by electrical measurements. Either of the two cells that are connected via gap junctions will be impaled by a microelectrode such that current can be injected to flow from one cell the other. When gap junction performance is discussed, the measured resistance is usually expressed as *conductance*. Amazingly, the conductance of a single gap junction channel depends on the individual member of the connexin protein family that forms the transmembrane channels, even though they are highly conserved [45].

Another approach involves the use of micropipettes to inject nonmembrane-permeable fluorescent dyes into the cytoplasm of a cell; the spread of the dye into adjacent cells connected via gap junction channels is then followed. Using this technique allows complete control over the size, surface charge and morphology of the probe, so that the junctional permeability can be extensively analyzed [46]. Along the same lines, but at a much lower experimental cost, it is possible to visualize gap junction conductance using the so-called scrape loading assay. Here, the cell layer is bathed in a buffer that contains the fluorescent gap junction probe. In order to introduce the dye into the cytoplasm of the cells, a needle is moved (scraped) through the cell layer. Those cells lying in the path of the needle are ripped by the moving pipette tip such that their membrane is ruptured. This allows the extracellular dye to enter the cytoplasm through the ruptured membrane, and then diffuse via gap junctions into neighboring cells. The diffusion performance of the dye reports on junctional coupling [47].

11.3 Impedance Spectroscopy

Impedance spectroscopy (IS), which is also referred to as electrical or electrochemical impedance spectroscopy (EIS), represents a versatile approach for probing and characterizing the dielectric and conducting properties of bulk materials, composite samples or interfacial layers. The technique is based on measuring the impedance – that is, the opposition to current flow – of a system while it is excited with low-amplitude alternating current or voltage. The impedance spectrum is obtained by scanning the sample impedance over a broad range of excitation frequencies, typically covering several decades. In continuous-wave IS, the impedance is measured sequentially at each individual frequency. In contrast, with pulse techniques the system is exposed to a superposition of multiple sine waves with different frequencies, at the same time. The impedance spectrum of the system is then extracted from the transient response by means of Fourier transform algorithms. Both data acquisition modes have their unique advantages. Whereas continuous-wave recordings are easier to conduct and have better signal-to-noise ratios, pulsing techniques are much faster, provide a better time resolution and are, thus, more suitable to study dynamic systems that change within seconds.

As long ago as the 1920s, research groups first began to investigate the impedance of tissues and biological fluids, and it was known even then that different tissues exhibit distinct dielectric properties, and that the impedance undergoes changes

during pathological conditions or upon changing the cellular environment [48, 49]. Thus, the value of impedance measurements for the analysis of biological samples such as organs, tissues, cell aggregates or even single cells was obvious from an early stage. One of the most important advantages of IS compared to other techniques is its noninvasiveness, as the technique relies entirely on low-amplitude currents and voltages that ensure damage-free examination with a minimum disturbance of the cells or tissues. This noninvasive nature of the method, combined with its high information content, has made it a valuable tool for biomedical research *in vitro*. More recently, *in vivo* applications in clinical settings have also been developed, including electrical impedance tomography (EIT), which is being used increasingly as a noninvasive monitoring tool for heart and/or lung function during surgery [50]. In the following sections, however, attention will be focused on the use of impedance spectroscopy to study cell junctions of animal cells *in vitro*. A broad survey of the use of EIS for biological samples such as tissues and organs is provided elsewhere [50].

11.3.1 Fundamental Relationships in Impedance Analysis

The electrical impedance, Z, is a complex quantity that describes the ability of the system under study to resist the flow of alternating current. In a typical IS experiment a sinusoidal voltage $U(t)$ with angular frequency $\omega = 2\pi f$ is applied to the system and the resulting steady-state current $I(t)$ associated with the voltage is measured. According to Ohm's Law, the impedance is given by the ratio of these two quantities:

$$Z = \frac{U(t)}{I(t)} \tag{11.1}$$

The impedance measurement is typically conducted within a linear voltage–current regime of the sample (i.e. the measured current amplitude is proportional to the amplitude of the applied voltage), so that the resulting current will also be a sine wave with the same frequency ω as the applied voltage signal, but it may be phase-shifted relative to the voltage by the phase angle φ. By introducing a complex notation, Equation 11.1 translates into

$$Z = \frac{U_0}{I_0}\exp(i \cdot \varphi) = |Z|\exp(i \cdot \varphi) \tag{11.2}$$

with U_0 and I_0 representing the amplitudes of voltage and current, respectively, and with $i = \sqrt{-1}$. Thus, at each frequency of interest the impedance is described by two quantities: the magnitude $|Z|$, which is the ratio of the amplitudes of U_0 and I_0, and the phase angle φ between voltage and current.

Instead of presenting the complex impedance Z in polar coordinates, $|Z|$ and φ, it can also be expressed in Cartesian coordinates, with a real (R) and an imaginary (X) component:

$$Z = R + i \cdot X \tag{11.3}$$

$$\begin{aligned} &\text{with} \quad R = \mathrm{Re}(Z) = |Z| \cdot \cos(\varphi) \\ &\text{and} \quad X = \mathrm{Im}(Z) = |Z| \cdot \sin(\varphi) \end{aligned} \tag{11.4}$$

The real part is called the *resistance R*, and corresponds to the impedance contribution arising from current that is in-phase with the applied voltage. The imaginary part is termed the *reactance X*, and describes the impedance contribution from current which is 90° out-of-phase with the voltage. With respect to biological samples, the resistive portion R of the impedance mirrors either the concentration of ions available for current flow within the sample, or their limited ability to migrate under the influence of the applied electric field. The latter is commonly caused by a geometric confinement of the ionic current pathways by means of insulating cell membranes or cell junctions that occlude or narrow down the aqueous spaces available for current flow. The reactance X arises from the presence of storage elements for electrical charges like, for instance, capacitors or coils in electrical circuits. In biological tissues, the dielectric (or insulating) cell membrane acts as a capacitor, separating two conducting fluids (extracellular and intracellular) by its hydrophobic core, whereas an inductive behavior of biological systems is only described in very rare cases. Thus, as a rule of thumb for animal cells and tissues, it can be said that the measured resistance arises from extracellular or intracellular fluids and their geometric dimensions, whereas the capacitive reactance originates from the cell membranes.

In some cases it is more convenient to use the inverse quantities of Z, R and X, which are referred to as *admittance* $Y = 1/Z$, *conductance* $G = \mathrm{Re}(Y)$ and *susceptance* $B = \mathrm{Im}(Y)$, respectively. In the linear voltage–current regime, the two representations are interchangeable and contain the same information. Accordingly, IS is occasionally entitled admittance spectroscopy.

In IS the impedance is measured over a range of frequencies covering several decades between mHz and GHz, depending on the type of sample and the problem being studied. The frequency regime typically studied to analyze cell junctions ranges from 1 Hz to 1 MHz.

11.3.2
Data Representation and Analysis

When the complex impedance of a system of interest (real and imaginary part) is recorded as a function of frequency, the complete presentation of the data requires a specialized method of data plotting. Most frequently, a so-called Bode diagram is used; here, the impedance magnitude $|Z|$ and the phase shift between current and voltage φ are plotted as a function of frequency on a logarithmic or semi-logarithmic scale, respectively. Figure 11.2a shows such a Bode diagram for an arbitrary electrical network shown in the inset of Figure 11.2b as an example. Alternatively, the imaginary component of the impedance X is plotted versus the real component R in a so-called impedance locus or Wessel diagram (Figure 11.2b). The latter presentation does not provide any further information on the frequency. Normally, an arrow is added to include the direction in which the sampling frequency

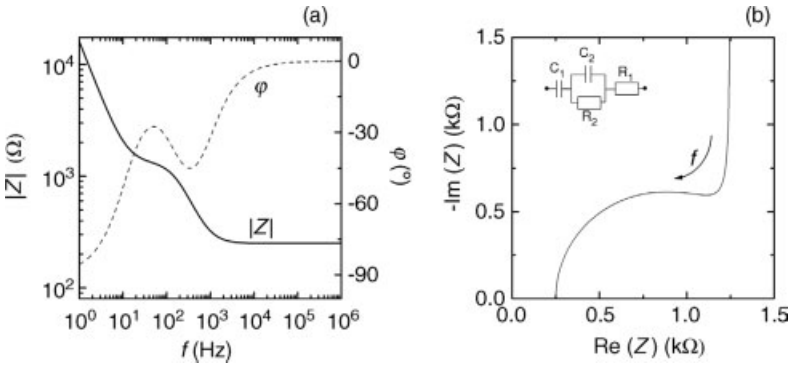


Figure 11.2 Different representations of impedance spectra. (a) Bode diagrams provide impedance magnitude $|Z|$ and phase shift φ between voltage and current as a function of frequency; (b) Impedance loci or Wessel diagrams display the imaginary part of Z as a function of the real part of Z. The arrow indicates the direction of increasing frequency (f). The insert shows the electrical circuit that was used to calculate the impedance data.

increases. The advantage of using impedance loci for data presentation is that the shape of the curve provides information on the electrical structure (perhaps even substructures of the sample), and also whether the behavior of the system deviates from that of ideal electrical networks. For example, a resistive current pathway in parallel to a capacitive one gives rise to a semicircle within an impedance locus that is centered on the x-axis. From the center and the radius of the semicircle one can directly extract the resistance and capacitance of this particular structure (cf. Figure 11.2b).

Although the shape of the curve in impedance loci may be useful for direct analysis, the most common method of analyzing experimental impedance spectra is by means of equivalent circuit modeling. Here, the system under study is described by an electrical network (as shown in the insert of Figure 11.2b) that mirrors the (predicted) electrical structure of the system. The system's equivalent circuit is composed of series or parallel connections of impedance elements (resistors, capacitors), as known from electronic circuitry, plus additional impedance elements that have been empirically derived for ionic systems, for example Warburg impedance or constant phase elements (CPEs). These elements have no correspondence in electronic systems. As described above, the impedance spectrum contains much information about the electrical properties of the system, and with experience it is possible to make a 'qualified guess' of a proper model based on the features in the diagrams. For a given equivalent circuit, the frequency-dependent impedance (transfer function) is then derived from the individual components and their interconnection using Ohm's law and Kirchoff's laws. The best estimates for the parameters – that is, the unknown values of the resistors and capacitors within the equivalent circuit – are then iteratively computed by ordinary least-squares algorithms, such as the Levenberg–Marquardt approach. If the impedance and phase spectra of the chosen model fit the

Table 11.1 Individual impedance contributions of ideal and empirical equivalent circuit elements.

Component of equivalent circuit	Parameter	Impedance Z	Phase shift φ
Resistor	R	R	0
Capacitor	C	$1/(i\cdot\omega\cdot C)$	$-\pi/2$
Coil	L	$i\cdot\omega\cdot L$	$+\pi/2$
Constant phase element (CPE)	A, $n(0 \leq \alpha \leq 1)$	$1/(i\cdot\omega)^{n}\cdot A$	$-n\cdot\pi/2$
Warburg impedance σ	σ	$\sigma\cdot(1-i)\cdot\omega^{-0.5}$	$-\pi/4$

data well, the parameter values are used to describe the electrical properties of the system and its changes throughout an experiment.

In order to find an equivalent circuit model that accurately predicts the impedance of biomaterials, it is often necessary to include nonideal circuit elements – that is, elements for which the parameters are themselves frequency-dependent. Such empirical elements account for ionic phenomena such as adsorption and diffusion that cannot be realized with standard electronic impedance elements. A list of all common circuit elements used to describe biomaterials in terms of their impedance and their phase shift is provided in Table 11.1. The CPE, which represents a nonideal capacitor and is one of these empirical impedance elements, was originally introduced to describe the interface impedance of noble metal electrodes immersed in electrolyte solutions. Although the physical basis of CPE behavior is not fully understood in detail, it is thought to be associated with surface roughness and specific ion adsorption to interfaces. Another empirical element is the Warburg impedance σ, which accounts for the impedance contribution arising from the diffusion limitation of many electrochemical reactions. The parameters that determine the individual impedances of these elements are listed in Table 11.1.

At this point it is important to place a word of caution concerning the equivalent circuit modeling approach. Different equivalent circuit models (which deviate with respect to either the components or the network structure) may produce equally good fits to the experimental data, yet ascribe the sample a very different physical structure. In such a case, independent experiments (microscopy and other spectroscopic approaches) are required to obtain further insight into the electrical structure of the sample and to identify the most appropriate model. It may also be tempting to increase the number of elements in a model to obtain a better agreement between experiment and model. However, the model may then become redundant because the components can no longer be quantified independently. Thus, an overly complex model can easily provide artificially good fits to the impedance data but, at the same time, highly inaccurate values for the individual parameters. Thus, it is sensible to use the equivalent circuit with the minimum number of elements that still describes all details of the impedance spectrum (the nonredundant model) [51].

Another approach towards analyzing impedance data (although less often applied) is based on deriving the current distribution in the system by means of differential

equations and boundary values. Solving the differential equations provides the impedance transfer function with the respective model parameters. Fitting of the transfer function to the recorded data then allows extraction of the best estimates for the model parameters. It should be noted that both approaches are essentially only different formalisms.

11.4 Impedance Analysis of Cell Junctions

11.4.1 General Remarks about Experimental Issues

11.4.1.1 Two-Probe versus Four-Probe Measurement

In order to analyze the impedance characteristics of a given electrochemical system, it must be interfaced with electrodes that are required for current injection and voltage sensing at appropriate frequencies. These electrodes are placed at opposite ends of the sample in order to provide the electrical structure of the entire system as a readout. In many cases the measurement can be made with four electrodes, two for current injection and two for voltage sensing. Here, the voltage sensing electrodes are usually placed very close to the sample surface so that only the voltage drop across the sample is measured without contributions from the bathing fluid or other parts of the measurement chamber. The current-injecting electrodes are often placed at the very end of the experimental chamber to avoid any disturbance of the sample when electrochemical reactions occur at the electrode surfaces. When low-amplitude signals are used and no significant faradaic currents (electrochemical reactions at the electrode surface) occur, it is often more convenient to work with two electrodes only. Under these conditions, either electrode is used for both current injection and voltage sensing. The major advantage of such a two-probe measurement is the reduction of electrodes, including the necessary cables and connectors, that must be integrated into the experimental set-up. Moreover, some approaches simply do not allow the use of four electrodes at the required location, as will be detailed below. The disadvantage of two-probe measurements is, however, that the electrical properties of the bathing fluid, the chamber and the electrodes themselves will be included in the experimental data. Thus, it is necessary to find ways to reduce their impact, to perform a meaningful and justified subtraction, or to include them into the model that is used for data analysis.

11.4.1.2 Introducing Electrodes for Impedance Readings into an Animal Cell Culture

In order to apply electrochemical impedance techniques to study cell junctions of animal cells, the initial experimental challenge is to introduce the necessary electrodes without causing major disturbances to normal cell division or differentiation. The measurement requires a homogeneous electric field to be applied across the cell layer or to the individual cells. In recent years, three systems have been established that fulfill these conditions in principle. The most popular set-up makes use of highly

permeable membranes that are manufactured either from polymers such as polycarbonate, or from aluminum oxide, and mechanically support the cell layer at the interface between two fluid compartments. When the electrodes are introduced into these fluid compartments below and above the cell layer, an impedance analysis of the cells on the filter can be carried out. It is important that the electrodes provide a homogeneous electric field across the entire cell layer. As the filter membranes on which the cells are grown often range from several millimeters to centimeters in diameter, it is insufficient to introduce point-like Pt or Ag/AgCl electrodes in either compartment. For meaningful impedance readings, it is necessary to use 2-D electrodes for current injection in order to ensure a homogeneous current penetration through the system. In the past, both, four-probe and two-probe electrode configurations have been used to study cells attached to permeable membranes [40, 52].

A slightly different approach that is, however, only suitable for intact tissues or tissue fragments, uses grids made from Pt or Ag/AgCl to serve simultaneously as the bottom electrode and as a mechanical support for the biological sample. Small pieces of tissue are fixed mechanically to these grids after having been excised from an animal or organ [53]. The electrodes in the upper compartment above the cell layer are easier to realize as they can simply be dipped into the bulk electrolyte above the cell layer. As the pore size of these grids is significantly larger than the size of the individual animal cells (5–20 µm in diameter), these systems cannot be used for cultured cells as these are normally seeded into the culture vessels as single-cell suspensions that will form a continuous cell layer with time.

The third and most recent approach uses thin-film electrodes made from inert noble metals such as gold or inert metal oxides such as indium-tin oxide (ITO). The cells are grown directly on the surface of the electrodes after a layer of adhesive proteins has been preadsorbed; such preadsorption is either performed intentionally before the cells are introduced, or occurs spontaneously from the culture medium. This technique, known as electric cell–substrate impedance sensing (ECIS), was pioneered by Giaever and Keese during the 1980s [54, 55], and today is on the verge of becoming a routine laboratory technology. The measurement principle is illustrated in Figure 11.3 where, as indicated in the insert, the distance between the electrode surface and the cell body is only on the order of 20–200 nm. As the electrical potential within this thin cleft is position-dependent, it is impossible to design a four-probe measurement for these systems. Thus, impedance readings of cells or cell layers that have been grown directly on the surface of thin-film electrodes will always contain contributions from the measuring electrodes and the bathing fluid that must be taken into account. Hence, the experimental set-up and electrode layout must be designed such that these contributions do not mask the impedance of the sample.

The coplanar electrode design shown in Figure 11.3 is composed of two electrodes of very different surface area (factor 1000). The smaller electrode is referred to as the *working electrode*, whereas the larger electrode is called the *counter electrode*. The area ratio between the counter- and working electrodes is typically 500 to 1000. The rationale behind this electrode layout is to make one electrode the bottleneck for the current, so that the total impedance of the system is dominated by the small electrode

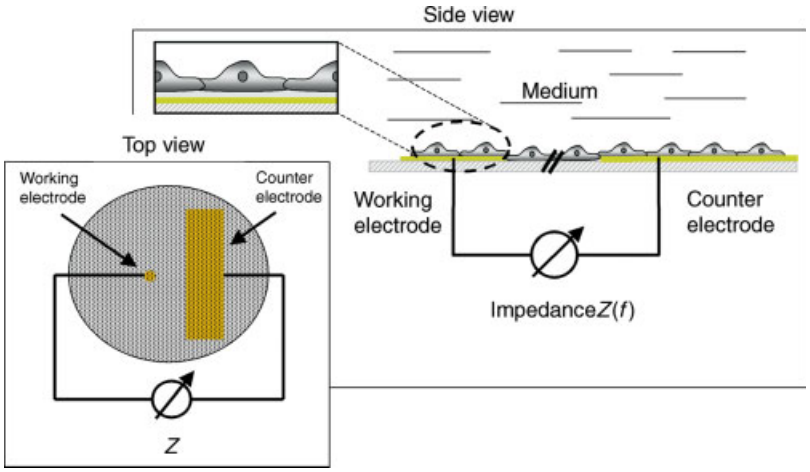


Figure 11.3 Schematic illustrating the principle of impedance analysis of adherent animal cells by means of thin-film electrodes. The cells are grown directly on the surface of the electrodes. The working electrode is made approximately 1000-fold smaller in surface area than the coplanar counter electrode, so that the cell-covered working electrode dominates the readout.

with its small population of cells. Contributions from the counter electrode can be neglected because of its larger area, even though it is also entirely cell-covered. The use of thin-film electrodes, on which the cells are grown, results in several technical advantages:

- Instead of using a dipping electrode that reaches into the bulk electrolyte from above, the second electrode is also a thin-film electrode deposited on the growth surface at sufficient distance from the first electrode (cf. Figure 11.3). A coplanar electrode arrangement avoids the need to open the chamber during the measurement, and also helps to maintain physiological, nonharmful conditions. At first sight, one would suspect that sneak currents might flow underneath the cell bodies in parallel to the surface, from one electrode to the other, without passing the cell layer. However, as the cells are rather close to the surface with only a nanometer-sized, electrolyte-filled cleft between lower cell membrane and electrode surface, this suspected 'sneak pathway' provides a much higher resistance than the current pathway across the cell layer, and is not relevant.
- The area of cell layer under examination largely determines the sensitivity of the measurement. The smaller the electrode (and cell layer), the more sensitive is the readout for the cellular parameters. Whereas, the filter set-up is difficult to miniaturize down to the single cell level, this is technically not problematic for thin-film electrodes. Circular electrodes with diameters of 20 µm (roughly the diameter of an animal cell) can be produced using standard photolithography techniques.

- As the electrodes can be manufactured using standard photolithography procedures, they can be customized for individual assays or experimental needs. Moreover, several electrodes can be arranged to a multiwell format on a common substrate so that several experiments can be performed in parallel. Nowadays, 96-well devices are commercially available.
- Noble metal electrodes in contact with an ionic solution behave, electrically, almost like an ideal capacitor. Small deviations, such as a phase angle smaller than the expected 90°, have been often observed and reported but are not of major importance. When the cells attach and spread on an almost perfectly capacitive electrode, the impedance contribution arising from the thin cleft between the cell membrane and the electrode surface is dependent on the frequency. This frequency dependency of the impedance arising at the site of cell–matrix junctions allows discrimination to be made between this particular impedance and the contribution arising from cell–cell junctions. The latter (most notably the tight junctions) are entirely resistive with respect to their electrical properties and, thus, are frequency-independent. This difference is decisive when assigning cell–cell and cell–matrix junctions their individual impedances. In contrast, when the cells are attached to a porous filter membrane which behaves electrically like a small resistor, the impedance contributions from cell–matrix and cell–cell junctions are both frequency-independent and can no longer be separated. Thus, the use of capacitive thin-film electrodes provides a significant analytical advantage over measurements on porous membranes. With the latter, the impedance contributions of cell–matrix and cell–cell junctions cannot be specified individually but only in combination.

The impedance experiments reported in the following sections were all conducted by using circular thin-film electrodes with diameters between 250 μm and 6 mm. All electrodes were prepared from 100 nm-thick gold or indium-tin oxide films, with the cells being grown directly on the electrode surfaces after an adhesive protein film had been established. The electrodes were held inside a cell culture incubator with a humidified atmosphere, 37 °C and 5% (v/v) CO_2 (the use of CO_2 was necessary to establish a stationary pH inside the bulk electrolyte).

11.4.1.3 **Experimental Set-Up**

A block diagram of the experimental ECIS set-up, as used for experiments described in the following sections, is shown in Figure 11.4. Here, five working electrodes are arranged side-by-side with one common counter electrode, such that five individual measurements can be performed. The electrode array is placed inside an ordinary cell culture incubator. Outside the incubator, a computer-controlled relay switch allows each of the individual working electrodes to be addressed. The impedance analyzer records the data, while the external function generator is an optional device that can be used to manipulate the cells on the electrode surface by invasive electric fields (electroporation, wounding). The impedance data are acquired in continuous-wave mode using noninvasive sinusoidal voltages of 10 mV_{rms} amplitude. The frequency range for analysis depends on the cell type under study, the surface area of the electrode, and the required time

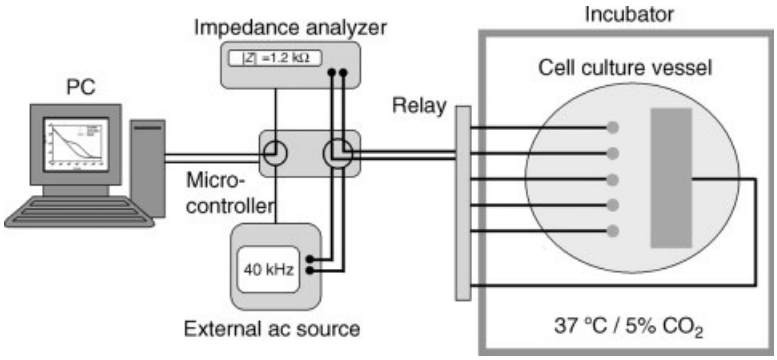


Figure 11.4 Experimental set-up to perform impedance analysis of animal cells grown on the surface of thin-film electrodes such as gold or indium-tin oxide. One or more working electrodes and a common counter electrode are deposited on the bottom of a cell culture dish. The electrode-containing dish is placed inside an ordinary cell-culture incubator to provide physiological conditions. A computer-controlled relay is used to address individual electrodes, the impedance analyzer records the frequency-dependent impedance of the system, and an external ac source may be used for invasive manipulations of the cells on the electrode.

resolution. Full spectral information is typically provided when the impedance is scanned from 1 Hz to 1 MHz.

11.4.2
Time-Resolved Impedance Measurements at Designated Frequencies

The information content of impedance measurements over a range of frequencies during the establishment of a cell monolayer starting from a suspension of single cells is best demonstrated by a 3-D presentation. To account for the complex nature of the impedance Z, the choice can be made to study the impedance magnitude $|Z|$ and the phase angle φ or resistance and reactance or any other representation of the real and imaginary components, as listed above. In the literature about impedance measurements on cultured animal cells by means of thin-film electrodes, one presentation has evolved that provides certain practical advantages [56]. Here, the complex impedance is decomposed in resistance $R = \mathrm{Re}(Z)$ and reactance $X = \mathrm{Im}(Z)$, followed by transformation of the reactance into an equivalent capacitance according to $X = (2\pi \cdot f \cdot C)^{-1}$. In other words, the measured impedance is interpreted as if it had been recorded for a system that behaves like a resistor and a capacitor in series, with the resistance and capacitance of the elements being frequency-dependent. In order to avoid confusion between the parameters of the entire sample and of individual parts of the sample, the total impedance, resistance and capacitance of the entire sample are from now on represented by Z, R and C without subscripts, whereas subordinate components are labeled with appropriate indices.

The time course of the total resistance R and total capacitance C as a function of frequency, when a suspension of epithelial MDCK cells is seeded on the electrodes at

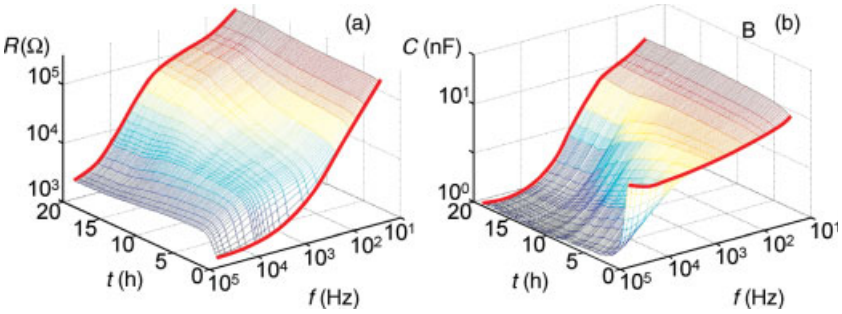


Figure 11.5 Total resistance *R* (a) and capacitance *C* (b) of a gold-film electrode ($d = 250\,\mu m$) when a suspension of MDCK cells is seeded on the electrodes at time zero. The seeding density was sufficiently high that no cell division was needed to establish a confluent cell layer. The red lines highlight the resistance and capacitance spectra for a cell-free and a fully cell-covered electrode at the beginning or end of the experiment, respectively.

time zero, is shown in Figure 11.5a and b [57]. The density of the cells was adjusted such that all adhesion sites on the surface were occupied by the settling cells, without any further cell division. At time zero, the resistance and capacitance spectra corresponded to those of an empty electrode (area $= 5 \times 10^{-4}\,cm^2$), whereas the spectra at the very end of the observation time (20 h) corresponded to the spectra of fully established cell layers with all cell junctions being expressed. (Note that the resistance and capacitance are plotted on a logarithmic scale.) When examining the resistance plot, it is easy to recognize that the time course of *R* is strongly dependent on the monitoring frequency. Whereas, the resistance increases immediately after cell seeding at high frequencies, it follows a biphasic pattern at intermediate frequencies and a retarded increase at the low-frequency end. The data for the total capacitance shows a less involved (but somewhat similar) pattern, with the largest and immediate changes at the high-frequency end and only very minor changes at the low-frequency end. An overlay of the resistance and capacitance spectra of a cell-free (filled symbols) and a cell-covered electrode (open symbols) as extracted from the 3-D profiles, is shown in Figure 11.6a and b, respectively.

The question is how such frequency-dependent resistance and capacitance characteristics arise. The answer to this point is shown in Figure 11.7, where the contributions to the total resistance/capacitance of a cell-free electrode can be described by an equivalent circuit of just two elements (Figure 11.7a): (i) the impedance of the electrode/electrolyte interface (hereafter referred to as the electrode impedance (Z_{el}); and (ii) the constriction resistance (R_{constr}) in series. As discussed above, the impedance of the electrode Z_{el} cannot be modeled accurately by an ideal capacitor, and is therefore represented by the complex impedance Z_{el}. The constriction resistance arises from constricting the electric field from the extended bulk phase down to the size of the electrode. This scales with $1/r$, where r is the radius of the electrode. Although the resistance of the bulk electrolyte is an inherent part of

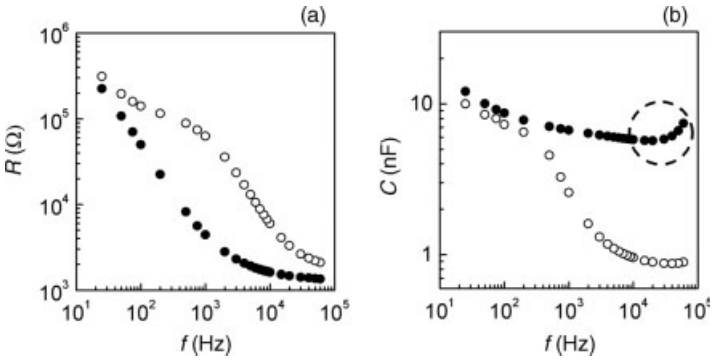


Figure 11.6 Direct overlay of the frequency-dependent resistance (a) and capacitance (b) spectra as recorded for a circular gold-film electrode with a diameter of 250 µm before (filled symbols) and after (open symbols) a monolayer of MDCK cells is established on the electrode surface. The presented spectra have been sliced out of the 3-D presentation shown in Figure 11.5 at times $t = 0$ h and $t = 20$ h. The circle in panel (b) indicates a parasitic increase of the measured capacitance which is due to cables and wiring.

R_{constr}, for the electrode sizes studied here the current constriction dominates this parameter. When referring to the spectra shown in Figure 11.6, the electrode impedance Z_{el} determines the total resistance of a cell-free electrode (filled symbols) at low frequencies, whereas R_{constr} dominates at high frequencies. The total capaci-

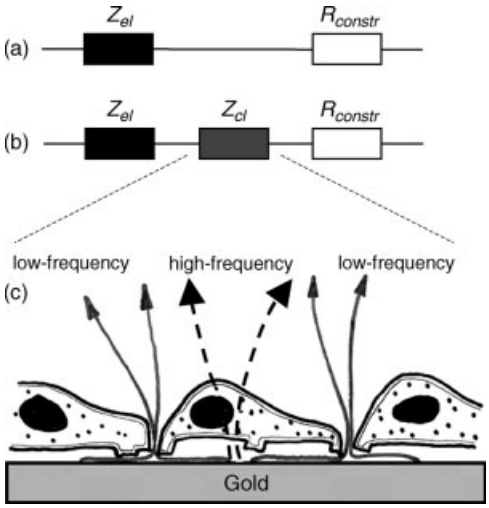


Figure 11.7 Equivalent circuit presentation of a cell-free (a) and a cell-covered (b) gold-film electrode. The cell-free electrode is completely described by a series combination of resistor R_{constr}, representing both, the constriction resistance due to the finite size of the electrode and the resistance of the bulk electrolyte, as well as the impedance of the electrode/electrolyte interface, Z_{el}. The impedance contributions of the cell layer, Z_{cl}, arise from the transcellular current pathway through the cells (broken arrows) and the paracellular current pathway around the cell bodies (solid arrows), as sketched in panel (c).

tance of the cell-free electrode arises solely from Z_{el}. It should be noted that the finite slope of the capacitance spectrum indicates that the electrode is not a perfect capacitor.

For a cell-covered electrode the impedance contribution associated with the presence of cell bodies on the electrode surface Z_{cell} must be added in series to the equivalent circuit of the cell-free electrode discussed above (see Figure 11.7b). The impact of Z_{cell} on the total resistance and capacitance is made very clear by comparing the corresponding spectra for a cell-free and a cell-covered electrode, as provided in Figure 11.6. In order to help understand the origin and contributions to Z_{cell}, Figure 11.7c illustrates the frequency-dependent current flow across the cell layer that can be explained in a two-case approach:

- At high frequencies ($f > 10$ kHz), the current can flow as a displacement current across the membranes, straight through the cell bodies (dashed line) as the plasma membranes behave electrically like capacitors. This is evident from the step-like drop in the spectrum of total capacitance towards higher frequencies (Figure 11.6b). Thus, at these high frequencies C holds information on the capacitance of the plasma membranes C_m, and can be used to monitor its changes during experimental challenges. However, there is another important aspect to this. As the drop in total capacitance at high frequencies is caused by the presence of the plasma membranes on the electrodes, the readings of total capacitance at these high frequencies can be used to determine the electrode coverage. As will be shown below, this relationship between changes in high-frequency capacitance and electrode surface coverage is extremely useful when monitoring cell attachment and spreading to the electrode surface, and in turn the formation of cell–matrix junctions [57].
- When the frequency of the ac signal is lowered well below 10 kHz, the plasma membranes become blocked due to their capacitive nature and the current must flow around the cell bodies in order to escape into the bulk electrolyte (straight lines). On bypassing the cell body, the current must pass through the narrow cleft between the lower cell membrane and the electrode surface, and further on through the cell–cell junctions. Thus, the signal holds information on the electrical properties of cell–matrix and cell–cell junctions, and so can be used to monitor these particular structures.

11.4.2.1 *De novo* Formation of Cell–Matrix and Cell–Cell Junctions

From theoretical considerations (as detailed above) and a set of validation experiments, we have learned that capacitance readings at high frequencies (ca. 10 kHz for the electrodes used here) are best for monitoring the *de novo* formation of cell–matrix contacts during attachment and spreading of initially suspended cells, whereas the resistance at frequencies below 1 kHz (usually 400 Hz for the electrodes used here) is best suited for reporting on the formation of cell junctions, in particular barrier forming tight junctions. A typical readout of the resistance R at 400 Hz and the capacitance C at 40 kHz during the establishment of a mature cell layer is summarized in Figure 11.8a and b for a duplicate experiment. Here, a suspension of epithelial

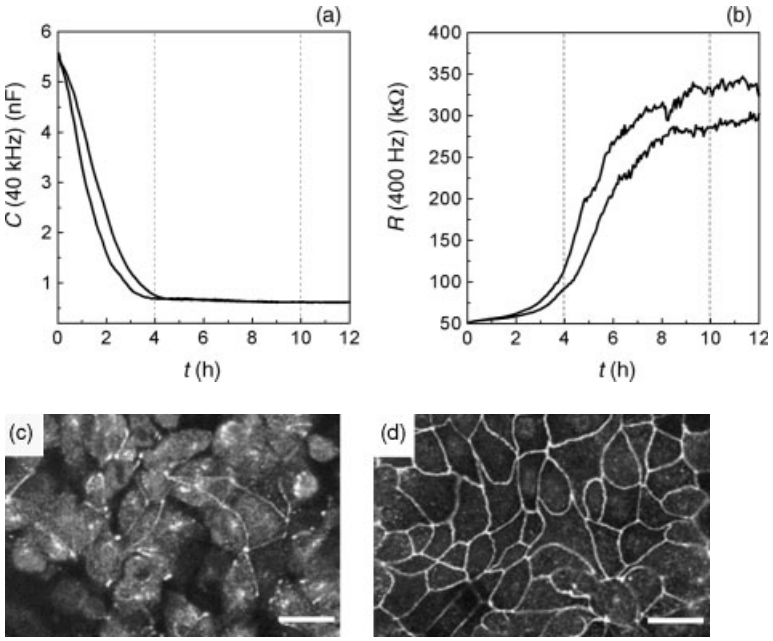


Figure 11.8 Time course of the total capacitance *C* at a monitoring frequency of 40 kHz (a) and the total resistance *R* at 400 Hz (b) during attachment, spreading and differentiation of initially suspended MDCK cells. The number of cells seeded on the electrodes was sufficiently high to form a confluent monolayer without cell division. The vertical dashed lines indicate the times at which immunolabeling of duplicate cultures was performed that are shown in panels (c) and (d). The fluorescence micrographs show the MDCK cell monolayer at the indicated times of 4 h (c) or 10 h (d) after seeding on the growth substrate. The cells were stained for the tight junction-associated protein ZO-1. Scale bar = 25 µm.

MDCK cells has been seeded on the electrode at time zero. The initial cell density (500 000 cm^{-2}) was sufficiently high so that a complete monolayer of cells could form without any further cell division. (MDCK cells are a well-known epithelial cell line that is used extensively worldwide to study tight junctions.) The capacitance drop presented in Figure 11.8 a reports to the attachment and spreading of cells on the electrode surface, which began immediately after the cell suspension had been introduced into the electrode-containing chamber. The cells had been suspended in a serum-containing medium, which led to the spontaneous and immediate adsorption of adhesive proteins from the serum onto the electrode surface. The time course of the total capacitance indicated that the *de novo* formation of cell–matrix junctions was complete, and that the surface was entirely covered with spread cells within 4 h. At approximately the time when cell adhesion was complete, the resistance at 400 Hz began to increase considerably for the next 6 h (Figure 11.8b), indicating the formation of barrier forming tight junctions between adjacent cells. Thus, from a cell biology viewpoint, the formation of tight junctions requires fully established cell–matrix contacts. The minor increase in resistance observed during the initial 4 h of the experiment was

caused by the constriction of current flow underneath the cells during attachment and spreading, and was therefore due to the formation of cell–matrix junctions. However, as the drop in capacitance C at 40 kHz correlates linearly with the fractional surface coverage of the electrode, C is to be the more useful parameter to monitor.

In order to back up these impedance data by microscopic studies, two aliquots of the same cell suspension were seeded on ordinary microscope slides and immunostained for the tight junction-associated proteins ZO-1. The fluorescence micrographs in Figure 11.8c and d provide typical snapshots of the distribution of ZO-1 in MDCK cells at 4 h and 10 h after cell inoculation, respectively. Whereas, after 4 h only very few cell–cell contact sites showed a positive staining for ZO-1, the protein was completely allocated to the cell borders after 10 h, consistent with the time course of R at a monitoring frequency of 400 Hz. Taken together, the results of these experiments show that it is possible to monitor the formation of cell–matrix and cell–cell contacts in one and the same cell population by utilizing noninvasive impedance measurements in real time, and that the impedance readings are in line with microscopic studies.

Additional experimental support for the claim that spreading of cells and formation of cell–cell junctions can be monitored individually simply by examining different ac frequencies is provided by an experiment in which, again, MDCK cells were seeded on electrodes that had been precoated with an adhesive protein (fibronectin). In this experiment, one population of cells was suspended in a buffer that contained 1 mM Ca^{2+} as the only divalent cation, while a second population was suspended in buffer containing 1 mM Mg^{2+} as the only divalent cation. The rationale of this approach was that the formation of cell–matrix junctions is dependent on the presence of either Ca^{2+} or Mg^{2+} (or Mn^{2+}), whereas tight junction formation is known to depend on the presence of Ca^{2+}, which cannot be substituted by a different cation. Figure 11.9a shows the time course of the capacitance C at 40 kHz for

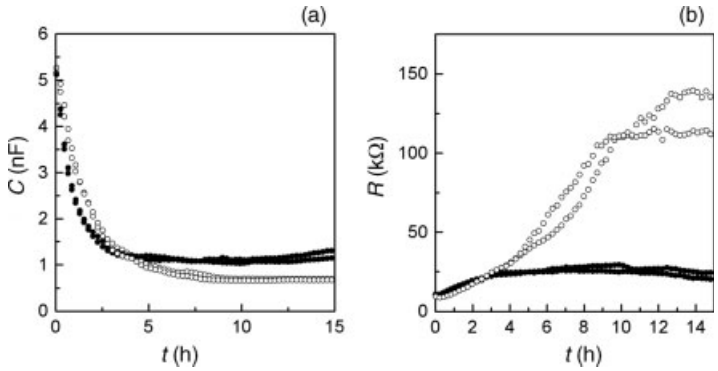


Figure 11.9 Time course of the total capacitance C at a monitoring frequency of 40 kHz (a) and the total resistance R at 400 Hz (b) during attachment, spreading and differentiation of initially suspended MDCK cells in different salt solutions. Cells were suspended in a buffer containing either 1 mM Mg^{2+} (filled symbols) or 1 mM Ca^{2+} (open symbols) as the only divalent cation.

duplicates of both conditions, while Figure 11.9b shows the resistance R at 400 Hz. In the presence of Mg^{2+}, the cells attached and spread slightly faster than in the presence of Ca^{2+}, whereas only in the presence of Ca^{2+} the resistance at 400 Hz increased significantly, indicating that tight junctions were formed only under the latter conditions. Although, from a cell biology viewpoint these experiments did not provide any new insight, they did demonstrate the capability of impedance analysis for monitoring cell junctions noninvasively and in real time. Recently, it was shown that this technology is well suited to unravel the different spreading rates of MDCK cells on various different protein coatings on the electrode surface, with unprecedented time resolution [57].

11.4.2.2 Modulation of Established Cell Junctions

In the experiments described above impedance measurements were used to follow the *de novo* formation of cell junctions when suspended cells were seeded onto the measuring electrodes. Another – perhaps more often conducted – experiment addresses the modulation of already established cell junctions within a confluent monolayer of cells upon exposure to a given stimulus. These stimuli can be biological (bacteria, viruses, cancer cells), physical (electric fields, mechanic shear) or chemical (toxins, drugs) in nature.

As an example of such an experiment, Figure 11.10 traces the time courses of resistance R at 100 Hz and capacitance C at 100 kHz when a confluent layer of epithelial MDCK cells is exposed to a 5 µM solution of Cytochalasin D (CD). CD is a fungal toxin that interferes with the actin cytoskeleton of animal cells. Actin filaments

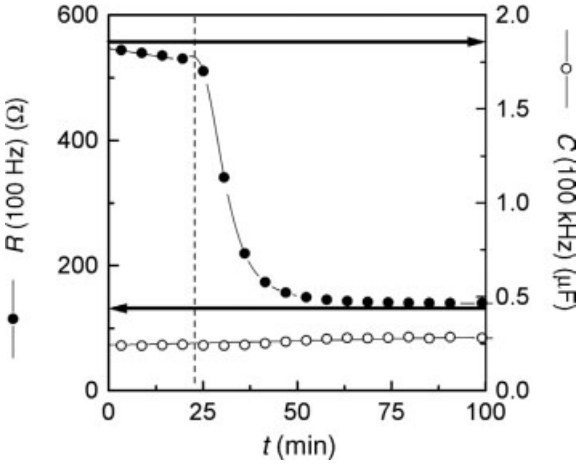


Figure 11.10 Time course of the total resistance R at a monitoring frequency of 100 Hz and total capacitance C at a frequency of 100 kHz before and during a confluent monolayer of MDCK cells was exposed to 5 µM Cytochalasin D (CD). The vertical dashed line indicates the time of CD addition to the cell population; horizontal arrows indicate the corresponding resistance or capacitance value of a cell-free electrode. The arrowheads point to the *y*-axis to which the arrow refers. Note: this experiment was performed with a working electrode with a surface area of 0.33 cm^2.

(f-actin) are polymers made from globular actin monomers (g-actin), and are highly dynamic structures that grow continuously at one end (polymerization) while shrinking at the other end (de-polymerization), depending on the requirements of the cells in a given physiological situation. Although CD interferes rigorously with actin polymerization, depolymerization is left unaffected so that the actin filaments become disassembled with time. With respect to the impedance readout, exposure to CD leads to a striking drop in the resistance at 100 Hz, whereas the capacitance changes only marginally. In Figure 11.10, the horizontal arrows indicate the values of R and C for the cell-free electrode, with the arrowheads pointing at the axis to which they refer. When translated into a molecular perspective, these time courses reveal that the cell–cell junctions must have completely disassembled during the exposure to CD, whereas the cells remain attached to the surface. Even though cell–matrix junctions might be compromised by the fungal toxin, they can still withstand the inner tension of the cell body and prevent cells from rolling up and losing their matrix anchorage.

The same experiment has been followed using fluorescence microscopy in order to correlate both readouts, comparing a layer of MDCK cells before and 50 min after exposure to 5 μM CD (see Figure 11.11a and b, respectively). For both images, the cells were stained with fluorescently labeled phalloidin. (Phalloidin binds to filamentous but not globular actin, and is therefore the classical stain for studying the actin cytoskeleton when using fluorescence microscopy.) After exposure to CD, the actin filaments visible inside the cell bodies of the control had entirely disappeared, as had the belt of actin filaments that ran around the periphery of the MDCK cells. Only actin aggregates were left in the CD-treated cells which, however, were still anchored to their growth substrate. Fluorescence micrographs of MDCK cells exposed to either 5 μM CD for 50 min or control conditions are shown in Figure 11.11c and d, respectively, although now the tight junction-associated protein ZO-1 is labeled with fluorescent antibodies. Under control conditions the junctional staining completely circumscribed the periphery of the individual cells. In contrast, after CD exposure the junctional staining was punctuated and discontinuous, indicating a loss of junctional tightness. To summarize, these microscopic studies on molecules that contribute to the cellular junctions are perfectly in line with the impedance readings on the time course of junctional disintegration, and their interpretation.

At this point it should be noted that monitoring the capacitance at high frequency only reports on electrode coverage. As long as the *de novo* formation of cell-matrix junctions is considered, coverage of the electrode is indicative of the formation of cell–matrix junctions. However, in an established cell monolayer, cell–matrix junctions may be moderately altered, without causing complete removal of the cell layer from the electrode surface. Under these conditions capacitance readings are insensitive, but improvements can be achieved if the impedance data are recorded over an extended frequency range (instead of only one or two designated frequencies), and when the impedance spectrum of a completely cell-covered electrode is modeled in detail. If the experiment is conducted in this way, even minor changes in the cell–matrix adhesion sites can be identified and followed with time. The model used to analyze the impedance spectra of established cell monolayers is described in the following section.

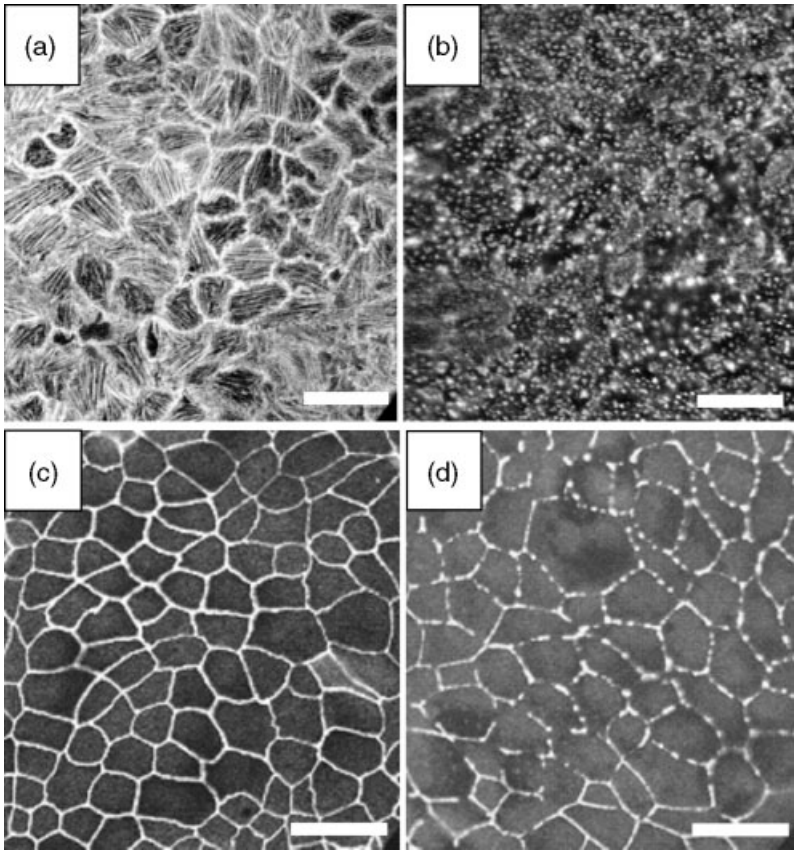


Figure 11.11 Micrographs of confluent MDCK cell layers with fluorescently labeled actin cytoskeleton (a, b) or with an immunofluorescence tag addressing the tight junction-associated protein ZO-1 (c, d). Panels (a) and (c) show control populations; cell layers in panels (b) and (d) were exposed to 5 μ*M* Cytochalasin D for 50 min. Scale bars = 25 μm.

11.4.3
Modeling the Complex Impedance of Cell-Covered Electrodes

In order to interpret the impedance spectra of cell-covered film-electrodes, a physical model has been developed that accounts for the impedance contributions associated with the presence of cells on the electrode surface, as described qualitatively above [58, 59]. In this model (Figure 11.12), the cells are treated as disk-shaped objects with a radius r_c, and are considered to hover at an average distance h above the electrode surface. Hence, two current pathways can be considered:

- At low ac frequencies the current will mainly flow from the electrode through the thin fluid-filled cleft between the cell and the electrode, and leave the cell sheet

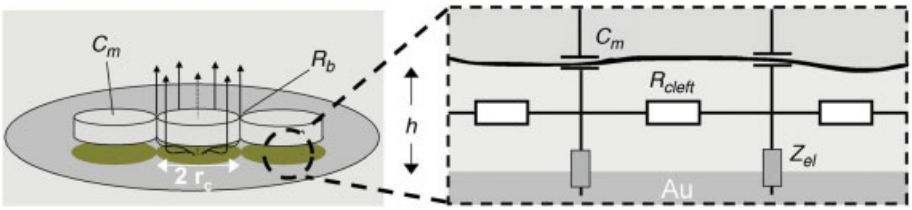


Figure 11.12 Sketch illustrating the physical model used to analyze and predict the complex impedance of cell-covered gold-film electrodes. The resistance R_b represents the resistance of the current pathway between two adjacent cells through the cell–cell junctions; the resistance R_{cleft} represents the resistance of the aqueous cleft underneath the cells at the site of cell–matrix junctions. The capacitance of the plasma membranes is accounted for by C_m. In the model, the cell bodies are treated as cylindrical disks with radius r_c that are separated from the electrode surface by distance h.

through the cell–cell junctions between adjacent cells (straight arrows). The model parameter R_{cleft} accounts for the impedance arising at the site of cell–matrix junctions, whereas R_b represents the resistive properties of the cell–cell junctions.

- At high frequency, the cell membrane, which is modeled as a capacitor C_m, allows a displacement current to pass through the cells (broken arrow). The resistive component of the plasma membrane impedance due to the presence of ion channels is neglected in the calculation as it is significantly higher than the paracellular resistance in most cases.

At intermediate frequencies the current makes use of both pathways and splits up in frequency and position-dependent ratios. Thus, the cell–electrode junction behaves like a 2-D core-coat conductor, with the inner electrolyte core and plasma membrane as well as the electrode surface as outer dielectric coats. Solving the 2-D cable equation with proper boundary conditions provides the following transfer function for the impedance of a cell-covered electrode Z_{cell}:

$$\frac{1}{Z_{cell}} = \frac{1}{Z_{el}}\left(\frac{Z_{el}}{Z_{el}+Z_m} + \frac{\frac{Z_m}{Z_{el}+Z_m}}{\frac{\gamma r_c}{2}\cdot\frac{I_0(\gamma r_c)}{I_1(\gamma r_c)} + R_b\left(\frac{1}{Z_{el}}+\frac{1}{Z_m}\right)}\right) \tag{11.5}$$

$$\text{with} \quad \gamma r_c = r_c \cdot \sqrt{\frac{\rho}{h}\left(\frac{1}{Z_{el}}+\frac{1}{Z_m}\right)} = \alpha \cdot \sqrt{\left(\frac{1}{Z_{el}}+\frac{1}{Z_m}\right)} \tag{11.6}$$

$$\text{and} \quad \alpha^2 = R_{\text{cleft}}$$

Here, I_0 and I_1 are modified Bessel functions of the first type of order 0 and 1, R_b is the resistance of cell–cell junctions, ρ is the specific electrolyte resistance inside the cell–electrode junction, and Z_m accounts for the impedance of the transcellular high-frequency pathway across both, the upper and lower membranes $Z_m = 2/(\omega \cdot C_m)$. Finally, the impedance of a cell-free electrode Z_{el} is most accurately modeled by a CPE

behavior, thus $Z_{el} = {}^{1}/_{A(i\omega)^n}$ with the parameter n ($0 \leq n \leq 1$) as an indicator for the deviation from ideal capacitive behavior, with $n = 1$ [60].

Fitting the above-described model to experimental impedance spectra provides three parameters to describe the changes that occur within the cell layer: R_{cleft} describes alterations at the site of cell–matrix adhesion; R_b describes changes within cell–cell junctions; and C_m reports on changes of the membrane capacitance.

The nonredundant model outlined above has been extended to represent the cellular structure with respect to geometry and electrical structure. More accurately for instance, the above model for disk-like cells has been adapted for elliptical cells, however, the long and small axes of the cellular ellipsoid must be determined independently from microscopic images [61]. Another approach accounts individually for the capacitance of the apical, basal and lateral membrane [62]. However, this latter model introduces two more parameters into the transfer function that cannot be determined independently, thus making it redundant. As can be seen from the microscopic images presented in Figures 11.8 and 11.11, the circular approximation is well justified for MDCK cells.

11.4.4
Spectroscopic Characterization of Cell–Cell and Cell–Matrix Junctions

If a cell layer is studied by continuously recording impedance spectra over a sufficiently extended frequency range during an experimental challenge, the analysis of these spectra with the above-described model will provide the time course of the model parameters R_b, R_{cleft} and C_m, which allows the cell response to be characterized in more detail. For demonstration, the data of MDCK cells challenged with 5 μM CD are reconsidered. The cell response to the CD challenge was monitored by repeatedly recording the impedance spectra, from 1 Hz to 1 MHz. In Figure 11.10, only the time courses of resistance at 100 Hz and capacitance at 100 kHz, respectively, are presented. An analysis of the complete full-width impedance spectra provides the time courses of R_b, R_{cleft} and C_m, but such modeling is only meaningful if the cell layer is still confluent, without significant defects. When the MDCK cells were exposed to 5 μM CD the membrane capacitance did not alter significantly during the time of the study. The changes in R_b and R_{cleft} are shown in Figure 11.13. As already deduced from the resistance readings at 100 Hz, the epithelial barrier function disappeared completely on exposure to CD. R_b fell from almost 100 $\Omega \cdot cm^2$ to values not significantly different from 0. Although the capacitance readings at 100 kHz correctly indicated that the cells had not detached from the electrode surface, there is nevertheless a drastic change in cell–matrix adhesion. The values for the R_{cleft} fall, from 250 to 25 $\Omega \cdot cm^2$. Thus, whilst the matrix anchorage is still sufficiently strong to keep the cells attached to the surface, there has been a considerable reorientation of the cell–matrix adhesion sites due to the activity of CD. According to the definition of R_{cleft}, which is

$$R_{cleft} = r_c^2 \frac{\rho}{h} \tag{11.7}$$

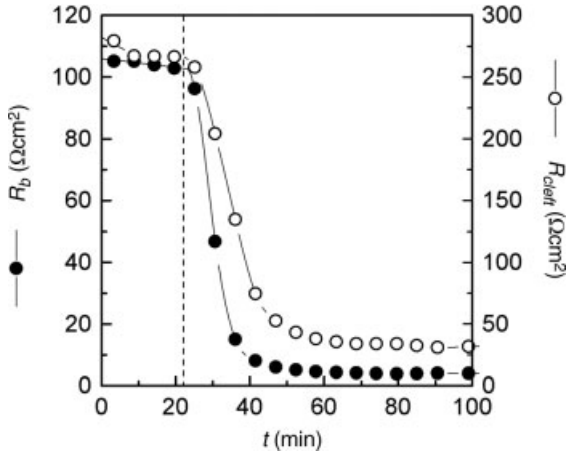


Figure 11.13 Time course of the model parameters R_b and R_{cleft} before and during a confluent monolayer of MDCK cells being exposed to 5 µ*M* Cytochalasin D. The model parameters were extracted from complete impedance spectra recorded over six frequency decades from 1 to 10^6 Hz. The membrane capacitance C_m (data not included) did not change during the course of the experiment.

the observed changes may be due to changes in r_c, ρ or h. As the cell radius cannot be changed very much in a continuous monolayer, the changes must arise from either changes in the specific electrolyte resistance ρ in the cell–electrode junction, or from an increase in the average distance between the lower cell membrane and electrode surface h, or both. This ambiguity cannot be resolved unless one of these two quantities is measured independently. If it is assumed that all changes in R_{cleft} are due to an increase in the average cell–substrate separation distance h (which has been determined as 25 nm under control conditions, using FLIC microscopy), then the observed changes in R_{cleft} translate into a change of h from 25 to 250 nm. However, this seems unlikely, and indicates that both, h and ρ may have changed. This example illustrates that, despite not being an imaging technique, impedance analysis is still sufficiently sensitive to report changes on the nanoscale, even at hidden interfaces such as the cell–electrode junction that are not accessible by scanning probe techniques.

Taken together, the impedance analysis of cell-covered film electrodes provides functional information concerning cell–cell and cell–matrix junctions in a time-resolved and noninvasive, but integral and nonimaging, manner. An approximate calculation using the above-described model, plus realistic estimates for the resolution of each model parameter, reveals that impedance analysis is more sensitive than light microscopy and capable of identifying differences on the nanometer scale. Moreover, the data can be recorded without opening the incubator door and with multiple samples in parallel, both of which are important considerations for screening processes.

Acknowledgments

The author would like to acknowledge the Kurt-Eberhard Bode foundation for generous financial support and Dr. R. Hütterer for careful proofreading.

References

1 Wegener, J. (2002) *Encyclopedia of Life Sciences*, Nature Publishing Group.
2 Alberts, B.A., Bray, D., Lewis, J., Raff, M. Roberts, K. and Watson, J.D. (1995) *Molecular Biology of the Cell*, Wiley Sons Ltd, New York.
3 Hartsock, A. and Nelson, W.J. (2007) *Biochimica et Biophysica Acta*, **1778** (3), 660.
4 Green, K.J. and Simpson, C.L. (2007) *The Journal of Investigative Dermatology*, **127**, 2499.
5 Green, K.J. and Gaudry, C.A. (2000) *Nature Reviews Molecular Cell Biology*, **1**, 208.
6 Insall, R. and Machesky, L. (2001) *Encyclopedia of Life Sciences*, John Wiley & Sons Ltd, Chichester.
7 Lo, S.H. (2006) *Developmental Biology*, **294**, 280.
8 Stupack, D.G. (2007) *Oncology (Williston Park)*, **21**, 6.
9 Chan, K.T., Cortesio, C.L. and Huttenlocher, A. (2007) *Methods in Enzymology*, **426**, 47.
10 Cereijido, M., Shoshani, L. and Contreras, R.G. (2000) *American Journal of Physiology - Gastrointestinal and Liver Physiology*, **279**, G477.
11 Powell, D.W. (1981) *The American Journal of Physiology*, **241**, G275.
12 Grebenkämper, K.G. and Galla, H.-J. (1994) *Chemistry and Physics of Lipids*, **71**, 133.
13 Niessen, C.M. (2007) *The Journal of Investigative Dermatology*, **127**, 2525.
14 Aijaz, S., Balda, M.S. and Matter, K. (2006) *International Review of Cytology*, **248**, 261.
15 Balda, M.S. and Matter, K. (1998) *Journal of Cell Science*, **111(**Pt 5), 541.
16 Matter, K., Aijaz, S., Tsapara, A. and Balda, M.S. (2005) *Current Opinion in Cell Biology*, **17**, 453.
17 Galla, H.J. and Wegener, J. (1996) *Chemistry and Physics of Lipids*, **81**, 339.
18 Garner, C.M. and Nash, J. (2001) Chemical synapses, in *Encyclopedia of Life Sciences*, John Wiley & Sons, Ltd, Chichester, doi: 10.1038/npg. els.0000037.
19 Schweizer, F.E. (2001) Synapses, in *Encyclopedia of Life Sciences*, John Wiley & Sons, Ltd, Chichester, doi: 0.1038/npg. els.0000207.
20 Evans, W.H. and Martin, P.E. (2002) *Molecular Membrane Biology*, **19**, 121.
21 Kumar, N.M. and Gilula, N.B. (1996) *Cell*, **84**, 381.
22 Wegener, J. (2005) *Encyclopedia of Biomedical Engineering* (ed. M. Akay), Wiley & Sons, Hobooken, NJ.
23 Martinez, E., Engel, E., Lopez-Iglesias, C., Mills, C.A., Planell, J.A. and Samitier, J. (2007) *Micron*, **39** (2), 111.
24 Wrobel, G., Holler, M., Ingebrandt, S., Dieluweit, S., Sommerhage, F., Bochem, H.P. and Offenhausser, A. (2007) *Journal of the Royal Society Interface*, **5** (19), 213.
25 Gingell, D. and Todd, I. (1979) *Biophysical Journal*, **26**, 507.
26 Verschueren, H. (1985) *Journal of Cell Science*, **75**, 279.
27 Braun, D. and Fromherz, P. (1997) *Applied Physics A*, **65**, 341.
28 Braun, D. and Fromherz, P. (1998) *Physical Review Letters*, **81**, 5241.
29 Lambacher, A. and Fromherz, P. (1996) *Applied Physics A*, **63**, 207.
30 Truskey, G.A., Burmeister, J.S., Grapa, E. and Reichert, W.M. (1992) *Journal of Cell Science*, **103.2**, 491.

31 Geggier, P. and Fuhr, G. (1999) *Applied Physics A: Materials Science and Processing*, **68**, 505.

32 Wegener, J., Janshoff, A. and Steinem, C. (2001) *Cell Biochemistry and Biophysics*, **34**, 121.

33 Heitmann, V., Reiss, B. and Wegener, J. (2006) *Piezoelectric Sensors* (eds C. Steinem and A. Janshoff), Springer, Berlin.

34 Thoumine, O., Ott, A. and Louvard, D. (1996) *Cell Motility and the Cytoskeleton*, **33**, 276.

35 Sagvolden, G., Giaever, I., Pettersen, E.O. and Feder, J. (1999) *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 471.

36 Braun, D. and Fromherz, P. (2004) *Biophysical Journal*, **87**, 1351.

37 Lohmann, C., Huwel, S. and Galla, H.J. (2002) *Journal of Drug Targeting*, **10**, 263.

38 Zink, S., Rosen, P. and Lemoine, H. (1995) *The American Journal of Physiology*, **269**, C1209.

39 Matter, K. and Balda, M.S. (2003) *Methods*, **30**, 228.

40 Wegener, J., Abrams, D., Willenbrink, W., Galla, H.J. and Janshoff, A. (2004) *Biotechniques*, **37**, 590.

41 Cereijido, M., Gonzalez-Mariscal, L. and Borboa, L. (1983) *The Journal of Experimental Biology*, **106**, 205.

42 Gitter, A.H., Bertog, M., Schulzke, J. and Fromm, M. (1997) *Pflügers Archiv: European Journal of Physiology*, **434**, 830.

43 Giocondi, M.C. and Le Grimellec, C. (1989) *Biochemical and Biophysical Research Communications*, **162**, 1004.

44 Korchev, Y.E., Negulyaev, Y.A., Edwards, C.R., Vodyanoy, I. and Lab, M.J. (2000) *Nature Cell Biology*, **2**, 616.

45 Goldberg, G.S., Valiunas, V. and Brink, P.R. (2004) *Biochimica et Biophysica Acta*, **1662**, 96.

46 Williams, K.K. and Watsky, M.A. (1997) *Current Eye Research*, **16**, 445.

47 el-Fouly, M.H., Trosko, J.E. and Chang, C.C. (1987) *Experimental Cell Research*, **168**, 422.

48 Schwan, H. (1993) *Medical Progress Through Technology*, **19**, 163.

49 Fricke, H. and Morse, S. (1926) *Journal of Cancer Research*, **10**, 340.

50 Grimnes, S. and Martinsen, Ø.G. (2000) *Bioimpedance and Bioelectricity Basics*, Academic Press, Cornwall.

51 Kottra, G. and Fromter, E. (1993) *Pflügers Archiv: European Journal of Physiology*, **425**, 535.

52 Erben, M., Decker, S., Franke, H. and Galla, H.J. (1995) *Journal of Biochemical and Biophysical Methods*, **30** (4), 227.

53 Kottra, G. and Fromter, E. (1984) *Pflügers Archiv: European Journal of Physiology*, **402**, 409.

54 Giaever, I. and Keese, C.R. (1993) *Nature*, **366**, 591.

55 Giaever, I. and Keese, C.R. (1984) *Proceedings of the National Academy of Sciences of the United States of America*, **81**, 3761.

56 Giaever, I. and Keese, C.R. (1986) *IEEE Transactions on Biomedical Engineering*, **BME-33**, 242.

57 Wegener, J., Keese, C.R. and Giaever, I. (2000) *Experimental Cell Research*, **259**, 158.

58 Giaever, I. and Keese, C.R. (1991) *Proceedings of the National Academy of Sciences of the United States of America*, **88**, 7896.

59 Arndt, S., Seebach, J., Psathaki, K., Galla, H.J. and Wegener, J. (2004) *Biosensors and Bioelectronics*, **19**, 583.

60 McAdams, E.T., Lackermeier, A., McLaughlin, J.A., Macken, D. and Jossinet, J. (1995) *Biosensors and Bioelectronics*, **10**, 67.

61 Lo, C.-M. and Ferrier, J. (1998) *Physical Review E*, **57**, 6982.

62 Lo, C.M., Keese, C.R. and Giaever, I. (1995) *Biophysical Journal*, **69**, 2800.

1
Top-Down Fabrication of Nanostructures

Ming Liu, Zhuoyu Ji, and Liwei Shang

1.1
Introduction

The "top-down" approach to nanofabrication involves the creation of "nanostructures" from a large parent entity. This type of fabrication is based on a number of tools and methodologies which consist of three major steps:

1) The deposition of thin films/coatings on a substrate.
2) Obtaining the desired shapes via photolithography.
3) Pattern transfer using either a lift-off process or selective etching of the films

Compared with general chemical fabrication and processing methods, top-down fabrication techniques for the creation of nanostructures are derived mainly from the techniques applied for the fabrication of microstructures in the semiconductor industry. In particular, the fundamentals and basic approaches are mostly based on micro-fabrications. In this chapter, methods of top-down nanofabrication will be discussed, with attention being focused primarily on methods of lithography, especially optical, electron-beam, X-ray and focused ion beam lithography. A brief introduction will also be provided on how to create nanostructures using various methods of thin film deposition and etching materials. Finally, the methods for pattern transfer through etching and lift-off techniques will be discussed.

In the past, top-down fabrication techniques have represented an effective approach for nanostructures and, when complemented with bottom-up approaches during the past few decades, have led to amazing progress having been made with a variety of nanostructures. The traditional top-down technology used to create nanostructures and nanopatterns is discussed in the following sections.

1.2
Lithography

Lithography, which is also often referred to as "photoengraving," was invented in 1798 in Germany by Alois Senefelder. It is the process of defining useful shapes on

Nanotechnology, Volume 8: Nanostructured Surfaces. Edited by Lifeng Chi

ISBN: 978-3-527-31739-4

the surface of a semiconductor wafer [1–5]. Typically, it consists of a patterned exposure into some form of photosensitive material that has already been deposited onto the wafer. Many techniques of lithography have been developed during the past fifty years, by using a variety of lens systems and exposure radiation sources that have included photons, X-rays, electrons, ions, and neutral atoms. In spite of the different exposure radiation sources used in the various lithographic methods, and the instrumental details, all of these techniques share the same general approaches and are based on similar fundamentals. *Photolithography* is the most widely used technique in microelectronic fabrication, particularly for the mass production of integrated circuits (ICs) [2], and has been the driving force behind the miniaturization of such circuits since they were first produced at Fairchild and at Texas Instruments during the early 1960s [6].

1.2.1 Photolithography

Photolithography (also called "optical lithography") is simply lithography using a radiation source with wavelength(s) in the visible spectrum. It has served as the dominant patterning technology in the semiconductor industry since the IC was invented almost sixty years ago. From the onset, optical lithography has always managed to keep pace with Moore's law [7, 8], including its recent acceleration. In order to keep pace with the shrinking feature size, a steady stream of improvements in the field of resolution, image placement, and pattern transfer have been introduced time after time, and these have enabled optical lithography to hold off the challenges of the competing lithography technologies.

The key historical events in photolithography have been as follow:

- 1826: Joseph Nicephore Niepce, in Chalon, France, takes the first photograph using bitumen of Judea on a pewter plate, developed using oil of lavender and mineral spirits.
- 1843: William Henry Fox Talbot, in England, develops dichromated gelatin, patented in Britain in 1852.
- 1935: Louis Minsk of Eastman Kodak develops the first synthetic photopolymer, poly(vinyl cinnamate), the basis of the first negative photoresists.
- 1940: Otto Suess of Kalle Division of Hoechst AG, develops the first diazoquinone-based positive photoresist.
- 1954: Louis Plambeck, Jr, of Du Pont, develops the Dycryl polymeric letterpress plate.

Optical lithography is a process used in microfabrication to selectively remove parts of a thin film (or the bulk of a substrate). It involves the use of an optical technique to produce images at smaller scales, which employs light to transfer a geometric pattern from a photomask to a light-sensitive chemical (photoresist, or simply "resist") on the substrate. The steps involved in the photolithographic process include: wafer cleaning; barrier layer formation; photoresist application; soft baking; mask alignment; exposure and development; and hard-baking.

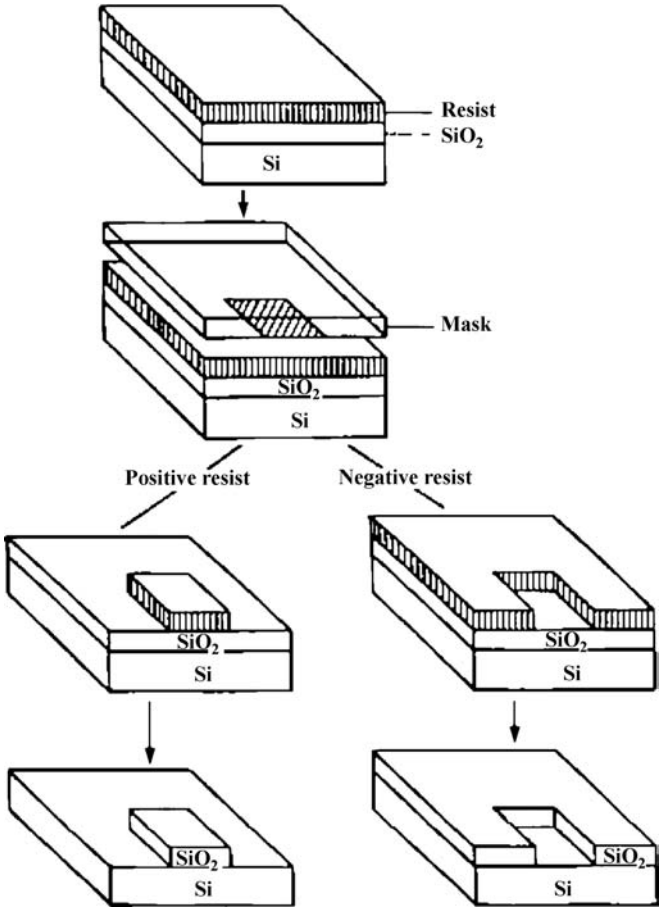


Figure 1.1 Schematic representation of the photolithographic process sequences, in which images in the mask are transferred to the underlying substrate surface.

The basic scheme of photolithography (as shown in Figure 1.1) involves three steps [9]: (i) a thin film of resist material is cast over the substrate; (ii) the substrate is then exposed to a pattern of intense light through a mask, during which time the resist material is selectively struck by the light; (iii) the exposed substrate is then immersed into the development solvent.

Depending on the chemical nature of the resist material, the photoresist is defined as either a positive or a negative type. For positive resists, the resist is exposed to ultraviolet (UV) light wherever the underlying material is to be removed. In these resists, exposure to the UV light alters the chemical structure of the resist, which causes it to become more soluble in the developing solvent than in the unexposed areas. The resist exposed under the UV light is then washed away by the developer solution. Overall, the process can be described as "whatever shows, goes away." In contrast, in the case of a negative resist, the exposed areas may be rendered less

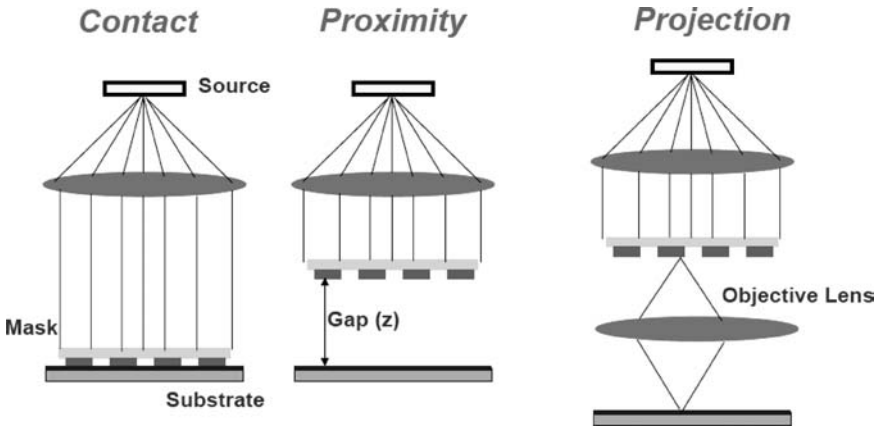


Figure 1.2 Schematic diagram of working modes of photolithography.

soluble in a certain developing solvent; this leads to the production of a negative tone image of the shadow mask, a process described as "whatever shows, stays behind."

In addition to conventional photoresist polymers, Langmuir–Blodgett films and self-assembled monolayers (SAMs) have also been used as resists in photolithography [10, 11]. In such applications, photochemical oxidation, crosslinking, or the generation of reactive groups are used to transfer patterns from the mask to the mono-layer [12, 13].

A *master mask* is necessary in the process of photolithography, and in general this is scribed using an optical method and produced by chemical etching. In this case, the light passes through the mask to define the actual structure in the material; according to the position of the mask with respect to the sample, three types of exposure lithography can be defined, namely contact, proximity, and projection lithography (see Figure 1.2).

Up until to the early 1970s, most lithography was carried out as either a contact or a close-proximity printing process, in which blue and near-UV light was passed through a photomask directly onto a photoresist-coated semiconductor substrate [14]. This apparently simple shadow imaging process has been described in many research reports and handbooks [15, 16].

1.2.1.1 **Contact Printing**

In contact-mode photolithography, the resist-coated silicon wafer is brought into intimate physical contact with the glass photomask. For this, the wafer is held on a vacuum chuck and the whole assembly rises until the wafer and mask make contact with each other. The photoresist is exposed to UV light while the wafer is in contact position with the mask, and this allows a mask pattern to be transferred into a photoresist with almost 100% accuracy, as well as providing the highest resolution (e.g., 1 μm features in 0.5 μm of positive resist). Unfortunately, however, the maximum resolution is seldom achieved owing to the presence of dust on the

substrates and the nonuniform thicknesses of both the photoresist and the substrate. The main problem with contact printing is that debris, trapped between the resist and the mask, can damage the mask and cause defects in the pattern.

1.2.1.2 **Proximity Printing**

The proximity exposure method is similar to contact printing, and involves introducing a gap about 10–25 μm wide between the mask and the wafer during the exposure stage. This gap minimizes (but may not eliminate) the mask damage. Although a resolution of about 2–4 μm is possible with proximity printing, increasing the gap will reduce the resolution by expanding the penumbral region caused by diffraction. The main difficulties associated with proximity printing include the control of a small and very constant space between the mask and wafer, which can be achieved only by using extremely flat wafers and masks.

1.2.1.3 **Projection Printing**

Generally speaking, projection techniques have a lower resolution capability than that provided by shadow printing. However, unlike shadow printing, in projection printing the lens elements are used to focus the mask image onto a wafer substrate, which is separated from the mask by several centimeters so that damage to the mask is entirely avoided. An image of the patterns on the mask is projected onto the resist-coated wafer, which is located several centimeters away. In order to achieve a high resolution, only a small portion of the mask is imaged, and the small image field is scanned or stepped over the surface of the wafer. Projection printers that step the mask image over the wafer surface are termed "step-and-repeat" systems, and are capable of approximately 1 μm resolution.

The first widespread use of projection printing for semiconductor manufacturing was fostered by a very well-accepted family of tools from Perkin-Elmer, the so-called "Micralign" projection aligners developed during the early 1970s. For the first time, these tools allowed a higher performance pattern definition by scanning and imaging only a fractional area of the wafer at any instant. Moreover, the optical resolution and pattern overlay performances were also significantly enhanced. The technique of projection printing was further developed when a new class of projection exposure tools – typically referred to as "steppers" – was introduced during the late 1970s [17]. For the first time, the pattern definition imaging on the semiconductor wafers was performed one chip at a time, in a step-and-repeat fashion. This had profound implications not only on the requirements for the photomask but also on the precision mechanical movements needed to accurately overlay a new pattern on underlying patterns already on the wafer substrate.

1.2.1.4 **Resolution Enhancement Techniques (RETs)**

Various RET approaches have undergone intense investigation during the past few years, and many reports have been made worldwide at leading lithography conferences. Moreover, most of these techniques have been introduced into high-volume wafer manufacture. The RETs interact in many ways, and Smith has discussed in some detail the impact of interacting factors when designing and refining the lithography

Table 1.1 The main categories of resolution enhancement technique in current use and undergoing investigation.

Resolution enhancement technique	Type	Advantage(s)	Disadvantage(s)
Phase-shift masks	Wavefront engineering	Improves DoF and exposure latitude	High mask cost, inspection and repair difficult
Modified illumination	Wavefront engineering	Improves DoF for dense line/space feature	Less improvement for holes or isolated lines affected by lens aberrations
Optical proximity correction (OPC)	Mask engineering	Improved CD control for various patterns	Additional design data processing masks more complex and expensive
Wafer control – antireflective layers	Resist engineering	Improved CD control reduces notching	Increased cost and process complexity may complicate etch
Pupil filtering	Wavefront engineering	Improved CD control and exposure latitude	Pattern-specific capability must be designed in by lens manufacturer
Multilayer and surface imaging resists	Resist engineering	Improved CD control, resolution, including antireflective functionality	Increased process complexity and cost

CD = critical dimension.

process [18]. The main categories of RETs currently being investigated and used are listed in Table 1.1; two of these techniques – phase-shifting mask (PSM) lithography and optical proximity correction (OPC) – are discussed in the following sections.

Phase-Shifting Mask Lithography Phase-shifting masks for optical lithography were developed by Levenson and coworkers at IBM during the early 1980s [19], although the independent development of phase shifting was also underway at the same time by Shibuya [20] and Smith [21, 22]. Phase-shifting masks are photomasks that take advantage of the interference generated by phase differences to improve the image resolution in photolithography, and exist in either alternating or attenuated forms.

Alternating Phase-Shifting Masks. In the case of an alternating PSM, a thin layer of transparent material of correct thickness is added onto the mask; this induces an abrupt change of the phase of the light used for exposure, and causes optical attenuation at desired locations. These phase masks (which are also known as "phase shifters") have produced features of −100 nm [23, 24] in photoresist. The correct thickness of the shifter is usually demonstrated by the physical thickness that provides an optical path length exactly one-half wavelength longer than the optical path length in the same thickness of air.

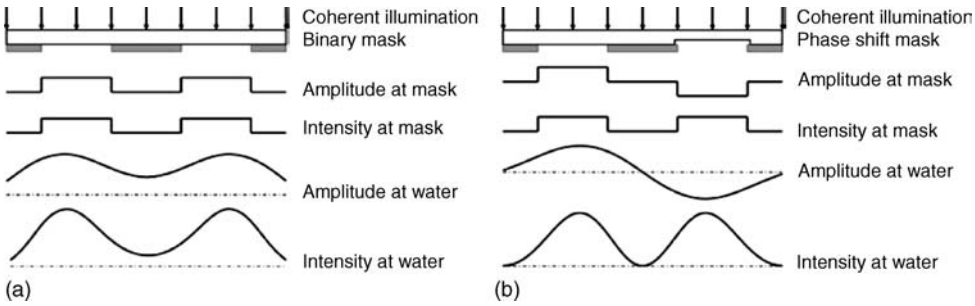


Figure 1.3 Alternating-phase shift masks. (a) Superposition of aerial image amplitudes for coherent illumination of a binary mask; (b) Superposition of aerial image amplitudes for coherent illumination of an alternating-phase shift mask.

The required phase shifter thickness is given by:

$$t = \frac{\lambda}{2(n-1)}$$

where n is the index of refraction of the shifter material. For typical conditions with $n = 1.5$, the phase shifter thickness is the same as the exposure wavelength.

The general concept of phase-shifting lithography is illustrated schematically in Figure 1.3. Phase masks may be used in both projection and contact mode photolithographic techniques, with the achievable photolithographic resolution being approximately $-\lambda/4n$, where λ is the wavelength of the exposure light and n is the refractive index of the photoresist. In fact, feature lines as narrow as 50 nm have been generated in this way [25, 26], with the resolution achieved corresponding approximately to $\lambda/5$. One improved approach to conformal near-field photolithography is to use masks constructed from "soft" organic elastomeric polymers.

Attenuated Phase-Shift Masks. Attenuated PSM lithography improves the pattern fidelity by "darkening" the edges of shapes through the destructive interference of light, using a mildly translucent photomask. Now commonly called "embedded attenuated phase masks," mask substrates are used that allow a small amount of light (~7–10%) to penetrate the normally opaque mask regions. Terazawa and coworkers have developed a significantly different implementation of the PSM concept, which is often referred to as the embedded attenuated phase-shifting mask (EAPSM) [27]. The attenuated PSM functions with the same basic interference principles as the alternating PSM, but the details are totally different: some sketches of the attenuated PSM enhancement mechanism are shown in Figure 1.4.

Due to their simpler construction and operation – particularly in combination with optimized illumination for memory patterns – attenuated PSMs are already widely used. Although alternating PSMs are more difficult to manufacture, and this has slowed their adoption, their use is becoming more widespread. For example, the alternating PSM technique is currently being used by Intel to print gates for their 65 nm and subsequent node transistors [28, 29].

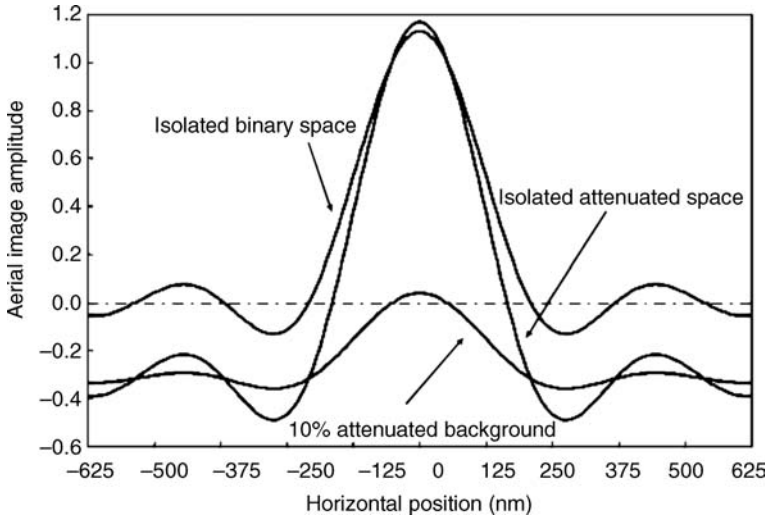


Figure 1.4 Attenuated-phase shift masks: Superposition of aerial image amplitudes for coherent illumination.

1.2.1.5 **Optical Proximity Correction (OPC)**

Optical proximity correction, as a photolithography enhancement technique, is frequently used to compensate for image errors caused by diffraction or process effects. The need for OPC applies mainly to the creation of semiconductor devices, the main reason for which is the limitation of light to resolve ever-finer details of patterns on the photomasks used to etch semiconductor passivation layers, and the difficulty in creating the building blocks of the transistors and other elements that constitute ICs. These projected images appeared as irregularities such as rounded corners, and with trace widths that were narrower than designed. If the diffraction effects could not be compensated, then the electrical properties of the fabricated unit would be significantly altered due to these distortions. OPC anticipates the irregularities of shape and size, and applies a corrective compensation to the photo mask images, which then produce a light beam that more closely approximates the intended shapes.

The two most common applications for OPC are line width differences between features in regions of different density (e.g., center versus edge of an array, or nested versus isolated lines), and line end shortening (e.g., gate overlap on a field oxide). In the case of the line width differences, scattering bars (subresolution lines placed adjacent to resolvable lines) or simple line width adjustments are applied to the design. For line end shortening, "dog-ear" (serif or hammerhead) features are attached to the line end in the design. OPC has a cost impact on photomask fabrication, as the addition of OPC features means more spots for defects to manifest themselves. Additionally, when using OPC the data size of the photomask layout will rise exponentially.

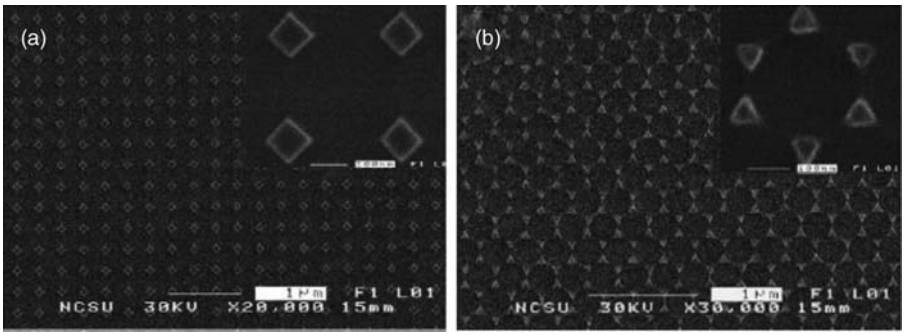


Figure 1.5 (a) Diamond shape pattern with the size of 80 nm × 80 nm and periods of 300 nm; (b) triangular shape pattern with the size is ∼73 nm and periods of 159 nm.

The OPC process is started by characterizing the patterning operation and all its inaccuracies from various sources, such as the mask build, wafer exposure, etch, and so on. In the now commonplace "model-based OPC," this mathematical description of the process is used in iterative optimization routines to predistort the mask shapes to compensate for known, systematic, and modeled patterning inaccuracies.

The OPC technique improves the "effective resolution" of a patterning process by overlapping the conditions with which different feature types can be imaged accurately. Nested features typically image on-size and with the best image quality at a different exposure dose than do isolated features. However, both feature types can be imaged adequately in a single exposure by biasing the mask patterns appropriately. However, OPC does not alter the fundamental resolution limits of a lithography system.

The in-plane anisotropic nanostructures fabricated by using deep UV lithography are shown in Figure 1.5. In this case, Figure 1.5a shows an 80 nm × 80 nm (113 nm axis length) diamond-shaped pattern with a 300 nm period, while Figure 1.5b shows the triangle-shaped pattern [30].

1.2.2
Electron Beam Lithography

Electron-beam lithography (EBL; also termed E-beam lithography) involves scanning a beam of electrons in a patterned fashion across a surface covered with a resist [31]. The first EBL system was developed during the late 1960s, and was based on the principle of scanning electron microscopy (SEM). Because of its excellent high resolution and flexibility, EBL represents a specialized technique for creating the extremely fine patterns required by the modern electronics industry for use in ICs [32–35].

Typically, a EBL system consists of the following parts: (i) an electron gun or electron source that supplies the electrons; (ii) an electron column that "shapes" and focuses the electron beam; (iii) a mechanical stage that positions the wafer under the electron beam; (iv) a wafer handling system that automatically feeds wafers to the system and unloads them after processing; and (v) a computer system that controls the equipment.

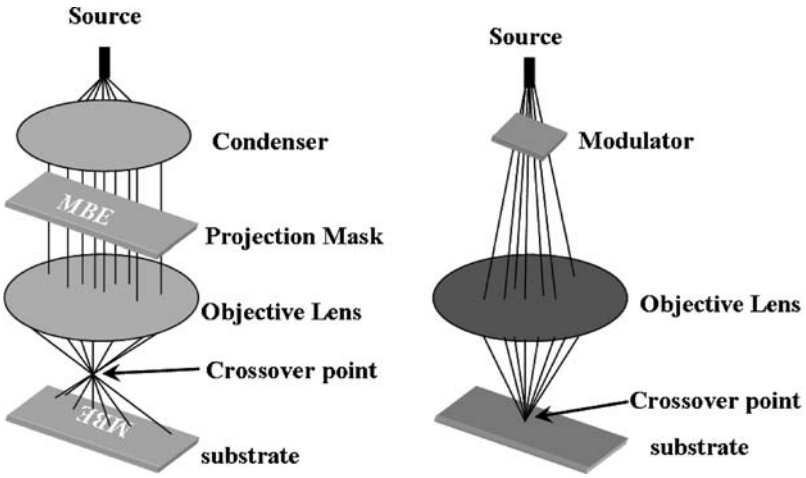


Figure 1.6 Schematic diagram of direct write and project printing.

In general, two distinct schemes are used in EBL:

- *Projection printing*, in which a relatively large-sized electron beam pattern is projected in parallel through the mask onto a resist-coated substrate by using a high-precision lens system.
- *Direct writing*, in which a small spot of the electron beam is written directly onto a resist-coated substrate, thus eliminating the expensive and time-consuming production of masks.

The principles of direct write and project printing are illustrated schematically in Figure 1.6.

EBL displays certain *advantages* over conventional photolithography techniques:

- It is capable of very high resolution, almost to the atomic level. Typically, EBL has a three orders of magnitude better resolution, although this is limited by the forward scattering of electrons in the resist layer, and back scattering from the underlying substrate.
- It is a flexible technique that can function with a wide variety of materials and an almost infinite number of patterns.

Unfortunately, however, EBL has certain *disadvantages*. Notably, it is slow in operation, being one or more orders of magnitude slower than optical lithography. Perhaps more importantly, however, it is expensive and complicated, with EBL systems costing many millions of dollars to purchase and requiring frequent servicing to maintain performance. Yet, despite these drawbacks, EBL currently represents the most powerful tool for the fabrication of features as small as 3–5 nm [36, 37].

Today, EBL is used principally in support of the IC industry, where it has three niche markets:

- In mask-making, typically the chrome-on-glass masks used by optical lithography tools. EBL is the preferred technique for masks because of its flexibility in providing a rapid turn-around of a finished part described only by a computer CAD file. The ability to meet stringent line width control and pattern placement specifications, on the order of 50 nm each, is a remarkable achievement.
- In direct writing for the advanced prototyping of ICs [38] and the manufacture of small-volume specialty products, such as gallium arsenide ICs and optical waveguides.
- For research into the scaling limits of ICs [39] and studies of quantum effects and other novel physics phenomena, at very small dimensions.

Since it is impossible to deflect an electron beam to cover a large area, in a typical EBL system mechanical stages are required to move the substrate through the deflection field of the electron beam column. Stages can be operated in a stepping mode in which the stage is stopped, an area of the pattern written, and the stage then moved to a new location where an adjacent pattern area is exposed. Alternatively, stages can be operated in a continuous mode where the pattern is written on the substrate while the stage is moving. Figure 1.7 shows SEM images of dots with a 15 nm diameter and a 35 nm pitch (64 mC cm^{-2}) fabricated in a 30 nm calixarene resist layer on a Si substrate at 50 keV acceleration voltage [40].

1.2.3 X-Ray Lithography

X-rays with wavelengths in the range of 0.04 to 0.5 nm represent another alternative radiation source with the potential for high-resolution pattern replication into

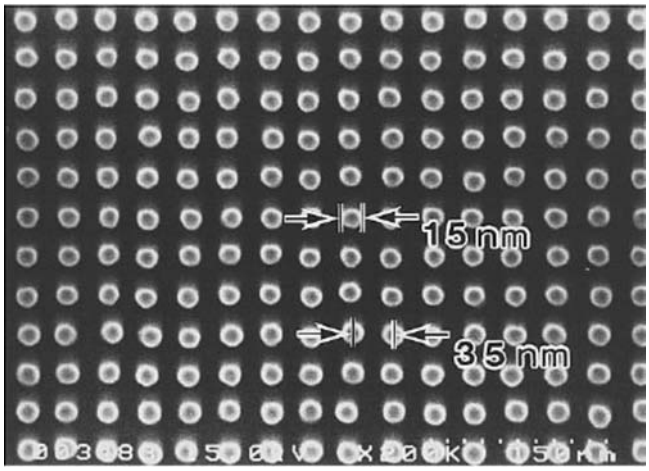


Figure 1.7 Scanning electron microscopy image of 15 nm dots at a pitch of 35 nm written in 30 nm-thick calixarene resist [40].

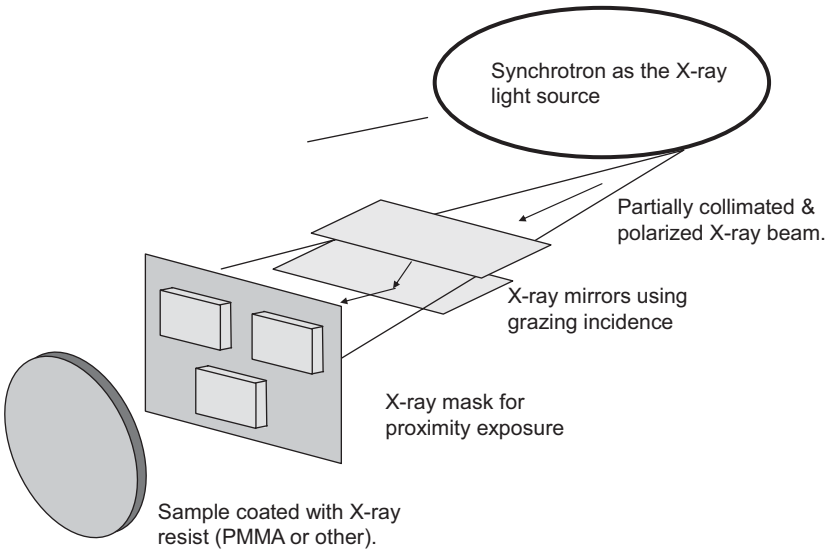


Figure 1.8 Schematic of X-ray lithography.

polymeric resist materials [41]. X-ray lithography was first shown to produce high-resolution patterns, using X-ray proximity printing, by Spears and Smith [42]. Although X-ray lithography uses the same procedure as optical lithography and EBL, an X-ray source is applied rather than using UV light or an electron beam. It would seem that X-ray lithography represents a next-generation lithography developed for the semiconductor industry, as its novel technology can be used in the same capacities as optical lithography, but with better results.

The principle of X-ray lithography is shown schematically in Figure 1.8. Basically, X-ray lithography is a shadow printing process in which patterns coated on a mask are transferred into a third dimension in a resist material, normally poly (methyl methacrylate) (PMMA). The essential ingredients in X-ray lithography include:

- A shadow mask, prepared on a thin membrane of X-ray transmitting material consisting of patterns made from an X-ray absorbing material.
- An X-ray sensitive material, which serves as the X-ray resist of high resolution and is suitable for the subsequent fabrication.
- An X-ray source of sufficient brightness in the suitable region to expose the resist through the mask.

The X-ray radiation sources represent important factors in this type of lithography, and may be generated in several ways, including: (i) by electron bombardment; (ii) by electron impact; (iii) via synchrotron sources; and (iv) as laser-generated plasmas as X-ray sources, among which synchrotron radiation sources are by far the brightest sources of soft X-rays [43–47].

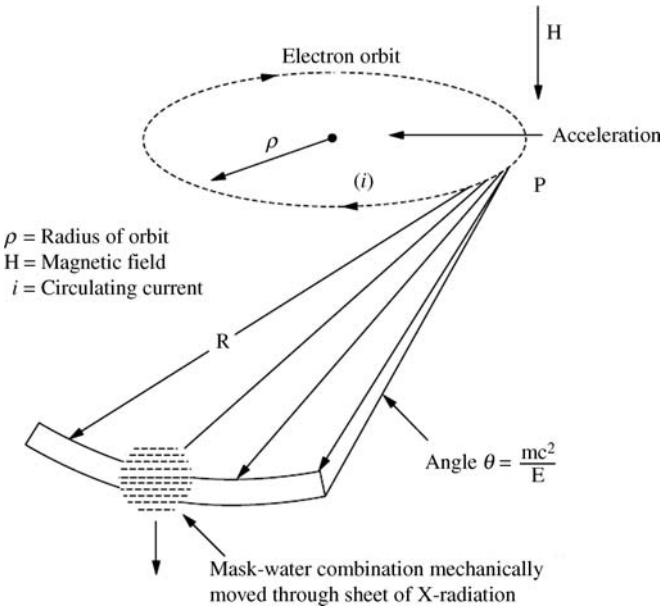


Figure 1.9 Schematic diagram of an X-ray exposure station with a synchrotron radiation source.

Synchrotron radiation is emitted by high-energy relativistic electrons in a synchrotron or storage ring, and then accelerated normal to the direction of their motion by a magnetic field. This leads to the production of a range of electrons over the entire electromagnetic spectrum, from radiowaves to infrared (IR) light, visible light, UV light, X-rays, and gamma rays. Synchrotron or storage rings, which produce a broad spectrum of radiation stemming from the energy loss of electrons in motion at relativistic energies, have been developed primarily for experiments in high-energy physics. A schematic of an X-ray exposure station with a synchrotron radiation source is shown in Figure 1.9, where the X-ray radiation opening angle, θ, is tangential to the path of the electron describing a line on an intersecting substrate.

X-ray proximity lithography is known to provide a one-to-one replication of the features patterned on the mask, and the resolution limit of the X-ray lithography is ~25 nm [48]. The first components to be created using X-ray lithography were surface acoustic wave devices [49], and this was soon followed by bipolar [50] and metal oxide semiconductor (MOS) transistors [51]. The first sub-100 nm Si transistors were produced using X-ray lithography [52], and a velocity overshoot was observed as well as transconductances greater than $1\,S\,mm^{-1}$ [53]. A wide variety of sub-100 nm quantum-effect devices have been fabricated using X-ray lithography [54], including those in which the Coulomb-blockade effect was first observed [55]; this has led to what is often referred to as the "single-electron transistor".

The SEM images of 35 nm-wide Au lines and 20 nm-wide W dots fabricated by electroplating and reactive ion etching, in combination with X-ray lithography, are shown in Figure 1.10 [56].

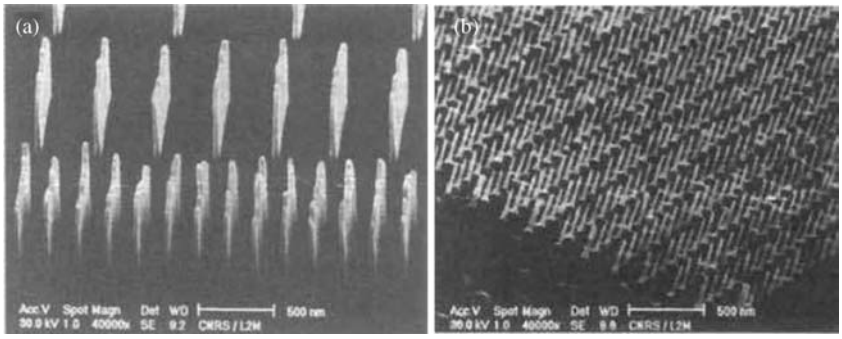


Figure 1.10 (a) 35 nm-wide Au lines grown by electroplating using a template fabricated by X-ray lithography. The mean thickness is about 450 nm, which corresponds to an aspect ratio close to 13; (b) 20 nm-wide W dots obtained after reactive ion etching of a 1250 nm-thick W layer [16].

1.2.4
Focused Ion Beam (FIB) Lithography

During the past few decades, a wide variety of nanofabrication techniques using photons, electrons, and ions have been investigated, and today focused ion-beam (FIB) technology represents one of the most promising techniques for nanofabrication, due to its great flexibility and simplicity.

Since the introduction of FIB technology to the semiconductor industry during the early 1980s, various applications have been developed for both the removal (direct ion milling, FIB chemical etching) and deposition (ion implantation, FIB chemical deposition) of a number of conductor and isolator materials, with sub-micron precision [57]. The FIB technique has also been rapidly developed into a very attractive tool for lithography, etching, deposition, and doping [58]. Because of the matching of ion and atom masses, the energy transfer efficiency of the ion beam to resist is significantly greater than with electron beams. Coupled with the fact that a focusing ion system typically operates at elevated potentials (up to 150 kV), the effective sensitivity charge per unit area of ion beams is two to three orders of magnitude higher than for electron beams, ion-beam lithography has long been recognized as offering an improved resolution [59, 60], and has also shown promise for high-resolution microfabrication [61]. When compared to EBL, FIB lithography has the advantages of a high resist exposure sensitivity, negligible ion scattering in the resist, and low back-scattering from the substrate [62].

Among the different FIB processes, direct-write milling and the dry development of FIB-implanted resists have been widely investigated for both the microfabrication and nanofabrication of advanced IC devices [63, 64].

1.2.4.1 **Direct-Write Milling**

The direct-write milling of the substrate by FIB represents the simplest process for pattern fabrication. In this method, the resists are eliminated and the dose of ions can be varied as a function of position on the wafer. The technique also utilizes heavy-ion

species such as Ga^+ and Au^+. Direct FIB milling has also been applied for lithography mask repair and circuit microsurgery, with resolution down to 100 nm [65]. Moreover, any opaque defects, such as any excess metal on the chromium-based masks, can simply be milled off, while clear defects can be repaired by milling a light-scattering structure (prism) into the area to be rendered opaque. FIB milling has also been applied to bilayer-structure lithography, in which a thin film of gold is usually deposited on top of the conventional resist [64].

1.2.4.2 Dry Development FIB Lithography

Dry development FIB lithography will also yield high aspect ratio structures, with nanometer resolution [63, 66, 67]. Likewise, FIB lithography may be combined with dry development processes by using the well-known top surface imaging (TSI) technique [68], which eliminates the need for wet processing and thus avoids any pattern deformation due to swelling. The limited penetration range of ions is a perfect match for the TSI processes, in which the surface of the resist is selectively manipulated by exposure to silicon-containing chemicals, so as to withstand oxygen dry development in unexposed areas [68, 69]. Other TSI processes utilize the dry development of ion-beam-irradiated resists for negative image formation in exposed areas [60, 70]. In these studies, PMMA resist regions implanted with different ion species (such as Ga^+ and Si^+) have demonstrated significant reductions in the etching rates during the oxygen reactive ion etching (RIE) process. The ion-beam-inhibited etching phenomenon can be explained on the basis of the formation of stable oxide layers during the etching process (i.e., Ga_2O_3 and SiO_2) [64]. Another explanation for etch resistance occurring in the implanted resist regions is the concept of a *physical hardening* of the resist [67]. According to this interpretation, the incident ions may break chemical bonds within the photoresist resin by sputtering the hydrogen and oxygen atoms away, which in turn results in the formation of a stable, carbon-rich "graphitized" structure.

1.3 Two-Dimensional Nanostructures: Thin-Film Deposition

Generally, the term "thin film" is applied to layers which have thicknesses on the order of microns or less, but which may be as thin as a few atomic layers. The deposition of thin films has been the subject of intensive study for almost a century, and a wide range of appropriate methods have been developed and improved. Today, electronic semiconductor devices and optical coatings are the main applications to benefit from thin-film construction. Although many excellent textbooks and monographs have been published on this topic [71–73], in this section a brief introduction of the fundamentals will be provided, and the typical experimental approaches of various well-established techniques of film deposition summarized.

Depending on the depositing process, film growth methods can be generally divided into two broad categories, namely *physical deposition* and *chemical deposition*. The former processes (often just called thin-film processes) are atomistic deposition

processes in which material is vaporized from a solid or liquid source in the form of atoms or molecules, and include evaporation, molecular beam epitaxy (MBE), sputtering, pulsed laser deposition, and cathodic arc deposition. In contrast, the latter processes are based on chemical reactions that include chemical vapor deposition (CVD), plasma-enhanced chemical deposition (PECVD) and atomic layer deposition (ALD). Recently, a considerable number of novel processes that utilize a combination of different methods have been developed. This combination allows a more defined control and tailoring of the microstructure and properties of thin films. Typical processes include ion beam-assisted deposition (IBAD) and plasma-enhanced CVD (PECVD).

1.3.1
Fundamentals of Film Growth

The growth of thin films, as with all phase transformations, involves the processes of nucleation and growth on the substrate or growth surfaces. The nucleation process plays a very important role in determining the crystallinity and microstructure of the resultant films. Lattice mismatch has a marked effect on film morphology, as strain resulting from lattice mismatch contributes to the interface energy, which is a key parameter in determining the growth mode. In general, nucleation in film formation is a heterogeneous process, although the surface free energies for the substrate and film materials also influence the mode of growth. Depending on the resulting film morphology, the growth modes have been placed into three categories [74]:

- Frank–van der Merwe (FM) or layer-by-layer growth
- Volmer–Weber (VW) or 3-D island growth
- Stranski–Krastanow (SK) or 3-D island-on-wetting-layer growth.

These three basic modes of initial nucleation in film growth are illustrated in Figure 1.11.

In the FM growth mode, the interatomic interactions between the substrate and film materials are stronger and more attractive than those between the different atomic species within the film material; this results in the first complete monolayer being formed before deposition of the second layer occurs. The most important examples of layer growth mode are the epitaxial growth of single crystal films.

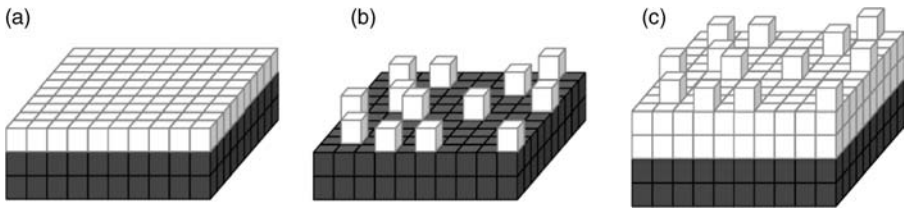


Figure 1.11 Basic modes of initial nucleation in the film growth. (a) Frank–van der Merwe mode (two-dimensional growth mode); (b) Volmer–Weber mode (Island growth mode); (c) Stranski–Krastanov mode.

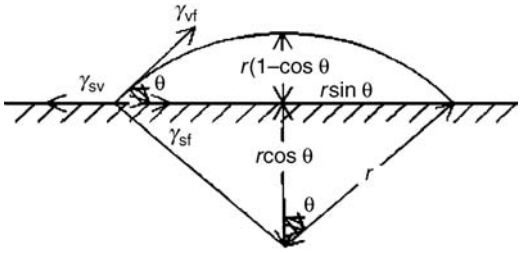


Figure 1.12 Schematic illustrating heterogeneous nucleation process with all related surface energy in equilibrium.

The VW growth mode contrasts with the FM growth mode, with the growth species being bound more strongly to each other than to the substrate; in this way, the islands that are formed initially will finally coalesce to form a continuous film. Many systems of metals on insulator substrates, alkali halides, graphite and mica substrates display this type of nucleation during the initial film deposition.

The SK growth mode occurs for interaction strengths somewhere in between FM and VW, where the layer growth and island growth are combined in intermediate fashion. This type of growth mode typically involves the stress that is developed during the formation of either nuclei or films.

When a new phase forms on a surface of another material, the process is termed "heterogeneous nucleation." The film deposition involves predominantly heterogeneous processes that include heterogeneous chemical reactions, evaporation, adsorption and desorption on growth surfaces, which in turn involves a heterogeneous nucleation at the initial stage. Provided that the growth species in the vapor phase impinge on the substrate surface, these growth species will diffuse and aggregate to form a nucleus with a cap shape, as illustrated in Figure 1.12. During formation of the thin film the Gibbs free energy and the surface energy will be increased, with the change in chemical energy, ΔG, which is associated with formation of the nucleus, given by:

$$\Delta G = \mathrm{a}_3 r^3 \Delta u_v + a_1 r^2 \gamma_{vf} + \mathrm{a}_2 r^2 \gamma_{fs} - \mathrm{a}_2 r^2 \gamma_{sv}$$

where r is the mean dimension of the nucleus, Δu_v is the change of Gibbs free energy per unit volume, and γ_{vf}, γ_{fs} and γ_{sv} are the surface or interface energies of the vapor–nucleus, nucleus–substrate, and substrate–vapor interfaces, respectively. The geometric constants are given by:

$$a_1 = 2\pi(1-\cos\theta)$$
$$a_2 = \pi \sin^2\theta$$
$$a_3 = 3\pi(2-3\cos\theta+\cos^2\theta)$$

where θ is the contact angle, which is dependent only on the surface properties of the surfaces or interfaces involved, and is defined by Young's equation:

$$\gamma_{sv} = \gamma_{fs} + \gamma_{vf}\cos\theta$$

The formation of new phase results in a reduction of the Gibbs free energy, but an increase in the total surface energy. The nucleus is stable only when its size is larger than the critical size, γ^* and the critical energy barrier, ΔG^*, is illustrated respectively:

$$\gamma^* = \frac{-2(a_1\gamma_{vf} + a_2\gamma_{fs} - a_2\gamma_{sv})}{3a_3\Delta G_v}$$

$$\Delta G^* = \frac{4(a_1\gamma_{vf} + a_2\gamma_{fs} - a_2\gamma_{sv})^3}{27a_3^2\Delta G_v}$$

As with the FM growth mode, the substrate is completely wetted by the depositing materials, in which case the contact angle will be equal to zero; thus, the corresponding Young's equation could be described as:

$$\gamma_{sv} = \gamma_{fs} + \gamma_{vf}$$

For island growth, the contact angle, θ, must be larger than zero. According to Young's equation, we then have:

$$\gamma_{sv} < \gamma_{fs} + \gamma_{vf}$$

In case the deposit does not wet the substrate at all, the contact angle will be 180 °C, a process which is commonly referred to as *homogeneous nucleation*.

The lattice constants of the deposit will most likely differ from those of the substrate, which commonly leads to the development of stress in the newly forming film, and a resultant island-layer growth mode. Island-layer growth involves *in situ*-developed stress, and is slightly more complicated than the above two growth modes. The thin film would proceed following the mode of layer growth initially, but when the deposit became elastically strained – due to, for example, a lattice mismatch between the deposit and the substrate – then strain energy would be developed. As each layer of the deposit continues, more stress and strain energy would be developed. Given that there is no plastic relaxation, the strain energy would be proportional to the volume of the deposit. Then, as the growth of the film continued the stress would reach a critical point and could not be released; the strain energy per unit area of deposit would then be large with respect to γ_{vf}, permitting nuclei to form above the initial layered deposit. In this case, the surface energy of the substrate would exceed the combination of both surface energy of the deposit and the interfacial energy between the substrate and the deposit:

$$\gamma_{sv} > \gamma_{fs} + \gamma_{vf}$$

1.3.2 Physical Vapor Deposition (PVD)

Physical vapor deposition (PVD) is fundamentally a vaporization coating technique that involves the transfer of material on an atomic level. The process can be

described according to the following sequence of steps: (i) the material to be deposited is converted into a vapor by physical means; (ii) the vapor is transported across a region of low pressure from its source to the substrate; and (iii) the vapor undergoes condensation on the substrate to form the thin film. Typically, PVD processes are used to deposit films with thicknesses in the range of a few nanometers to thousands of nanometers. However, they can also be used to form multilayer coatings, graded composition deposits, very thick deposits, and freestanding structures.

PVD thin film technology covers a rather broad range of deposition techniques, including electron-beam or hot-boat evaporation, reactive evaporation, and ion plating. PVD techniques also include processes based on sputtering, whether by plasma or by an ion beam of some sort. PVD is also used to describe the deposition from arc sources which may or may not be filtered. In general, the methods can be divided into two groups, namely *evaporation* and *sputtering*. Evaporation refers to thin films being deposited by thermal means, whereas in sputtering mode the atoms or molecules are dislodged from the solid target through the impact of gaseous ions (plasma). Both methods have been further developed into a number of specific techniques.

1.3.2.1 **Vacuum Evaporation**

The formation of thin films by evaporation was first recognized about 150 years ago [75], and has since acquired a wide range of applications, notably in the past fifty years since industrial-scale vacuum techniques were developed [76]. Many excellent books and reviews have been produced describing evaporated films [77]. Vacuum evaporation (or vacuum deposition) is a PVD process in which the atoms or molecules from a thermal vaporization source reach the substrate without colliding with residual gas molecules in the deposition chamber. Vacuum deposition normally requires a vacuum of better than 10^{-4} Torr; however, even at such low pressure there will still be a large amount of concurrent impingement on the substrate by potentially undesirable residual gases that can contaminate the film. If film contamination is problematic, a high (10^{-7} Torr) or ultrahigh ($<10^{-9}$ Torr) vacuum environment can be used to produce a film with the desired purity, depending on the deposition rate, the reactivities of the residual gases and the depositing species, and the tolerable impurity level in the deposit. The evaporation of elemental metals is fairly straightforward: heated metals have high vapor pressures, and in a high vacuum (HV) the evaporated atoms will be transported to the substrate. A schematic image of a typical evaporation system is shown in Figure 1.13.

A typical evaporation system will consist of an evaporation source to vaporize the desired material, and a substrate located at an appropriate distance, facing the evaporation source. Both, the source and the substrate, are located in a vacuum chamber.

The saturation or equilibrium vapor pressure of a material is defined as the vapor pressure of that material in equilibrium with the solid or liquid surface, in a closed container. At equilibrium, as many atoms return to the surface as leave the surface, such that the equilibrium vapor pressure of an element can be estimated as:

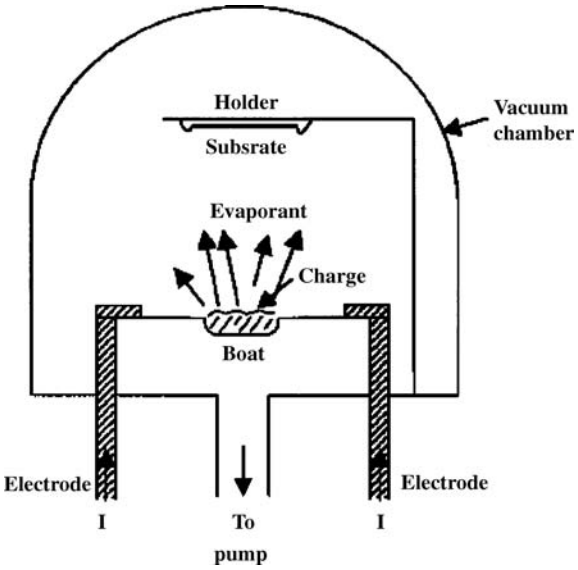


Figure 1.13 Schematic image of a typical evaporation system.

$$\ln P_e = -\frac{\Delta H_e}{RT} + C$$

where ΔH_e is the molar heat of evaporation, R is the gas constant, T is the temperature, and C is a constant.

Although most elements vaporize as atoms, some – such as Sb, Sn, C, and Se – have a significant portion of the vaporized species as clusters of atoms. For materials which evaporate as clusters, special vaporization sources (called *baffle sources*) can be used to ensure that the depositing vapor is in the form of atoms. The evaporation of compounds is more complicated, as compounds may undergo chemical reactions such as pyrolysis, decomposition and dissociation, and the resultant vapor composition may often differ from the source composition during evaporation at elevated temperatures. It should be noted that, as a material is heated, the first components to be volatilized are the high-vapor-pressure surface contaminates, absorbed gases, and high-vapor-pressure impurities.

A material vaporizes freely from a surface when the vaporized material leaves the surface with no collisions above the surface. The free surface vaporization rate is proportional to the vapor pressure, and is given by the Hertz–Knudsen vaporization equation [78, 79]:

$$\frac{dN}{dt} = C\,(2\,\pi mkT)\text{-}\frac{1}{2}\,(p^*\text{-}p)\ \mathrm{s}^{-1}$$

where dN is the number of evaporating atoms per cm^2 of surface area, C is a constant that depends on the rotational degrees of freedom in the liquid and the vapor, p^* is the vapor pressure of the material at temperature T, p is the pressure of the vapor above

the surface, k is Boltzmann's constant, T is the absolute temperature, and m is the mass of the vaporized species.

According to Raoult's law, the constituents of a mixture of elements or compounds vaporize in a ratio that is proportional to their vapor pressures; that is, a constituent with a high vapor pressure will vaporize more rapidly than a material with a low vapor pressure [78, 79]. Thus, the chemical composition of the vapor phase is most likely to be different from that in the source, although adjusting the composition or molar ratio of the constituents in the source may help in this respect. However, the composition of the source would change as the evaporation proceeds, with the higher vapor pressure material steadily decreasing in proportion to the lower vapor pressure material in the melt. As a result, the composition in the vapor phase will change. For a multi-component system, the chemical composition of evaporated film is likely to produce a gradation of film composition as the evaporant is selectively vaporized. Therefore, it is in general difficult to deposit complex films using an evaporation method.

However, vacuum deposition does have advantages in some cases:

- Line-of-sight deposition allows the use of masks to define area of deposition.
- Large-area sources can be used for some materials (e.g., "hog trough" crucibles for Al and Zn).
- High deposition rates can be obtained.
- Deposition rate monitoring is relatively easy.
- The vaporization source material can exist in many forms, such as chunks, powder, wire, and chips.
- A vaporization source material of high purity is relatively inexpensive.
- High-purity films are easily deposited from high-purity source materials, as the deposition ambient can be made as noncontaminating as is desired.
- The technique is relatively inexpensive compared to other PVD methods.

1.3.2.2 **Molecular Beam Epitaxy (MBE)**

Perhaps the most sophisticated PVD process is that of MBE or vapor-phase epitaxy (VPE) [80–82], which take place in either high-vacuum or ultra-high-vacuum environments (10^{-8} Pa) [83–85]. The most important aspect of MBE is the slow deposition rate (typically $<1000\,\mathrm{nm\,h^{-1}}$), which allows the films to grow epitaxially. The slow deposition rates require (proportionally) a better vacuum to achieve the same impurity levels as other deposition techniques. Hence, MBE can be considered a special case of evaporation for single-crystal film growth, with the highly controlled evaporation of a variety of sources in ultra-high-vacuum.

Figure 1.14 shows, schematically, a number of effusion cells aligned radially with the substrates. Since MBE is a variant of evaporation, instead of an open crucible the source material can be heated in an equilibrium source known as the Knudsen cell. In this case, an atomic beam (in the molecular flow regime; hence the name MBE) exits the cell through an orifice that is small compared to the source size. Such equilibrium sources are much more stable than open sources, because they are heated resistively or by an electron beam. In high vacuum, the evaporated atoms do not experience

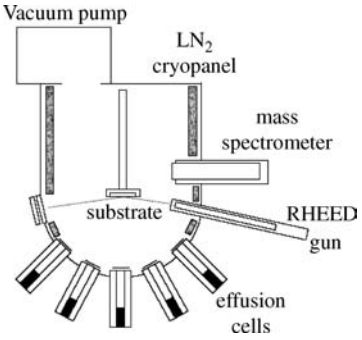


Figure 1.14 Schematic diagram of a number of effusion cells aligned radially with the substrates. LN_2 = liquid nitrogen; RHEED = reflection high-energy electron diffraction.

collisions, and therefore will take a line-of-sight route from the source to the substrate. The mean free path of atoms or molecules (~100 m) far exceeds the distance between the source and the substrate (typically ~30 cm) inside the deposition chamber; therefore, the atoms or molecules striking on the single crystal substrate will result in formation of the desired epitaxial film.

Atoms arriving at the substrate surface may undergo absorption to the surface, surface migration, incorporation into the crystal lattice, and thermal desorption, but which of these competing pathways dominates the growth will depend heavily on the temperature of the substrate. At a low temperature, atoms will "stick" where they land without arranging properly, leading to poor crystal quality; however, at a high temperature the atoms will desorb (re-evaporate) readily from the surface, leading to low growth rates and a poor crystal quality. In the appropriate intermediate temperature range, the atoms will have sufficient energy to move to the proper position on the surface and be incorporated into the growing crystal. The growth mechanism of MBE is shown schematically in Figure 1.15.

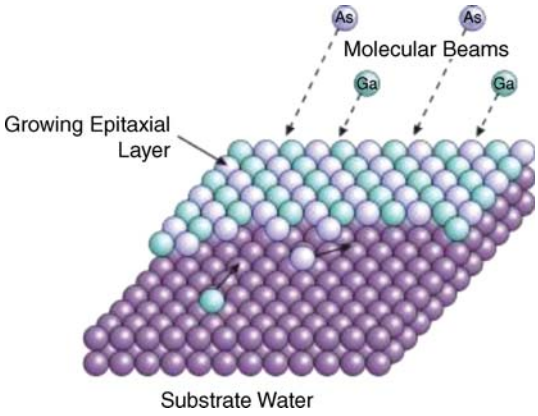


Figure 1.15 The molecular beam epitaxy growth mechanism.

During the MBE process, a range of structural and analytical probes can be used to monitor film growth *in situ*, in real time, on a sub-nanometer scale:

- Reflection high-energy electron diffraction (RHEED), using forward scattering at the grazing angle; this shows a maximum when there is a completed monolayer, and a minimum when there is a partial layer, which produces more scattering.
- Low-energy electron diffraction (LEED) takes place in backscattering geometry, and can be used to study surface morphology, but not during growth.
- Auger electron spectroscopy (AES), records the type of atoms present.
- Modulated beam mass spectrometry (MBMS) allows the chemical species and reaction kinetics to be studied.
- Reflectance difference spectroscopy (RDS).
- Scanning tunneling microscopy (STM) and atomic force microscopy (AFM).

1.3.2.3 **Sputter Deposition**

Sputtering is one of the most important PVD methods, as it involves the physical vaporization of atoms from a surface by momentum transfer from bombarding, energetic, atomic-sized particles. Sputter deposition permits a better control of the composition of multi-element films, and a greater flexibility in the types of materials that may be deposited.

Although first reported by Wright in 1877, the sputter deposition of films became feasible only because a relatively poor vacuum is needed for its operation. Despite the fact that Edison patented a sputter deposition process for depositing silver onto wax photograph cylinders in 1904, the process was not used widely in industry until the advent of magnetron sputtering in 1974. The application of sputter deposition led to an acceleration in the development of reproducible, stable long-lived vaporization sources for production purposes. Following the use of a magnetic field that would confine the motion of the secondary electrons close to the target surface, planar magnetron sputtering has become the most widely used sputtering configuration, having been derived originally from the development of the microwave klystron tube in World War II, from the investigations of Kesaev and Pashkova (in 1959) on confining arcs, and of Chapin (in 1974) on developing the planar magnetron sputtering source [86–89]. The operating principles of both direct current (DC) and radio-frequency (RF) sputtering systems are illustrated schematically in Figure 1.16 [71].

Effective sputter deposition can be achieved in:

- a good vacuum ($<10^{-5}$ Torr) using ion beams;
- a low-pressure gas environment, where sputtered particles are transported from the target to the substrate without gas-phase collisions (i.e., a pressure less than about 5 mTorr), using a plasma as the ion source of ions; and
- a higher-pressure gas, where gas phase collisions and "thermalization" of the ejected particles occurs but the pressure is low enough that gas-phase nucleation is not important (i.e., a pressure greater than about 5 mTorr but less than about 50 mTorr).

Currently, plasma-based sputtering is the most common form of sputtering, in which a plasma is present and positive ions are accelerated to the target which is at

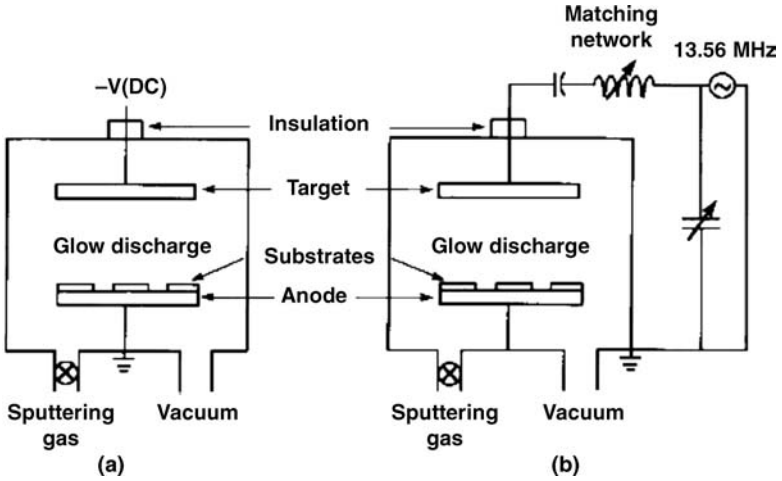


Figure 1.16 Schematic diagram of the principles of (a) direct current (DC) and (b) radiofrequency (RF) sputtering systems.

a negative potential with respect to the plasma. At higher pressures, the ions suffer physical collisions and charge-exchange collisions, so that there is a spectrum of energies of the ions and neutrals bombarding the target surface. At low pressures, the ions reach the target surface with an energy which is given by the potential drop between the surface and the point in the electric field where the ion is formed. In vacuum-based sputtering, however, an ion or plasma beam is formed in a separate ionization source, accelerated, and then extracted into a processing chamber which is maintained under good vacuum conditions. In this process, the mean bombarding energy is generally higher than in the plasma-based bombardment, and the reflected high-energy neutrals are more energetic.

Sputter deposition could be used to deposit films of elemental materials, and also to deposit alloy films and maintain the composition of the target material. This is possible by virtue of the fact that the material is removed from the target in a layer-by-layer fashion, which is one of the main advantages of the process. This allows the deposition of some rather complex alloys such as W : Ti for semiconductor metallization [90], Al : Si : Cu for semiconductor metallization [91], and Metal–Cr–Al–Y alloys for aircraft turbine blade coatings.

The deposition of films of compound materials by sputtering can be achieved either by sputtering from a compound target, or by sputtering from an elemental target in a partial pressure of a reactive gas (i.e., "reactive sputter deposition"). In most cases, the sputter deposition of a compound material from a compound target results in a loss of some of the more volatile material (e.g., oxygen from SiO_2); however, this loss is often made up by deposition in an ambient containing a partial pressure of the reactive gas – a process known as "quasi-reactive sputter deposition." In the latter case, the partial pressure of reactive gas that is needed is less than that used for reactive sputter deposition.

The advantages of sputter deposition include:

- Any material can be sputtered and deposited, including elements, alloys, or compounds.
- The sputtering target provides a stable, long-lived vaporization source.
- Vaporization is from a solid surface and can occur up, down, or sideways.
- In some configurations, the sputtering target can provide a large-area vaporization source.
- In some configurations, the sputtering target can provide specific vaporization geometries, for example, a line source from a planar magnetron sputtering source.
- The sputtering target can be made conformal to a substrate surface, such as a cone or sphere.
- Sputtering conditions can easily be reproduced from run to-run.
- There is little radiant heating in the system compared to vacuum evaporation.
- In a reactive deposition, the reactive species can be activated in a plasma.
- When using chemical vapor precursors, the molecules can be either fully or partially dissociated in the plasma.
- The utilization of sputtered material can be high.
- *In situ* surface preparation is easily incorporated into the processing.

1.3.3
Chemical Vapor Deposition (CVD)

Chemical vapor deposition (CVD) is a method of forming a thin solid film on a substrate by the reaction of vapor-phase chemicals which contain the required constituents. The decomposition of source gases is induced by various energy forms such as chemical, thermal, plasma or photon, and reacted on and/or above the temperature-controlled surface to form the thin film. *Thermal CVD* is the deposition of atoms or molecules by the high-temperature (range from 300 to 900 °C) reduction or decomposition of chemical vapor precursor species which contain the material to be deposited [92–94]. Normally, the reduction is accomplished by hydrogen at an elevated temperature, while the decomposition is accomplished by thermal activation. The deposited material may react with other gaseous species in the system to produce compounds (e.g., oxides, nitrides). In general, CVD processing is accompanied by volatile reaction byproducts and unused precursor species. The CVD process has been studied extensively and is very well documented [95–97], largely due to the close association with solid-state microelectronics.

In CVD the source materials are brought in a gas phase flow into the vicinity of the substrate, where they decompose and react to deposit the film onto the substrate. Any gaseous byproducts are then pumped away, as shown schematically in Figure 1.17.

The CVD process can be generalized in a sequence of steps:

- Reactants are introduced into reactor;
- The gas species are activated and/or dissociated by mixing, heat, plasma or other means.

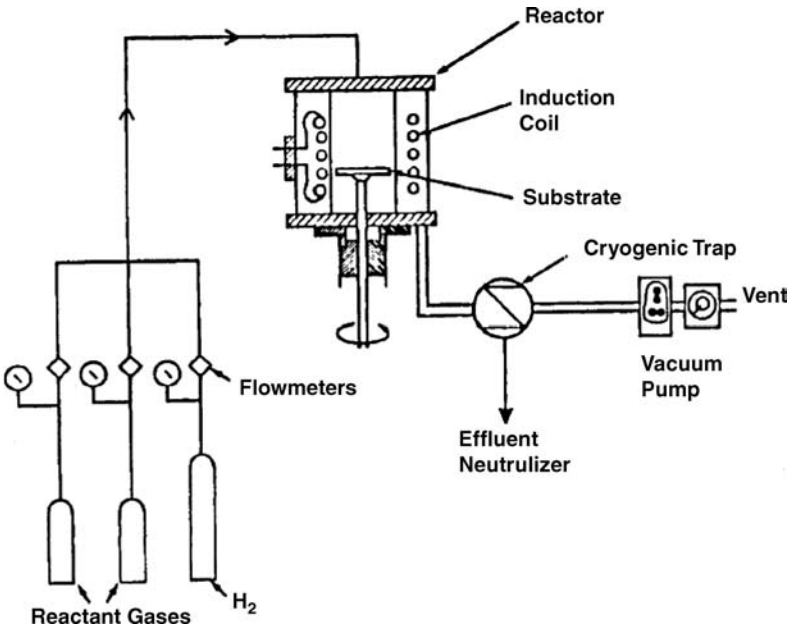


Figure 1.17 The chemical vapor deposition (CVD) process. The source materials are brought in gas phase flow into the vicinity of the substrate, where they decompose and react to deposit film on the substrate. Both, gas-phase transport and surface chemical reactions are important for film deposition.

- The reactive species are adsorbed on the substrate surface.
- The adsorbed species undergo chemical reaction or react with other incoming species to form a solid film.
- The reaction byproducts are desorbed from the substrate surface.
- The reaction byproduct is removed from the reactor.

Due to the versatile nature of CVD, the number of potential chemistries leading to the commonly used films is huge; details of those chemistries that have been, and still are, widely used are listed in Table 1.2. The gas-phase (homogeneous) reactions and surface (heterogeneous) reactions are intricately mixed, but with increasing temperature and partial pressure of the reactants the gas-phase reactions will become progressively more important. Extremely high concentrations of the reactants will cause the gas-phase reactions to become predominant, resulting in homogeneous nucleation. The wide variety of chemical reactions involved can be grouped into: pyrolysis; reduction; oxidation; compound formation; disproportionation; and reversible transfer, depending on the precursors used and the deposition conditions applied.

1.3.3.1 **Reaction Kinetics**

CVD is a nonequilibrium process that is controlled by chemical kinetics and transport phenomena, with the reaction rates obeying Arrhenius behavior. In case

Table 1.2 Reactions used in chemical vapor deposition.

Reaction	Equation
Pyrolysis	SiH_4 (g) → Si (s) + 2 H_2 (g)
Reduction	$SiCl_4$ (g) + 2 H_2 (g) → Si (s) + 4 HCl (g)
Hydrolysis	$SiCl_4$ (g) + 2 H_2 (g) + O_2 (g) → SiO_2 (s) + 4 HCl (g)
Compound formation	$SiCl_4$ (g) + CH_4 (g) → SiC(s) + 4HCl(g)
	$TiCl_4$ (g) + CH_4 (g) → TiC(s) + 4HCl(g)
Disproportionation	2 GeI_2(g) → Ge(s) + $^{\bullet}GeI_4$(g) at 300 °C
Reversible transfer	As_4(g) + As_2(g) + 6GaCl(g) + 3H_2(g) → 6GaAs(s) + 6HCl(g)

of the deposition rate determined at several temperatures, the activation energy E_a can be extracted from the Arrhenius formula. The magnitude of the activation energy gives hints to possible reaction mechanisms.

For most CVD reactions, two temperature regimes can be found (Figure 1.18). Here, when the temperature is low the surface reaction rate is low, and there is an overabundance of reactants; the reaction is then in the *surface reaction-limited regime*. The rate of silicon nitride deposition from SiH_2Cl_2 at 770 °C is approximately 3.3 nm min^{-1}, but this is compensated by the fact that the deposition can take place simultaneously on up to 100 wafers.

As the temperature increases, the reaction rate on the surface is increased exponentially such that, above a certain temperature, all of the source gas molecules will react at the surface. The reaction is then in the *mass transport-limited regime*, because the rate is dependent on the supply of a new species to the surface. The fluid dynamics of the reactor then plays a major role in both deposition uniformity and the reaction rate.

1.3.3.2 **Variants of CVD Methods [98]**

The conventional CVD process, which is based on thermally activated CVD, uses inorganic precursor sources, with the deposition process being initiated by thermal

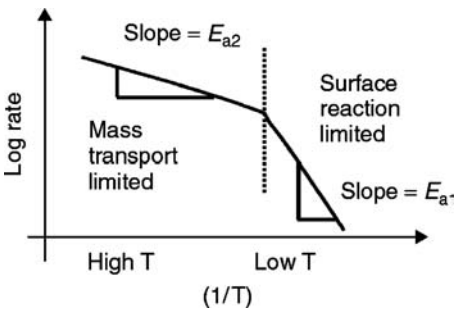


Figure 1.18 Surface reaction-limited versus mass transfer-limited CVD reactions.

energy and occurs at atmospheric pressure, low pressure, or ultrahigh vacuum. The deposition often requires relatively high temperatures (typically 500–1400 °C), depending on the type of inorganic precursor used (e.g., halides, hydrides). According to the forms of energy introduced to the system in order to activate the chemical reactions desired to deposit solid films onto substrates, a variety of CVD methods and CVD reactors have been developed, including PECVD and photo-assisted CVD (PACVD) which use plasma and light, respectively, to activate the chemical reactions. Atomic layer epitaxy (ALE) represents a special mode of CVD where a "monatomic layer" can be grown in sequence by employing sequential saturating surface reactions, while metal–organic CVD (MOCVD) uses a metal–organic precursor rather than an inorganic precursor as used in conventional CVD methods.

These CVD variants are useful when there is a need to control the growth of epitaxial films, and to fabricate tailored molecular structures. Other CVD variants, such as pulsed-injection MOCVD and aerosol-assisted CVD use special precursor generation and delivery systems, unlike conventional CVD; for example, flame-assisted vapor deposition (FAVD) uses a flame source to initiate the chemical reaction and/or heat the substrate. Electrochemical vapor deposition (EVD) represents another variant of CVD that has been tailored to deposit dense films onto porous substrates. In addition, chemical vapor infiltration (CVI) is a form of CVD that has been adapted for the deposition of a dense ceramic matrix during the fabrication of ceramic fiber-reinforced ceramic matrix composites.

Currently emerging low-cost, non-vacuum CVD-based techniques (e.g., aerosol-assisted and flame-assisted CVD) have the potential to be scaled up for large area or mass production. Although most of these variants can also be carried out at either atmospheric or reduced pressure, PACVD must be conducted at low pressure (typically 1.3 to 1333 Pa) in order to generate the plasma.

1.3.3.3 **Advantages of CVD**

Although CVD is a complex chemical system, it has its distinctive advantages. Notably, the process is gas phase in nature; hence, given a uniform temperature within the coating retort and uniform concentrations of the depositing species, then the rate of deposition will be similar on all surfaces. Consequently, variable shaped surfaces such as screw threads, blind holes and channels or recesses, if provided with reasonable access to the coating powders or gases, can be coated evenly without any build-up on the edges.

In some cases, it is possible to form ductile CVD layers, (e.g., chromizing of low-carbon mild steel). (*Note*: "Chromizing" refers to a type of coating process developed for coinage production, in which chromium is deposited onto mild steel blanks by CVD and diffused into the surface to generate a layer that is, effectively, a ferritic stainless steel.) Given that the processing temperature will normally anneal any hilly ferrous substrate onto which the CVD layer is deposited, it would then become practicable to form, press or bend these components successfully after coating. This is in direct contrast with stainless steel components, which become significantly work-hardened during any forming or pressing, and may cause the rapid wearing of any tooling used in the process.

The high temperatures used during CVD results in a considerable amount of diffusion of the coating into the substrate; consequently, if the thermal expansion coefficients are compatible between the coating and substrate, then the adhesion will be excellent. In many cases the substrate can be heat-treated after CVD coating, with no distress to the coating.

1.3.4 Atomic Layer Deposition (ALD)

Atomic layer deposition (ALD), a unique thin film-deposition technique, is based on the sequential use of a gas-phase chemical process with atomic-scale precision, and differs significantly from other thin film-deposition methods. By keeping the precursors separate throughout the coating process, the thickness of the film grown can be controlled down to the atomic/molecular scale per monolayer [99, 100]. Some excellent reviews have been produced by Ritala and Leskela [101, 102] on the subject of ALD which, in the literature, is also referred to as ALE, atomic layer growth (ALG), atomic-layer CVD (ALCVD), and molecular layer epitaxy (MLE). In comparison with other thin film-deposition techniques, ALD is relatively new and was first used to grow ZnS films [103]. Additional details on ALD were reported during the early 1980s [104–106]. ALD can be considered as a special modification of CVD, or even as a combination of vapor-phase self-assembly and surface reaction. It is similar in terms of its chemistry to CVD, except that in ALD the CVD reaction is broken into two half-reactions; this allows the precursor materials to be kept separate during the reaction, so that they can react with a surface one-at-a-time, in sequential manner.

ALD was developed during the late 1970s, and subsequently introduced worldwide as ALE [107]. The ALD deposition method was mainly developed in response to the need for thin-film electroluminescent (TFEL) flat-panel displays, as these require high-quality dielectric and luminescent films on large-area substrates. Subsequently, interest in ALD increased stepwise during the mid-1990s and 2000s, with interest focused on silicon-based microelectronics. In this respect, reviews produced by Ritala [108] and Kim [109] represent recent key references. Today, both the equipment required for ALD, and the processes, have moved through two generations and are approaching a third generation to be hall-marked by its higher productivity, reliability, and other enhancements. Currently, ALD is considered to be the deposition method with the greatest potential for producing very thin, conformal films and, in particular, an ability to control the thickness and composition of the films at the atomic level. One major driving force in this area has been the recent interest in using ALD to scale down microelectronic devices. In fact, the recently demonstrated ability of ALD to produce outstanding dielectric layers and attracts for the semiconductor industry for use in high-K dielectric materials has led to an acceleration of ALD development. A typical atomic layer deposition system is illustrated schematically in Figure 1.19.

1.3.4.1 ALD Process

The growth of material layers by ALD consists of repeating the following characteristic four steps:

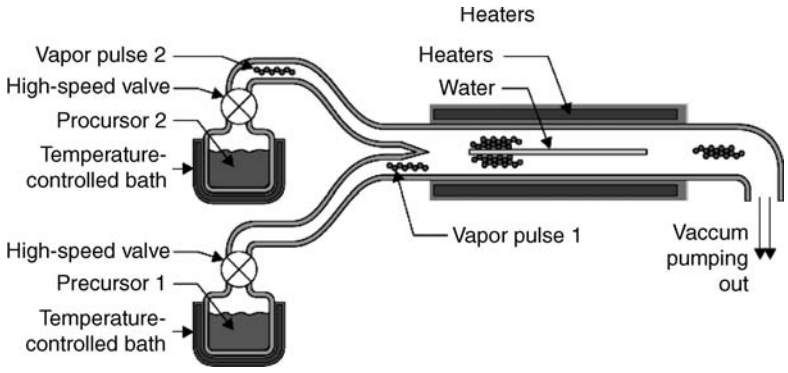


Figure 1.19 Schematic image of a typical atomic layer deposition system.

- Exposure of the first precursor.
- Purging or evacuation of the reaction chamber to remove the nonreacted precursors and the gaseous reaction byproducts.
- Exposure of the second precursor, or another treatment to activate the surface again for the reaction of the first precursor.
- Purging or evacuation of the reaction chamber.

1.3.4.2 Types of ALD Reaction

Today, ALD may be carried out by either thermal reactions or by plasma-assisted processes (which are also partially thermally activated). Almost all these reactions comprise a two-step ALD process, with single unbalanced heuristic chemical reactions being used unless the reaction chemistry is unclear without an explicit two-step description. Some of the reactions selected from important examples of films used in the development of ALD technology and applications are detailed in the following sections.

Thermal ALD

1) Depositing Compounds

In this class of reactions, metal halides are reacted with hydrides, such as H_2O or NH_3. Although $TiCl_4$ and WF_6 are well-known halide sources with good vapor pressures, many other metallic halides may often exist as solids and may sublime at convenient source temperatures, but have low vapor pressures. A thermal ALD reaction of note using metal halides is the $HfCl_4$ chemistry:

$$HfCl_4 + 2H_2O \rightarrow HfO_2 + 4HCl.$$

Recently, HfO_2 has emerged as a promising high-K dielectric because of its chemical stability with silicon relative to ZrO_2. A basic $HfCl_4/H_2O$ process has been demonstrated [110]. Other metal oxides with known halide reactions include: ZrO_2 [111], Ta_2O_5 [112], Al_2O_3 [113], and TiO_2 [114, 115].

The use of metal–organic (MO) chemistry reactions represents another means of forming compounds. Indeed, although the early ALD processes were often developed with halide/hydride chemistry, the use of trimethyl aluminum (TMA) for aluminum-bearing films was an important exception.

Moving from halide sources to MO sources eliminates trace Cl or F, but introduces the presence of trace C or N, but this is a necessary trade-off as organic chemistry provides a wide diversity of materials. Among other precursor characteristics, it may be desirable to develop liquid precursors with a higher pressure and, concurrently, with a greater thermal stability.

Although the ALD process for Al_2O_3, using the TMA/H_2O chemistry [116], was the "standard" process for many years, semiconductor device leakage characteristics were subsequently found to be better with TMA/O_3 chemistry [117].

$$Al(CH_3)_3 + HOH \rightarrow Al_2O_3 + CH_4.$$

$$Al(CH_3)_3 + O_3 \rightarrow Al_2O_3 + H, C, O \text{ containing by products.}$$

2) Depositing Elemental Films

Reactions forming elemental films using halide–hydride chemistry have been developed, in which alternating pulses of a silicon halide and silicon hydride can lead to the formation of elemental silicon [118].

An early demonstration of the ALD formation of elemental W films was using metal halides with silane reduction chemistry [119]:

$$WF_6 + SiH_4 \rightarrow W + SiF_xH_y + SiF_4 + H_2.$$

As the strong bonding energy of the SiF_x byproduct compound essentially leaves no Si to react with the metal, WSi_x is not formed. This chemistry is similar to that known for CVD W processes. Other refractive metals can be formed using silane or other hydride-based reduction reactions.

3) Depositing Noble Metals with O_2 Chemistry

This type of chemical reaction [120] may run counter to initial intuition, as oxygen "combustion" is used to create an elemental material. However, RuO is an intermediate, and once RuO_x forms on the substrate surface it is further reduced by Ru $(Cp)_2$ to form an elemental material:

$$RuO_x + Ri(Cp)_2 \rightarrow Ru\text{-}Ru(Cp) + CO_2 + H_2O$$

$$Ru\,(Cp) + O_2 \rightarrow RuO_x + CO_2 + H_2O$$

where $(Cp)_2 = (C_5\,H_5)_2$. After each of the O_2 half-reactions, the surface is terminated in RuO_x; however, after each pair of half-reactions an additional layer of Ru is added to the bulk film. The concentration of the O may have to be controlled in order to avoid oxidation of the underlying layers.

Plasma-Assisted ALD

1) Depositing Elemental Films

An early Si elemental ALD thermal process which used SiH_2, Cl_2/H_2 [121] was demonstrated by using metal or silicon halides with atomic H; however, the reaction proceeded only above 800 °C, which was too high for many applications. Thus, ALD saturation and elemental Si (or Ge) were achieved by using atomic H for reduction at approximately 540 °C [122]. Still later, others utilized the same reaction principle to deposit non-Group IV elements, and produced elemental Ti and Ta [122].

$$SiCl_4 + {}^*H \rightarrow Si\text{-}H + HCl$$

$$TiCl_4 + {}^*H \rightarrow Ti\text{-}H + HCl$$

where * indicates a radical or plasma environment. Hence, plasma-assisted ALD may occur by direct plasma, remote apparatus configurations, or combinations thereof.

2) Depositing Metal Compounds

Plasma-assisted metal precursors use metal halides or MO compounds to make metal nitrides:

$$TiCl_4 + {}^*NH_3 \rightarrow TiN\text{-}H\ CO_2 + HCl$$

The reaction takes place at approximately 100 °C lower than its thermal counterpart. Metal nitrides or oxides may be formed using halide precursors and plasma containing either oxidants (O_3, H_2O) or nitridants (NH_3) [123], as well as otherwise nonreactive gases such as O_2 and N_2/H_2.

1.3.4.3 Advantages and Disadvantages

When using ALD, the film thickness will depend only on the number of reaction cycles, which in turn makes the thickness control accurate and simple. Unlike CVD, there is less need for reactant flux homogeneity, which gives a large area (large batch and easy scale-up) capability, excellent conformality and reproducibility, and also simplifies the use of solid precursors. The growth of different multilayer structures is also straightforward. Taken together, these advantages make the ALD method highly attractive for microelectronics, and notably for the manufacture of future-generation integrated circuits. The other advantages of ALD include the wide range of film materials that is available, as well as the high density and low impurity level. A lower deposition temperature may also be used in order not to affect sensitive substrates.

1) Advantages
   - Stoichiometric films with large area uniformity and 3-D conformality.
   - Precise thickness control.
   - Low-temperature deposition possible.
   - Gentle deposition process for sensitive substrates.

1) Disadvantages
   - Deposition rate slower than CVD.
   - The number of different material that can be deposited is fair compared to MBE.

1.4 Pattern Transfer

In the manufacture of nanostructures via top-down fabrication methods, one vital step is that of pattern transfer, where the pattern is defined through two steps: (i) lithographic resist patterning; and (ii) subsequent etching of the underlying material. Etching and lift-off represent the two ways of developing the transfer the pattern onto the substrates. As shown in Figure 1.20, the transfer can be either additive (lift-off) or subtractive (etch), although in practice the subtractive processes are preferred as they have a greater reliability and so a higher yield. Subtractive processing involves either etching or the removal of material; this can be achieved either by using suitable wet

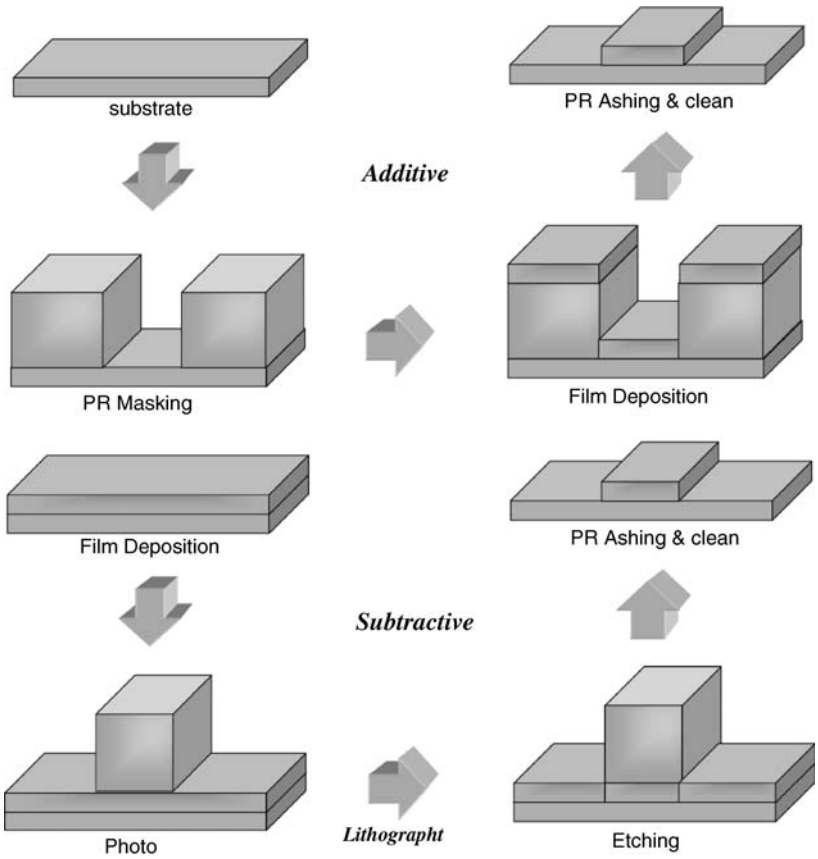


Figure 1.20 Pattern transfer by lift-off and etching way. PR = photoresist.

chemicals, or by dry etching in a vacuum system with the assistance of ions formed by an electrical discharge in a gas. Although the resist pattern can always be removed if found to be faulty on inspection, once the pattern has been transferred onto a solid material by etching, then any reworking is much more difficult, and often impossible.

1.4.1 Etch

Etching is the process of removing regions of the underlying material that are no longer protected by photoresist after development. There are two major types of etching: wet etching and dry etching.

1.4.1.1 Wet Etching

In wet etching, liquid chemicals or etchants are used to remove materials from the wafer, during which time the substrates are immersed in a reactive solution (etchant). As the layer to be removed is "etched away" by chemical reaction or by dissolution, the reaction products must be soluble so that they can be carried away by the etchant solution.

A basic wet etching process may be broken down into three basic steps: (i) diffusion of the etchant to the surface for removal; (ii) reaction between the etchant and the material being removed; and (iii) diffusion of the reaction byproducts from the reacted surface. The mechanisms fall into two major categories:

- metal etching (electron transfer):

$$\mathrm{M\,(s)} \rightarrow \mathrm{M}^{n+}\,\mathrm{(aq)} + ne^{-}$$

- insulator etching (acid–base reaction):

$$\mathrm{SiO_2} + 6\mathrm{HF} \rightarrow \mathrm{H_2SiOF_6\,(aq)} + 2\mathrm{H_2O}$$

Wet etching has found widespread use because of its special advantages of: (i) low cost; (ii) high reliability; (iii) high throughput; (iv) excellent selectivity in most cases with respect to both mask and substrate materials; (v) greater ease of use; (vi) higher reproducibility; and (vii) better efficiency in the use of etchants. Many of the materials used in microelectronics can be etched using wet etching methods.

Silicon Dioxide Etch HF-based etchants are widely used for etching silicon dioxide; hence, for pure HF-etching the overall reaction could be described as [124, 125]:

$$\mathrm{SiO_2} + 6\,\mathrm{HF} \rightarrow \mathrm{H_2} + \mathrm{SiF_6} + 2\mathrm{H_2O}.$$

Here, a 5 : 1 buffered hydrofluoric acid (BHF) solution (also known as buffered oxide etch, BOE) is a commonly used SiO_2 etchant formulation, and the reaction involved in the process is:

$$\mathrm{SiO_2} + 4\mathrm{HF} + 2\mathrm{NH_4F} \rightarrow (\mathrm{NH_4})_2\mathrm{SiF_6} + 2\mathrm{H_2O}$$

$$\mathrm{SiO_2} + 3\mathrm{HF_2}^{-} + \mathrm{H}^{+} \rightarrow \mathrm{SiF_6^{2-}} + 2\mathrm{H_2O}.$$

Here, "5 : 1" refers to five parts by weight of 40 wt% ammonium fluoride (the "buffer") to one part by weight of 49 wt% HF; this results in a total of about 33% NH_4F and 8.3% HF by weight [126], and the pH-value is about 3. HF is a weak acid, and except when present in very small concentrations it does not completely dissociate into H^+ and F^- ions in water [127]. Judge [128] and Deckert [129] have each shown the etch rate of both silicon dioxide and silicon nitride to increase linearly with the concentrations of both HF and HF_2^-. However, for concentrations below 10 M, whilst being independent of the concentration of F^- ions alone, the HF_2^- complex attacks oxides much faster than HF. Thus, the etch rate increases more than linearly with respect to the HF concentration.

Poly-Silicon Etch Silicon is usually wet-etched using a mixture of nitric acid (HNO_3) and HF [2, 24, 25], which may be masked by the photoresist.

A simplified description of the reaction is that the HNO_3 in the solution oxidizes the silicon, followed by the oxidized compound being etched by the HF (formed from the fluoride ions in this acidic solution). Many metal-etches also remove material in a two-step manner, the overall reaction being [125, 130]:

$$3Si + 4HNO_3 + 18HF \rightarrow 3H_2SiF_6 + 4NO(g) + 8H_2O$$

Although the etching of silicon with HNO_3 and HF is an isotropic approach, on occasion it does not meet the requirements for microelectronics. Orientation-dependent silicon wet etchants have also been developed. For example, KOH is used for the orientation-dependent etching (ODE) of single-crystal silicon. ODEs attack {111}-type planes, which have a high bond density, much more slowly than other planes [124, 131]. Occasionally, isopropyl alcohol may be added to the KOH solutions; this decreases the etch rate but improves uniformity, thus reducing the requirement for stirring [132]. In this case the gross reaction is:

$$Si + 2OH^- \rightarrow SiO_2(OH)_2^{2-} + 2H_2(g)$$

Metal Wet Etch The wet etching of aluminum and aluminum alloy layers may be achieved using slightly heated (35–45 °C) solutions of phosphoric acid, acetic acid, nitric acid, and water. This is a multistep etch process, which could also be masked by a photoresist. The aluminum is first oxidized by the nitric acid, while the phosphoric acid and water simultaneously etch the resulting oxide [133].

A solution with a constituent of H_2O_2 (30%; i.e., hydrogen peroxide, 30% by weight) is used to wet-etch tungsten and its alloys (where HF is the active ingredient in the etchant) and also etched oxides. Raising the proportion of HF in the solution causes an increase in the etch rate. In this etching, a film of tungsten oxide is formed that is dissolved in the hydrogen peroxide [134]. This etchant can also be used to etch tungsten–titanium alloys, but not pure titanium.

A wet process is also used for the cleaning of wafers before use. For example, Piranha – a hot solution of H_2SO_4 and H_2O_2 mixed in any ratio – has been used for decades for wafer cleaning [135–137]. With a lower ratio of H_2SO_4 to H_2O_2, as for other acidic hydrogen peroxide solutions, Piranha will first strip off the photoresist

and any other organics by oxidizing them, and then remove any metals by forming complexes that remain in solution [138, 139]. This does not adversely affect the silicon dioxide and silicon nitride, and has only a minor effect on the bare silicon of forming a thin layer of hydrous silicon oxide. However, after the Piranha clean and rinse the silicon oxide can easily be removed with quick (10 s) dip into 10 : 1 or 25 : 1 HF. Other materials on the wafer may be wet-etched by using the appropriate etching solutions.

As with any process, wet etching has its own certain disadvantages. Typically, wet etching is isotropic in nature, and the etched feature will have curved walls and its width will differ from that of the opening in the resist. If the aspect ratio (the ratio of depth to width) of the desired feature is small, then the isotopic nature of the etching is often not important; however, in the closely packed structures found in very large-scale integration (VLSI) integrated circuits it is not acceptable. An additional problem is that, after wet chemical etching, disposal of the partly used reagent may raise environmental issues. A further problem is that monolayer-thick layers of hydrocarbons can inhibit wet etching, such that the *in situ* control of etch depth is made difficult.

Silicon of ⟨110⟩ orientation offers an interesting possibility for the anisotropic wet etching of perfectly vertical walls when the mask is aligned so that slow-etching (111) planes form the sidewalls. The side walls and bottom surfaces shown in Figure 1.21b contain a large number of facets when etched at 70 °C in KOH–water without the addition of 2-propanol; however, the microchannel sidewalls are clearly still vertical. In contrast, Figure 1.21a shows a less-pronounced surface structure with a definite slope in the side walls when etched in a solution containing 2-propanol [140].

1.4.1.2 **Dry Etching**

Unlike wet etching processes, dry etching does not utilize any liquid chemicals or etchants to remove materials from the wafer; rather, the substrates are immersed in a reactive gas (plasma), and in the process only volatile byproducts are generated. As with wet etching, dry etching also follows the resist mask patterns on the wafer; that

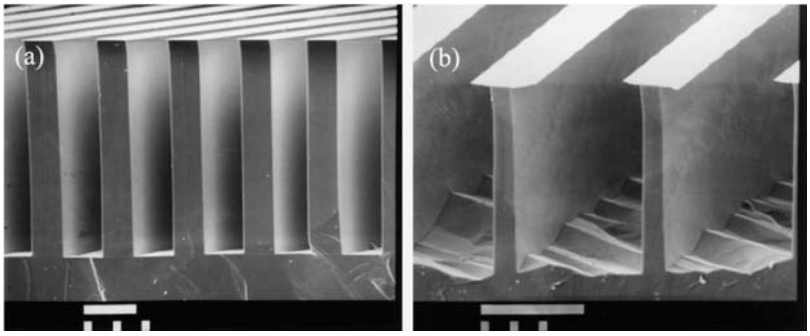


Figure 1.21 Scanning electron microscopy images. (a) 410 mm-deep microchannels etched at 85 °C with a Si_3N_4 masking layer over the thermal SiO_2 layer; (b) 170 μm-deep microchannels etched at 70 °C with a 1 mm-thick SiO_2 masking layer, showing side hanging of the masking layer over the microchannels.

is, it only etches away materials that are not covered by mask material (and are therefore exposed to its etching species), while leaving areas covered by the masks almost (but not perfectly) intact. These masks were previously deposited on the wafer using a wafer fabrication step known as "lithography."

Dry etching may be accomplished by any of the following: (i) through chemical reactions that consume the material, using chemically reactive gases or plasma; (ii) by physical removal of the material, usually by momentum transfer; or (iii) by a combination of both physical removal and chemical reactions.

In a plasma discharge, a number of different mechanisms for gas-phase reactions are operative. Discharge generates both ions and excited neutrals, and both are important for etching.

$$\text{Ionization}: \ e^- + Ar \rightarrow 2e^- + Ar^+$$

$$\text{Excitation}: \ e^- + O_2 \rightarrow e^- + O_2{}^*$$

$$\text{Dissociation}: \ e^- + SF_6 \rightarrow e^- + SF_5{}^* + F^*$$

Silicon Dioxide Etch Fluorocarbon–Plasma ($CF_4 + CHF_3 + He$) is used for silicon dioxide dry etching. In this case, it appears that the CF_x ($x \leq 3$) radicals are chemisorbed onto the SiO_2 and become dissociated; the radicals then supply carbon to form CO, CO_2, and COF_2 gases from the oxygen in the film. They also supply fluorine to form SiF_4 gas [141]. The overall reactions that occur are as followings [124]:

$$3SiO_2 + 4CF_3 \rightarrow 2CO + CO_2 + 3SiF_4$$

$$SiO_2 + 2CHF_2 \rightarrow 2CO + H_2 + SiF_4$$

Plasma HF-vapor is another method using for the dry etching of silicon dioxide.

As with liquid-based HF etches, the HF vapor etches silicon dioxide and has been used to remove native oxide from silicon before the growth of epitaxial silicon and other processes, such as the XeF_2 etching of silicon. In the process, the HF/H_2O vapor condenses into droplets on the surfaces of the oxide samples during a 1-minute etch, such that a faster etching is caused where the droplets had formed. The related reaction in this process is [142]:

$$SiO_2 + 4HF \rightarrow SiF_4 + H_2O.$$

Silicon Etch Fluorine-, chlorine-, and bromine-based processes represent standards for silicon etching, and result in reaction products of SiF_4, $SiCl_4$ and $SiBr_4$, respectively. Fluorine-based processes are safer to use, but are seldom fully anisotropic, while chlorine-based processes result in vertical sidewalls inherently (the same applies to bromine-based processes). Importantly, both chlorine and bromine are highly toxic, and it is essential that the equipment used for Cl_2 or HBr etching must be equipped with a loadlock.

First synthesized in 1962 [143], XeF_2 has the unusual capability to etch silicon at a significant rate, without requiring a plasma to generate reactive species, and has been used for silicon etching [143, 144]. Notably, XeF_2 has one major advantage over wet silicon etchants in that will gently etch without the application of any forces. In addition, it has the advantage over plasma etching of being extremely selective over almost all of the traditional masking layers, including silicon dioxide, some silicon nitrides, and photoresists. XeF_2 has also been used to micromachine free-standing structures made from aluminum and polysilicon protected by a layer of oxide [145].

During the etch process, the XeF_2 molecules are physisorbed onto the silicon surface and dissociate to release volatile xenon atoms, while the fluorine atoms remain to react with the silicon to form volatile SiF_4. The overall reaction is:

$$Si + 2XeF_2 \rightarrow SiF_4 + 2Xe.$$

The etching of silicon with plasma represents an anisotropic approach, normally using Cl_2 + He, HBr + $C1_2$, as the plasma source. When Cl_2 + He is applied to etch silicon, a previous SF_6 step is typically used to break through the native oxide. In this case, the chlorine atoms are chemisorbed one at a time onto the silicon surface, eventually forming volatile $SiCl_4$ [146]. This method has been used to etch 80 pm-deep trenches with fairly vertical sidewalls [147]; the overall reaction is:

$$4Cl + Si \rightarrow SiCl_4.$$

HBr + $C1_2$ plasma represents yet another anisotropic silicon plasma etch source, but this has a better selectivity of silicon over oxide, whereby the bromine atoms most likely react with silicon in a manner similar to chlorine (as described above).

The chlorine etching of undoped silicon occurs very slowly in the absence of ion bombardment [147]. Unlike F-atom silicon etches, Cl- and Br-based etches tend to be vertical [148].

Silicon Nitride Etches Fluorine-Atom-Plasma (SF_6 + He or CF_4 + CHF_3 + He) is used to plasma-etch silicon nitride, and this can be masked with a photoresist. The etch is anisotropic and results in fairly vertical sidewalls. In this case, the fluorine atoms are adsorbed onto the surface one at a time, in a surface reaction, and volatile products are formed. The overall reaction is [124]:

$$12FCl + Si_3N_4 \rightarrow SiF_4 + 2N_2.$$

Plasma Metal Etches Cl_2 + $BC1_3$ + $CHCl_3$ + N_2 is used to dry etch aluminum, which is an anisotropic etch due to the side-wall inhibitor formed from the $CHC1_3$ [149]. Due to poor selectivity, when etching the thick layers of Al, a thicker photoresist, a plasma-hardened photoresist, or a more durable masking layer must be used. Usually, a higher temperature is used to keep the etch product volatile so that it leaves the wafer [124] and does not coat the chamber or exhaust the plumbing. Cl_2 rather than Cl appears to be the main etchant [149], and the etch product becomes $AlC1_3$ at higher temperatures [149, 150]. The dominant overall reaction

below 200 °C is:

$$12Al + 3Cl_2 \rightarrow Al_2Cl_3$$

SF_6 is used for tungsten plasma etching, with the etch process function being fairly isotropic. CF_4 is often added to the feed gas, and increases the anisotropy as side-wall polymers form, although the etch rate will be slowed down. In this case, the chuck can be heated to enhance the etch rate. The overall reaction is:

$$6F + W \rightarrow WF_6$$

Shown in Figure 1.22 are the SEM images of surface structures of silicon that have been etched with SF_6/O_2 plasma in an RIE etching manner. The clearly visible difference between these figures relates to the fact that the substrate in Figure 1.22a was pretreated with BHF and acetone, whereas that in Figure 1.22b was pretreated only with BHF. The width of the needles is almost 200 nm. The morphology of the silicon surfaces etched in Figure 1.22b was found to be uniform, and the formation of columnar nanostructures with diameters <100 nm and depths >300 nm were possible [151].

The dry etching method has the following advantages:

- The elimination of handling dangerous acids and solvents.
- The use of only small amounts of chemicals.
- The production of isotropic or anisotropic etch profiles.
- Directional etching can be achieved, without using the crystal orientation of Si.
- Lithographically defined photoresist patterns are faithfully transferred into the underlying layers
- High resolution and cleanliness.
- Less undercutting.
- No unintentional prolongation of etching.
- Better process control.
- Ease of automation (e.g., cassette loading).

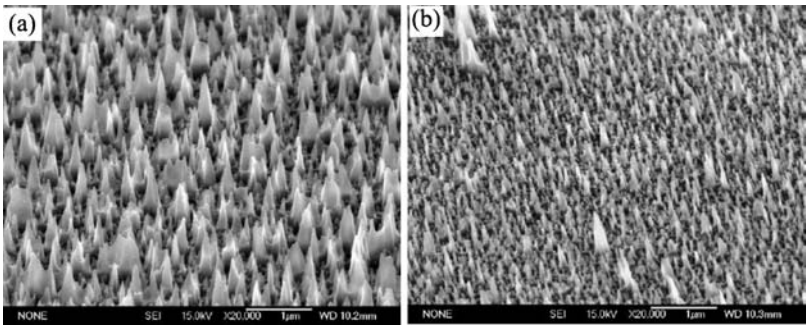


Figure 1.22 Scanning electron microscopy images of needle-like nanostructures of the silicon surface by using RIE in a parallel-plate plasma system. (a) Pretreatment with BHF and acetone; (b) Pretreatment only with BHF.

1.4.1.3 A Comparison of Wet and Plasma Etching

Wet etching is usually isotropic (which is desirable in some cases), it may provide a selectivity that is dependent on the crystallographic direction, and it can be very selective over the masking and underlying layers. In contrast, plasma etching uses fresh chemicals for each etch (this results in a less chemical-related etch-rate variability) and it can be vertically anisotropic (as well as isotropic), thus allowing the patterning of narrow lines. One drawback of wet etching is that, when removing a sacrificial layer in micromachining, there will be a capillary-force pull down of any free-standing structures [152]. However, this can be overcome by using a supercritical-liquid drying process [153] or by switching to a dry-etched sacrificial layer [154, 155].

In many applications, the choice of wet versus plasma etching is a question of convenience – whether certain equipment or an etch bath is available, or a suitable masking material is at hand. However, when sloped etch profiles are required, or when under-cutting is needed, then isotropic etching must be used. One particular benefit is that the isotropic wet etching of silicon can be achieved at fairly high rates, at microns or even tens of microns per minute.

1.4.2 Lift-Off Processes

The pattern-transfer technique – *lift-off*– refers to the process of creating patterns on the wafer surface through an additive process, as opposed to the more familiar patterning techniques that involve subtractive processes, such as etching. Lift-off is most commonly employed in patterning metal films for interconnections.

The steps of the technique are shown schematically in Figure 1.20. The resist is first exposed to radiation via the pattern-carrying mask, and the exposed areas of the resist are then developed (as shown in Figure 1.20). A film is then deposited over the resist and substrate. Prior to film deposition – and particularly for sputtering or evaporation processes – a "post-develop bake" is recommended, which will drive off any excess solvent so that there will be less out-gassing during the film deposition. The film thickness must be smaller than that of the resist. By using an appropriate solvent (such as acetone), the remaining parts of the resist and the deposited film atop these parts of the resist can be lifted off (see Figure 1.20). The lift-off technique is capable of high resolution, and often used for the fabrication of discrete devices.

Depending on the type of lift-off process used, patterns can be defined with extremely high fidelity and for very fine geometries. Lift-off, for example, is the process of choice for patterning electron-beam-written metal lines, because the film remains only where the photoresist has been cleared. The defect modes are the opposite what might be expected for etching films, as the defects may occur in the underlaying photoresist layer; for example, particles that underlay the photoresist may lead to open or other unwanted shapes on the substrate, whereas in case of metal lift-off the scratches may lead to unwanted areas of the metal layer remaining on the wafer.

Any deposited film can be lifted-off, provided that:

- During film deposition, the substrate does not reach temperatures that are high enough to burn the photoresist.
- The film quality is not absolutely critical; a photoresist will outgas very slightly in vacuum systems, which may adversely affect the quality of the deposited film.
- The adhesion of the deposited film on the substrate is very good.
- The film can be easily wetted by the solvent.
- The film is thin enough and/or grainy enough to allow solvent to seep underneath; the thickness of the film being lifted off should be preferably kept at less than one-third of the total photoresist thickness.
- The film is not elastic, and is thin and/or brittle enough to be torn along the adhesion lines.

References

1 Thompson, L.F. (1983) *Introduction to Microlithography* (eds L.F. Thompson, C.G. Willson, and M.J. Bowden), The American Chemical Society, Washington, DC, p. 1.

2 Moreau, W.M. (1988) *Semiconductor Lithography: Principles and Materials*, Plenum, New York.

3 Suzuki, K., Matsui, S., and Ochiai, Y. (2000) *Sub-Half-Micron Lithography for ULSIs*, Cambridge University Press, Cambridge.

4 Gentili, M., Giovannella, C., and Selci, S. (1993) *Nanolithography: A Borderland between STM, EB, IS, and X-Ray Lithographies*, Kluwer, Dordrecht, The Netherlands.

5 Brambley, D., Martin, B., and Prewett, F.D. (1994) *Adv. Muter: Opt. Electron.*, **4**, 55.

6 Bernard, F. (2002) *Microelectron. Eng.*, **61–62**, 11.

7 Moore, G.E. (1995) *Proc. SPIE*, **2440**, 2–17.

8 Mack, C. (1996) *Opt. Photo. News*, **April**, 29–33.

9 Dammel, R. (1993) *Diazonaphthoquinone-Based Resists*, vol. **TT11**, SPIE Optical Engineering Press.

10 Kumar, A., Abbot, N.A., Kim, E., Biebuyck, H.A., and Whitesides, G.M. (1995) *Acc. Chem. Res.*, **28**, 219.

11 Ulman, A. (1991) *An Introduction to Ultrathin Organic Films: From Langmuir-Blodgett to Self-Assembly*, Academic Press, San Diego, CA.

12 Huang, J., Dahlgren, D.A., and Hemminger, J.C. (1994) *Langmuir*, **10**, 626.

13 Chan, K.C., Kim, T., Schoer, J.K., and Crooks, R.M. (1995) *J. Am. Chem. Soc.*, **117**, 5875.

14 Bruning, J.H. (1997) *Proc. SPIE*, **3051**, 14.

15 Levinson, H.J. and Arnold, W.H. (1997) Optical lithography, *Handbook of Microlithography and Microfabrication*, vol. **1** (ed. P. Rai-Choudhury), SPIE, Bellingham, WA.

16 Sheats, J.R. and Smith, B.W.(eds) (1998) *Microlithography: Science and Technology*, Marcel Dekker, New York.

17 Bruning, J.H. (1980) *J. Vac. Sci. Technol.*, **17**, 1147.

18 Smith, B.W. (2002) *J. Microlith. Microfab. Microsyst.*, **1**, 95.

19 Levenson, M.D., Viswanathan, N.S., and Simpson, R.A. (1982) *IEEE Trans. Electron Devices*, **ED29**, 1812.

20 Shibuya, M. (1987) Projection master for transmitted illuminated. Japanese Patent Gazette No. 62-50811

21 Smith, H.I., Anderson, E.H., and Schattenburg, M.L. (1989) Lithography mask with a p-phase shifting attenuator. U.S. Patent 4,890,309

22 Flanders, D.C. and Smith, H.I. (1982) Spatial period division exposing. U.S. Patent 4,360,586

23 Tananka, T., Uchino, S., Hasegawa, N., Yamanaka, T., Terasawa, T., and Okazaki, S. (1991) *Jpn. J. Appl. Phys. Part I*, **30**, 1131.

24 Rogers, J.A., Paul, K.E., Jackman, R.J., and Whitesides, G.W. (1998) *J. Vac. Sci. Technol. B*, **16**, 59.

25 Aizenberg, J., Rogers, J.A., Paul, K.E., and Whitesides, G.M. (1998) *Appl. Opt.*, **37**, 2145.

26 Aizenberg, J.J., Rogers, A., Paul, K.E., and Whitesides, G.M. (1997) *Appl. Phys. Lett.*, **71**, 3773.

27 Terazawa, T., Hasegawa, N., Fukuda, H., and Katagiri, S. (1991) *Jpn. J. Appl. Phys.*, **30**, 2991.

28 Tritchkov, A., Jeong, S., and Kenyon, C. (2005) *Proc. SPIE*, **215**, 5754.

29 Perlitz, S., Buttgereit, U., Scherubl, T., Seidel, D., Lee, K.M., and Tavassoli, M. (2007) *Proc. SPIE*, **6607**, 66070Z.

30 Luo, Y. and Misra, V. (2006) *Nanotechnology*, **17**, 4909.

31 McCord, M.A. and Rooks, M.J. (2000) *Handbook of Microlithography, Micromachining and Microfabrication*, (ed. P. Rai-Coudhury), vol. 2, SPIE, Washington.

32 Brewer, G.R. (1980) *Electron-Beam Technology in Microelectronic Fabrication*, Academic Press, New York.

33 Chen, W. and Ahmed, H. (1993) *Appl. Phys. Lett.*, **62**, 1499.

34 Craighead, H.G., Howard, R.E., Jackel, L.D., and Mankievich, P.M. (1983) *Appl. Phys. Lett.*, **42**, 38.

35 Chou, S.Y. (1997) *Proc. IEEE*, **85**, 652.

36 Vieu, C., Carcenac, F., Pepin, A., Chen, Y., Mejias, M., Lebib, A., Manin-Ferlazzo, L., Couraud, L., and Lunois, H. (2000) *Appl. Surf. Sci.*, **164**, 111.

37 Yesin, S., Hasko, D.G., and Ahmed, H. (2001) *Appl. Phys. Lett.*, **78**, 2760.

38 Rosenfield, M.G., Thomson, M.G.R., Coane, P.J., Kwietniak, K.T., Keller, J., Klaus, D.P., Volant, R.P., Blair, C.R., Tremaine, K.S., Newman, T.H., and Hohn, F.J. (1993) *J. Vac. Sci. Technol. B*, **11**, 2615.

39 Rishton, S.A., Schmid, H., Kern, D.P., Luhn, H.E., Chang, T.H.P., Sai-Halasz, G.A., Wordeman, M.R., Ganin, E., and Polcari, M. (1988) *J. Vac. Sci. Technol. B*, **6**, 140.

40 Fujita, J., Ohnishi, Y., Ochinai, Y., Nomura, E., and Matsui F S. (1996) *J. Vac. Sci. Technol. B* **14**, 4272.

41 Thompson, L.F. and Bowden, M.J. (1983) *Introduction to Microlithography* (eds L.F. Thompson, C.G. Willson, and M.J. Bowden), The American Chemical Society, Washington, DC, p. 15.

42 Spears, D.L. and Smith, H.I. (1972) *Solid State Technol.*, **15**, 21.

43 Smith, H.I. (1996) *Phys. Scr.*, **T61**, 2631.

44 Perlman, M.L., Rowe, E.M., and Watson, R.E. (1974) *Phys. Today*, **27**, 30.

45 Kunz, C. (1976) *Physik BI.*, **32**, 55.

46 Dagneaux, P., Depautcx, C., Dhez, P., Durup, J., Farge, Y., Fourme, R., Guyon, P.M., Jaegle, P., Leach, S., Lopez-Delgado, R., Morel, G., Pinchaux, R., Thiry, P., Vermeil, C., and Wuilleumier, F. (1975) *Ann. Phys. (Paris)*, **9**, 9.

47 Kunz, C. (1974) *Vacuum Ultraviolet Radiation Physics* (eds E.E. Koch, R. Haensel, and C. Kunz), Vieweg, Braunschweig, p. 753.

48 Kitayama, T., Itoga, K., Watanabe, Y., and Uzawa, S. (2000) *J. Vac. Sci. Technol. B*, **18**, 2950.

49 Smith, H.I., Spears, D.L., and Bemacki, S.E. (1973) *J. Vac. Sci. Technol.*, **10**, 913.

50 Bemacki, S.E. and Smith, H.I. (1974) *Proceedings 6th International Conference Electron and Ion Beam Science and Technology, 1974, San Francisco, CA* (ed. R. Bakish), The Electrochemical Society, Princeton, NJ.

51 Bemacki, S.E. and Smith, H.I. (1975) *IEEE Trans. Elec. Dev.*, **ED22**, 421.

52 Chou, S.Y., Smith, H.I., and Antoniadis, D.A. (1985) *J. Vac. Sci. Technol. B.*, **3**, 1587.

53 Shahidi, G.G., Antoniadis, D.A., and Smith, H.I. (1988) *IEEE Elect. Dev Lett.*, **EDG9**, 94.

54 Ismail, K., Bagwell, P.F., Orlando, T.P., Antoniadis, D.A., and Smith, H.I. (1991) *Proc. IEEE*, **79**, 1106.

55 Scott-Thomas, J.H.F., Field, S.B., Kastner, M.A., Smith, H.I., and Antoniadis, D.A. (1989) *Phys. Rev. Lett.*, **62**, 583.

56 Simon, G., Haghiri-Gosnet, A.M., Bourneix, J., Decanini, D., Chen, Y., Rousseaux, F., Launios, H., and Vidal, B. (1997) *J. Vac. Sci. Technol. B*, **15**, 2489.

57 Gerlach, R. and Utlaut, M. (2001) *Proc. SPIE, Charged Particle Detection, Diagnostics, and Imaging*, **4510**, 96.

58 Prewett, P.D. and Mair, G.L.R.(eds) (1991) *Focused Ion Beams from Liquid Metal Ion Sources*, John Wiley & Sons, New York.

59 Hall, T.M., Wagner, A., and Thompson, L.F. (1979) *J. Vac. Sci. Technol.*, **16**, 1889.

60 Seliger, R.L., Kubena, R.L., Olney, R.D., Ward, J.W., and Wang, V. (1979) *J. Vac. Sci. Technol.*, **16**, 1610.

61 Gierak, I., Septierl, A., and Vieu, C. (1999) *Nucl. Instrum. Methods Phys. Res. Sect. A*, **A421**, 91.

62 Matsui, S. and Ochiai, Y. (1996) *Nanotechnology*, **7**, 247.

63 Gamo, K. (1996) *Microelectron. Eng.*, **32**, 159.

64 Morimoto, H., Sasaki, Y., Saitoh, K., Watakabe, Y., and Kato, T. (1986) *Microelectron. Eng.*, **4**, 163.

65 Melngailis, J. (1987) *J. Vac. Sci. Technol.*, **B5**, 469.

66 Zachariasse, J. and Walker J. (1997) *Microelectron. Eng.*, **35**, 63.

67 Adesida, I. (1983) *Nucl. Instrum. Methods*, **209–210**, 79.

68 Harthey, M., Shaver, D., Shepard, M., Melngailis, J., Medvedev, V., and Robinson, W. (1991) *J. Vac. Sci. Technol.*, **B9**, 3432.

69 Herbert, P., Braddell, J., MacKenzie, S., Woodham, R., and Cleaver, J. (1994) *Microelectron. Eng.*, **23**, 263.

70 Kuwano, H. (1984) *J. Appl. Phys.*, **55**, 1149.

71 Ohring, M. (1992) *The Materials Science of Thin Films*, Academic Press, San Diego, CA.

72 Vossen, J.L. and Kern, W.(eds) (1991) *Thin Film Processes II*, Academic Press, San Diego, CA.

73 Nalwa, H.S.(ed.) (2002) *Handbook of thin Film Materials, Vol. I: Deposition and Processing of Thin Films*, Academic Press, San Diego, CA.

74 Chambers Scott A. (2000) *Surf. Sci. Rep*, **39**, 105.

75 Faraday, M. (1857) *Philos. Trans.*, **147**, 145.

76 Holland, L. (1957) *Vacuum Deposition of Thin Films*, Chapman & Hall, London.

77 Deshpandey, C.V. and Bunshah, R.F. (1991) *Thin Film Processes II* (eds J.L. Vossenand W. Kern), Academic Press, San Diego, CA.

78 Glang, R. (1970) *Vacuum Evaporation, Handbook of Thin Film Technology* (eds L.I. Maisseland R. Glang), McGraw-Hill, pp. 1–26.

79 Pulker, H.K. (1984) Chapter 6, Film Formation Methods: Coatings on Glass, in *Thin Films: Science and Technology Series*, No. 6, Ch. 6 Elsevier.

80 Koleshko, V.M. (1987) *Vacuum*, **36**, 689.

81 Perry, A.J. (1981) *Wear*, **67**, 381.

82 Buckley, D.H. (1981) *Surface Effects in Adhesion, Friction, Wear, and Lubrication, Tribology Series 5*, Elsevier, p. 613.

83 Herman, M.A. and Sitter, H. (1989) *Molecular Beam Epitaxy-Fundamentals and Current Status*, Springer-Verlag, Berlin.

84 Kasper, E. and Bean, J.C.(eds) (1988) *Silicon-Molecular Beam Epitaxy I and II*, CRC Press, Boca Raton, FL.

85 Parker, E.H.C.(ed.) (1985) *The Technology and Physics of Molecular Beam Epitaxy*, Plenum Press, New York.

86 Wehner, G.K. (1955) *Adv. Electron. Electron Physics*, **7**, 239.

87 Kay, E. (1962) *Adv. Electron. Electron Physics*, **17**, 245.

88 Maissel, L.I. (1966) The deposition of thin films by cathode sputtering, *Physics of Thin Films*, vol. 3 (eds G. Hassand R.E. Thun), Academic Press, p. 61.

89 Holland, L. (1961) Cathodic sputtering, in *Vacuum Deposition of Thin Films*, Ch. 14, Chapman & Hall.

90 Nowicki, R.S. (1982) *Solid State Technol.*, **21**, 127.

91 Gadepally, K.V. and Hawk, R.M. (1989) *Proc. Arkansas Acad. Sci.*, **43**, 29.

92 Morosanu, C.E. (1990) *Thin Films by Chemical Vapor Deposition*, Elsevier, Amsterdam.

93 Cooke, M.J. (1985) *Vacuum*, **35**, 67.

94 Pierson, H.O. (1992) *Handbook of Chemical Vapor Deposition: Principles, Technology and Applications*, Noyes Publications.

95 Jensen, K.F. and Kern, W. (1991) *Thin Film Processes II* (eds J.L. Vossenand W. Kern), Academic Press, San Diego, CA.

96 Hitchman, M.L. and Jensen, K.F.(eds) (1993) *CVD Principles and Applications*, Academic Press, San Diego.

97 Ser, P., Kalck, P., and Feurer, R. (2002) *Chem. Rev.*, **102**, 3085.

98 Choy, K.L. (2003) *Prog. Mater. Sci.*, **48**, 57.

99 Herrmann, C.F., DelRio, F.W., George, S.M., and Bright, V.M. (2005) Micromachining and Microfabrication Process Technology, (eds M.-A. Maherand H.D. Stewart), *Proceedings SPIE*, vol. 5715, SPIE, Bellingham, WA. Available at: http://ald.colorado.edu/J_Phys_Chem_100.pdf.

100 Wikipedia: The Free Encyclopedia (2006) Atomic Layer Deposition. Wikimedia Foundation, 24 April 2006.

101 Ritala, M. and Leskela, M. (2002) *Handbook of Thin Film Materials, 61.1: Deposition and Processing of Thin Films* (ed. H.S. Nalwa), Academic Press, San Diego, CA, p. 103.

102 Ritala, M. and Leskela, M. (1999) *Nanotechnology*, **10**, 19.

103 Suntola, T. and Simpson, M. (1990) *Atomic Layer Epitaxy* (eds T. Suntolaand M. Simpson), Blackie, New York, pp. 3–5.

104 Ahonen, M. and Pessa, M. (1980) *Thin Solid Films*, **65**, 301.

105 Pessa, M., Makela, R., and Suntola, T. (1981) *Appl. Phys. Lett.*, **38**, 131.

106 Suntola, T. and Hyvarinen, J. (1985) *Annu. Rev. Muter. Sci.*, **15**, 177.

107 Suntola, T. and Antson, J. (1977) Methods for producing compound thin films. U.S. Patent No. 4058430

108 Ritala, M. and Leskela, M. (2002) Deposition and processing of thin films, in *Handbook of Thin Film Materials*, vol. 1 (ed. H. Nalwa), Academic Press, San Diego, p. 103.

109 Kim, H. (2003) *J. Vac. Sci. Technol. B*, **26**, 2231.

110 Ott, A.W., Klaus, J.W., Johnson, J.M., and George, S.M. (1997) *Thin Solid Films*, **292**, 135.

111 Aarik, J., Aidla, A., Mändar, H., Uustare, T., Kukli, K., and Schuisky, M. (2001) *Appl. Surf. Sci.*, **173**, 15.

112 Ritala, M. and Leskelä, M. (1994) *Appl. Surf. Sci.*, **75**, 330.

113 Zang, H., Solanki, R., Roberds, B., Bai, G., and Banerjee, I. (2000) *J. Appl. Phys.*, **87**, 1921.

114 Yun, S.-J., Kang, J.S., Paek, M.C., and Nam, K.S. (1998) *J. Korean Phys. Soc.*, **33**, S170.

115 Aarik, J., Aidla, A., Uustare, T., and Sammelselg, V. (1995) *J. Cryst. Growth.*, **148**, 268.

116 Sammelselg, V., Rosental, A., Tarre, A., Niinisto, L., Heiskanen, K., Ilmonen, K., Johansson, L.-S., and Uustare, T. (1998) *Appl. Surf. Sci.*, **134**, 78.

117 Kim, Y.K., Lee, S.H., Choi, S.J., Park, H.B., See, Y.D., and Chin, K.H. (2000) *IEDM Tech. Dig.*, 369, (IEEE Cat No: 00CH38138).

118 Yokoyama, S., Ohba, K., and Nakajima, A. (2001) *Appl. Phys. Lett.*, **79**, 617.

119 Klaus, J.W., Ferro, S.J., and George, S.M. (2000) *Thin Solid Films*, **360**, 145.

120 Aaltonen, T., Alen, P., Ritala, M., and Leskela, M. (2003) *Chem. Vap. Deposition*, **9**, 45.

121 Nishizawa, J., Aoki, K., Suzuki, S., and Kikuchi, K. (1990) *J. Cryst. Growth*, **99**, 502.

122 Rossnagel, S., Sherman, A., and Turner, F. (2000) *J. Vac. Sci. Technol.*, **B18**, 2016.

123 Londergan, A. (2002) *Rapid Thermal and Other Short Time Processes III, Proceedings, Vol. 2002–11*, The Electrochemical Society, Inc.

124 Runyan, W.R. and Bean, K.E. (1990) *Semiconductor Integrated Circuit Processing Technology*, Addison-Wesley, Reading, MA.

125 Ghandi, S.K. (1983) Silicon and gallium arsenide, *VLSI Fabrication Principles*, John Wiley & Sons, New York.

126 J.T. Baker, Inc . (1993) Product Specifications for Product No. 5192, Buffered Oxide Etch, 5:1, J.T. Baker, Inc., Phillipsburg, NJ, Tech. support, 7 June 1995.

127 Kikyuama, H., Miki, N., Saka, K., Takano, J., Kawanabe, I., Miyashita, M., and Ohmi T. (1991) *IEEE Trans. Semicond. Manuf.*, **4**, 26.

128 Judge, J.S. (1971) *J. Electrochem. Soc.*, **118**, 1772.

129 Deckert, C.A. (1978) *J. Electrochem. Soc.*, **125**, 320.

130 Turner, D.R. (1960) *J. Electrochem. Soc.*, **107**, 810.

131 Kendall, D.L. (1990) *J. Vac. Sci. Technol. A*, **8**, 3598.

132 Amulya, K.N. and Goldemberg, J. (1990) *J. Electrochem. Soc.*, **137**, 3612.

133 Elliot, D.J. (1989) *Integrated Circuit Fabrication Technology*, 2nd edn, McGraw-Hill, New York, p. 355.

134 van den Meerakker, J.E.A.M., Scholten, M., and van Oekel, J.J. (1992) *Thin Solid Films*, **208**, 237.

135 Wolf, S. and Tauber, R.N. (1986) *Silicon Processing for the VLSI Era*, vol. 1, Lattice, Sunset Beach, CA.

136 Yang, M.G. and Koliwad, K.M. (1975) *J. Electrochem. Soc.*, **122**, 675.

137 Pintchovski, F., Price, J.B., Tobin, P.J., Peavey, J., and Kobold, K. (1979) *J. Electrochem. Soc.*, **126**, 1428.

138 Kem, W. and Puotinen, D.A. (1970) *RCA Review*, **30**, 187.

139 Amick, J.A. (1976) *Solid State Technol.*, **19**, 47.

140 Dwivedi, V.K., and Ahmad, R.G.S. (2000) *Microelectron. J.*, **31**, 405.

141 Dwivedi, V.K., Gopal, R., and Ahmad, S. (2000) *Microelectron. J.*, **31**, 405.

142 Kuiper, A.E.T. and Lathouwers, E.G.C. (1992) *J. Electrochem. Soc.*, **139** (9), 2594–2599.

143 Oxtoby, D.W. and Nachtrieh, N.H. (1986) *Principles of Chemistry*, Saunders College Pub., Philadelphia, p. 728.

144 Winters, H.F. and Cobum, J.W. (1979) *Appl. Phys. Lett.*, **34**, 70.

145 Hoffman, E., Warneke, B., Kruglick, E., Weigold, J., and Pister, K.S.J. (1995) 3D structures with piezoresistive sensors in standard CMOS, Proceedings, IEEE Micro Electro-Mechanical Systems 1995, Amsterdam, The Netherlands, January-February, vol. 1, p. 288.

146 Manos, D.M. and Flamm, D.L.(eds) (1989) *Plasma Etching: An Introduction*, Academic, Boston.

147 Keller, C.G. and Howe, R.T. (1995) Nickel-filled thermally actuated hexsil tweezers, Technical Digest, 8th International Conference on Solid-state Sensors and Actuators (Transducers '95), Stockholm, Sweden, p. 376.

148 Rossnagel, S.M., Cnomo, J.J., and Westwood, W.D.(eds) (1990) *Handbook of Plasma Processing Technology*, Noyes, Park Ridge, NJ.

149 Lieberman, M.A. and Lichtenherg, A.J. (1994) *Principles of Plasma Discharges and Materials Processing*, John Wiley & Sons, New York.

150 Kem, W. and Deckert, C.A. (1978) Chemical etching, *Thin Film Processes*, (eds J.L. Vossenand W. Kem), Ch. V-1, Academic, New York, p. 413.

151 Jung, S., Kim, K., Park, D., Sohn, B.H., Jung, J.C., Zin, W.C., Hwang, S., Dhungel, S.K., Yoo, J., and Yi, J. (2007) *Mater. Sci. Eng. C*, **27**, 1452.

152 Mastrangelo, C.H. and Hsu, C.H. (1993) *IEEE J. Microelectromech. Syst.*, **2**, 44.

153 Mulhem, G.T., Soane, D.S., and Howe, R.T. (1993) Supercritical carbon dioxide drying of microstructures, Technical Digest, 7th International Conference on Solid- State Sensors and Actuators (Transducers '93), Yokahama, Japan, June.

154 Sampsell, J.B. (1993) The digital micromirror device and its application to projection displays, Technical Digest, 7th International Conference on Solid-State Sensors and Actuators (Transducers '93), Yokahama, Japan, June 1993, p. 24.

155 Storment, C.W., Borkholder, D.A., Westerlind, V., Suh, J.W., Maluf, N.I., and Kovacs, G.T.A. Flexible, dry-released process for aluminium electrostatic actuators (1994) *IEEE J. Microelectromech. Syst.*, **3** (3), 296–299.

2
Scanning Probe Microscopy as a Tool for the Fabrication of Structured Surfaces

Claudia Haensch, Nicole Herzer, Stephanie Hoeppener, and Ulrich S. Schubert

2.1
Introduction

The invention of scanning probe microscopy (SPM) by Binnig and Rohrer [1, 2] during the 1980s represented an important milestone in the field of surface sciences. Soon after its invention, the capability of the SPM technique (and its variations) to manipulate matter by means of the scanning tip inspired the development of new approaches for nanofabrication. The desire to generate ever-smaller structures that cannot easily be fabricated using conventional structuring tools, as well as very large-scale integration and complementary metal-oxide-semiconductor approaches, fuelled a major interest in using SPM methods to create structures with nanometer resolution. Although not the primary example of manipulating surfaces by means of SPM, the studies of Dagata *et al*. should be mentioned as being an important step in this development. Thus, it was shown that by applying voltage pulses, first via scanning tunneling microscopy (STM) [3] and later via an atomic force microscopy (AFM) tips [4], oxide structures could be formed with nanometer resolution. Subsequently, Day *et al*. were the first to investigate site-selective oxidation on silicon [4], when they used two different silicon substrates for patterning, namely silicon with a thermally grown 10 nm-thick layer of silicon oxide and silicon with native oxide; the inscribed features were then analyzed using AFM technique. As a result, patterning structures were obtained with a height of 2.8 nm, and which were recessed about 3.8 nm after etching with hydrofluoric acid (HF). Notably, these results indicated not only the formation of silicon dioxide but also consumption of the silicon substrate during oxide formation, since the hole structures had been created after the etching process. Although the obtained line widths were only 85 nm, this could be further improved by using sharper tips. Shortly afterwards, Yasutake and coworkers described the patterning of Si(100) and hydrogen-terminated Si surfaces by the application of a negative voltage between an Au-coated Si_3N_4 tip and the Si [5]. Subsequent AFM and Auger electron spectroscopy investigations were carried out to investigate the height and chemical composition of the obtained features. A schematic representation of the local anodic oxidation (LAO) process is shown in Figure 2.1.

Nanotechnology, Volume 8: Nanostructured Surfaces. Edited by Lifeng Chi

ISBN: 978-3-527-31739-4

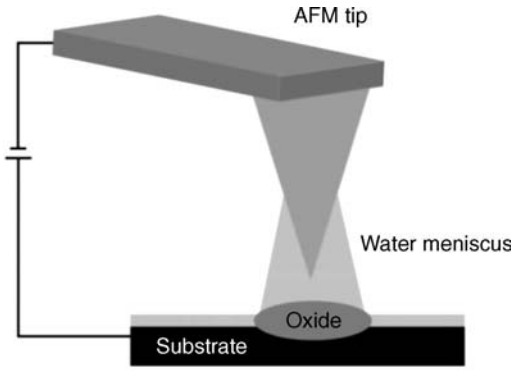


Figure 2.1 Schematic representation of the local anodic oxidation process.

In this case, the AFM tip serves as the cathode whilst the substrate acts as the anode. As this method is dependent on an electrical current, both the AFM tip and substrate must be conductive [6]. The water meniscus which is formed between the tip and the surface can be seen as a nanometer-sized electrical cell [7], and also provides the electrolytes for the LAO, which is essential for the patterning. The applied voltage induces an electric field with a value of $10^9\,\mathrm{V\,m^{-1}}$ as the threshold for the LAO process [8]. In fact, this electrical field is responsible for ionizing the water molecules to form reactive ionic species necessary for the LAO process [6]. In this way, the tip was used to create a nanometric electrochemical cell that consisted of the substrate, the tip itself, and the water meniscus (which is present under atmospheric conditions). At the anode – that is, the sample surface – an oxidation reaction takes place as described by Garcia *et al.* [7, 9], whilst at the cathode (in this case, the SFM tip) hydrogen is generated:

$$\mathrm{Si} + 2\mathrm{H}^+ + 2\mathrm{OH}^- \rightarrow \mathrm{Si(OH)_2} \rightarrow \mathrm{SiO_2} + 2\mathrm{H}^+ + 2\mathrm{e}^-$$

$$(\mathrm{M} + n\mathrm{H_2O} \rightarrow \mathrm{MO}_n + 2n\mathrm{H}^+ + 2n\mathrm{e}^-)$$

$$2\mathrm{H}^+_{\mathrm{aq}} + 2\mathrm{e}^- \rightarrow \mathrm{H_2}.$$

Investigations of the oxidized structures are carried out by taking measurements of the obtained height, since the silicon oxide features reveal an increase in height after the LAO [8]. This might be explained by the molecular volume of the oxides, which is normally larger than that of the substrate, and therefore the raised features would be formed during the LAO reaction [10]. In this situation, the hydrolysis of water from the water meniscus plays an essential role, and emphasizes the importance of the reliable formation of a water meniscus between the tip and the sample. The typical size of the water meniscus at different ambient conditions can be investigated using environmental scanning electron microscopy (ESAM), as depicted in Figure 2.2.

Whereas, the majority of LAOs were performed in a water meniscus, some examples have been reported where organic solvents were used as the reaction media for the inscription of nanometer-sized structures. Garcia and coworkers

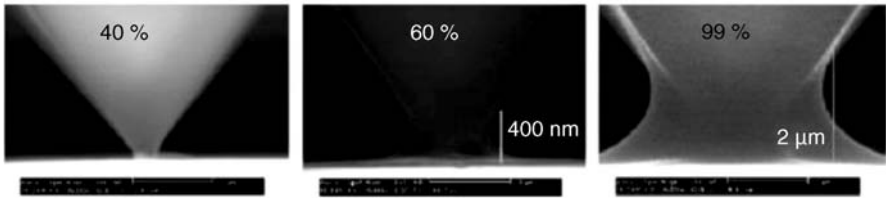


Figure 2.2 Environmental scanning electron microscopy investigation of the typical size of the water menisci at different relative humidities. Reproduced with permission from Ref. [11].

described a comparison of the LAO process in water and ethanol [12] where the LAO was found to be increased in time, supposedly as the result of a reduction in the trapped charges within the growing oxide. These ethanol menisci were later used for the fabrication of nanometer-sized carbide structures [13], whereby the inscribed nanowires were written with a 70 nm distance between each other and with diameters less than 45 nm, demonstrating the high accuracy of the LAO. Additional studies included the use of octane and 1-octene as organic liquids in the meniscus for the LAO of sub-10 nm-sized structures [14]. Further reports have described the LAO in hexadecane [15], in hydrocarbon solvents (e.g., *n*-octane, toluene, dioxane) [16] and in HF/ethanol [17]. The LAO has been reported for a wide variety of different substrates, including semiconductors, metals, and self-assembled monolayers (SAMs). Whereas, the LAO on silicon [18–29] was subject of the primary experiments, other semiconductors such as gallium arsenide [30–36], germanium [37–39] and silicon nitride [40–45] have been shown to be of interest for the fabrication of electronic devices. In addition to semiconductors, many different metals have been investigated for the LAO, including titanium [46–51], ferromagnetic metals [52–55], niobium [56, 57], molybdenum [58–60], aluminum [61, 62], and zirconium [63]. Furthermore, patterning was also achieved on both diamond [64–66] and graphene [67, 68] substrates. Since the first experiments related to LAO, several research groups have conducted investigations into the mechanism and reaction kinetics of this process [69–85]. Notably, such studies have addressed the dependence of the LAO process on various parameters, including the relative humidity, tip geometry, tip–substrate distance, applied voltage, and the oxidation time. Various explanations for the LAO behavior have been proposed by different models, including the Cabrera–Mott model, the power-law model, the direct-log kinetic model, and the space charge model [10]. Moreover, several groups have described a linear relationship between the thickness of the oxide layer and the applied voltage [6, 69]. When Avouris and coworkers discussed the rate of the LAO process, they proposed that this would decrease rapidly while the oxide layer was increasing, and that this effect could be explained by a self-limiting influence of the decreasing strength of the applied field and a build-up of stress [71]. These findings were subsequently supported by Sugimura and Nakagiri [9]; moreover, the resolution of the oxidized patterns could be controlled by the shape of the tip [70]. In addition to the commercially available conductive SPM tips that have in the past generally been used for LAOs, several groups have demonstrated the use of carbon nanotube (CNT) probes to improve the

resolution of the oxidized features [86–89]. This effect could be explained by a decrease in the water meniscus between the tip and the substrate. Kuramochi and coworkers, for example, demonstrated the fabrication of a lattice structure with line widths of 15 nm and a spacing of 35 nm, as well as concentric circles with 25 nm line widths and 25 nm spacing with multiwalled CNTs used as probes [90]. The size of the oxide features depend on the field strength applied between the tip and the substrate [79]. In subsequent studies, Kuramochi and coworkers quantified the role of the relative humidity compared to the size of the oxide structures [82] and showed that, whilst operating under constant humidity, the size would increase in line with the applied voltage and exposure time, but decrease in line with the speed of the tip. However, if the humidity was to be increased, the size of the features would increase if the applied voltage and exposure time were kept constant.

This electro-oxidation of the surfaces inspired a number of research projects that focused on investigating the nature and formation process of these structures. By using secondary ion mass spectroscopy (SIMS), it could be shown that the formed structures consisted of silicon oxide [91], and this was confirmed using X-ray photoelectron spectroscopy (XPS) [30]. This fact is of particular importance since the processing of silicon oxide is performed routinely via wet-etching processes, as implemented in microelectronics. Indeed, it was shown that the formed silicon oxide structures could be used in a similar way and could be developed into topographic features, for example by applying HF etching [4], anisotropic wet-etching (e.g., with hydrazine or aqueous solutions of potassium hydroxide) [92, 93], or dry-etching procedures [94]. This compatibility with standard structuring techniques would allow the rational and efficient design of functional devices, such that the SPM could be used to generate sophisticated, small features that cannot easily be produced using conventional methods. This led to the application of LAO lithography not only for the fabrication of etched masks [40] and high-density read-only storage devices [49], but also in the fabrication of device structures that have since achieved a remarkable level of sophistication.

The major drawbacks of this patterning technique are the slow acquisition of the image, the exposure rate, and the obtainable size of the LAO structures [94]. The area of the features is limited by the size of the piezoelectric scanner [95], which in turn limits the industrial application of the system. Consequently, several attempts have been made to produce centimeter-sized areas by using a parallel lithography method. As an example, Minne *et al*. demonstrated a reliable electro-oxidation process that employed an array of two cantilevers, although unfortunately an array of five cantilevers showed different qualities of the electro-oxidation patterns for every individual tip [94]. However, this problem could be overcome by the use of a 2×1 array of individually controlled cantilevers [96]. While using a modular micromachined parallel AFM tip array combined with large displacement scanners, the electro-oxidation of centimeter-sized areas is possible, and also in high resolution [97]. Another interesting approach towards upscaling the LAO process would be the use of metalized stamps [95, 98–100]. For this, e.g., a metalized digital video disc (DVD) polymeric support with multiple protrusions was used as cathode, so as to generate a large amount of features with line widths down to 100 nm [95].

And, only three years later the same group introduced an instrument that could perform anodic oxidation in parallel fashion [101], opening the possibility of fabricating silicon oxide patterns over square-centimeter regions in an operation time of less than one minute.

Villarroya *et al.* reported the fabrication of a new cantilever-based sensor system for biochemical detections to be operated within liquid environments, and which employed conventional micro-electro-mechanical system (MEMS) technology and AFM-based lithography. The key feature of the sensor layout here was the implementation of electrical elements for the deflection detection of the cantilever. This was realized by measuring the change in electrochemical current between the mobile cantilever of the sensor and another electrode that was fixed at the free extreme of the cantilever. In this way, a finger-like array of electrodes with high spring constants could be fabricated on the microfabricated sensor cantilever. As a result of the cantilever deflection the relative positions of the electrodes would be altered, and this would result in a change in the electrochemical current to be measured. Hence, advantage could be taken of the heavy dependence of the electrochemical potential on the effective facing cross-sections of both electrodes. In order to ensure a sufficient sensitivity of this detection principle, it was necessary that the distance between the electrodes was less than 100 nm, and that they were patterned by using the AFM electro-oxidation of aluminum oxide on the microfabricated cantilever itself. This led to an increase in the thickness of the aluminum oxide, which subsequently was removed to generate the finger-electrodes.

Minne *et al.* demonstrated the fabrication of 0.1 μm metal oxide semiconductor field effect transistors (MOSFETs) on amorphous silicon (α: Si) films [102]. The probe-induced oxide pattern was first transferred onto α: Si by plasma dry etching, after which the gate contact pad was masked by the photoresist and the gate masked by oxide, leaving the α: Si in these regions intact. The authors reported later on a parallel lithography approach in which the direct electro-oxidation process was coupled with arrays of cantilevers (maximum 50) [96, 97, 102]. Wilder and Quate introduced a cantilever within an integrated MOSFET as a current source for the on-chip control of the exposure current.

Moreover, electro-oxidation of the substrate allows the direct manipulation of the electronic properties of the substrate, with nanometer precision. Campbell *et al.* used the electro-oxidation of Ti films to fabricate metal-oxide-metal devices [103], by scanning a biased tip across a predefined area to define a wire structure which restricted the current flow. The tip was subsequently repositioned on the nonoxidized side of the wire and scanned towards the wire. Due to the constriction of the electro-oxidation process, the electrical resistance was increased as the tip was moved towards the oxide wire. The measurement of the device's resistance during the electro-oxidation process allowed a precise control of the width and the resistance of the junctions and, as a result, structures with dimensions of less than 10 nm and precisely tailored electrical properties could be obtained.

Ishii *et al.* used a AlGaAs/GaAs heterostructure containing a two-dimensional (2-D) electron gas as substrate for electro-oxidation [104]. The group observed that the electro-oxidation of the GaAs cap layer resulted in an increase of the resistance within

the 2-D electron gas. Later, Ensslin *et al*. showed that the 2-D electron gas was depleted by a local oxidation of the cap layer [105] if the electron gas was less than 50 nm beneath the surface, and used this process to produce a large variety of sophisticated devices. It appears that a depletion of the electron gas results from the fact that, during the electro-oxidation process, the surface/electron gas distance is reduced and thus the number of surface states is slightly increased. The majority of the donor electrons from the doping layer is then used to fill up these surface states, whilst only a small number of electrons move into the electron gas. The change in the internal electric field is regarded as the reason for the depletion of the electron gas, and one-dimensional (1-D) simulations using a Poisson–Schrödinger solver were used to confirm this mechanism. In this case, the line structures typically showed a width of 100 nm and a height of 8–10 nm. Moreover, it was found that an increase occurred in the resistance in the 2-D electron gas below the oxidized area, and the introduction of additional top and gate layers allowed efficient tuning of the device's characteristics. Similar to the electro-oxidation of the GaAs cap layer, this approach could also be used to pattern the thin Ti gate electrodes to create, for example, quantum point contacts [106]. Consequently, a wide variety of different structures was fabricated to study not only the conductance but also conductance fluctuations in quantum wires [105], four-terminal quantum dots (QDs) and a double quantum dot system with integrated charge readout [107]. Coulomb blockade oscillations in in-plane gate singe-electron transistors [108], the conductance in single-electron transistors and quantum point contacts [109], as well as quantum rings [110], magnetotransport in antidot arrays [111], Aharonov–Bohm oscillations in quantum ring structures [112] and Coulomb blockade resonances in single-electron transistors [113] are all examples demonstrating the impressive capabilities of this patterning approach.

These examples stress some of the advantages of SPM-based structuring techniques. In particular, the wide variety of structuring modes that can be used to modify the surface itself as well as its properties are very versatile. Besides the electro-oxidative modification schemes introduced here, many other interactions can be used to inscribe features onto a surface. Notably, mechanical, thermal, electrostatic and chemical interactions (or combinations of these) have been used to structure surfaces at the nanoscale by means of SPM-based approaches [114] and, as a result, the versatility and variety of materials and structures that can be produced has fuelled extensive interest in SPM-based lithography. The reported surfaces and materials that can be patterned include semiconducting materials such as silicon, metals and rare-earth metal oxide blends, and also molecularly functionalized surfaces including thiols, silanes, DNA, proteins, biomolecules, and nanoparticles [115, 116]. One other important advantage of SPM-based structuring methods is an ability to visualize the inscribed features at high resolution, directly after the patterning process [117, 118]. From a technical point of view, SPM-based techniques are simpler and do not require the expensive fabrication of masks (as do other nanostructuring methods), and consequently the technique is much cheaper, especially for prototyping purposes. With regards to instrumental requirements the technique is very cost-effective, as only a relatively cheap AFM/STM set-up is required [117]. Furthermore, the method offers a complete freedom of pattern choice, a very high spatial precision, and

produces structures with a resolution of <10 nm (a value that requires significant know-how if it is to be achieved using conventional structuring methods) [119–122]. A molecular precision of the measured objects can be achieved (and even improved) by using an ultrasharp tip [119]. Likewise, from a practical aspect it is possible to use the same tip to pattern the substrate and to image the inscribed structure *in-situ* after the writing process, thus implementing a direct control of the fabrication step. Whereas, STM-based lithography is mainly performed in ultra-high vacuum, structuring methods based on AFM can be carried out in an ambient environment and under liquid conditions [119, 122]; thus, AFM-based lithography shows great promise for the structuring of biomaterials *in-vitro*, and also for imaging under physiological conditions. Of interest to chemists is the possibility to investigate nanometer-sized features, such as molecular-scale chemical syntheses [115]. Systematic studies of the size-dependent properties of the inscribed nanostructures can be performed directly using AFM, based on an ability to make *in situ* changes to the nanometer-sized features. Another important aspect of these nanometer-sized structures is the possible investigation of molecular recognition processes, the electronic behavior of small clusters of molecules, and the manipulation and organization of biomaterials [115]. Moreover, additional studies may provide information on tip–surface interactions, structures and properties on a nanometer level [119].

Unfortunately, however, SPM-based lithography has certain well-known limitations when implementing fabrication applications. Notably, the fabrication of nanostructures is serial in nature and therefore rather slow [115, 119]; consequently, until now SPM-based patterning techniques have been the subject of research mainly at the academic level. Although the use of SPM-based lithography as a manufacturing tool for high-throughput applications still presents a major challenge, much effort has been made during the past decade to increase the writing speed with a single tip, and to pattern simultaneously with multiple tips [120, 123–132]. For example, Cruchon-Dupeyrat *et al*. introduced an automated vector-scanning scanning probe lithography (SPL) instrument, which could be programmed or linked to computer-assisted design software [120]. However, due to drift and creep of the piezo scanner, and to electronics and/or cantilever artifacts, the inscribed text files and filled structures fidelity of the process have not yet been perfected. Never the less, many of these problems could be solved by using a closed-looped feedback sensor, or by choosing a cantilever with the correct spring constant. More recently, centimeter-scale imaging and lithography with tip arrays of up to 50 cantilevers were reported by Minne *et al*. [97]. In this case, a modular micromachined parallel AFM array was combined with a large displacement scanner and used to produce nanometer-sized features over a large area. Likewise, Wouters and coworkers reported details of the large-scale constructive nanolithography of *n*-octadecyltrichlorosilane (OTS) monolayers via two different methods [127]. The first method used automated AFM where, with a single tip, approximately 1000 structures could be transferred onto the substrate, whilst in a second approach a four-cantilever array was used to inscribe different patterns. Both methods produced structures of high resolution that could be further modified by, for example, the self-assembly of nanomaterials. A cantilever array of five probes for the parallel SPL of *n*-octadecyltrimethoxysilane monolayers

was successfully demonstrated by Kakushima *et al*. [128]. In addition, the fabrication of an AFM array with a single-electron transistor has been reported, which can be combined with ultraviolet (UV) lithography for application in quantum devices. Mirkin and coworkers reported the fabrication of a nanoplotter with an array of microfabricated probes for application in parallel dip-pen nanolithography (DPN) [123]. During these investigations, two types of tip array were developed. The first type consisted of 32 silicon nitride cantilevers, separated by 100 μm; although this provided straightforward writing and imaging of the structures, the sharpness of the tips that could be produced was greatly limited by the conformal blanket deposition of the silicon nitride thin film. The second type of probe array, which consisted of eight boron-doped silicon tips separated by 310 μm, provided structures with line widths down to 60 nm and increased imaging capabilities, but the probe density was diminished. One of the first probes to have a large number of tip arrays (created by IBM) incorporated a 2-D array of 32×32 (1024) AFM cantilevers [125]. The information densities of the array were in the order of 650 to 1300 Gb cm^{-2}, and the unit showed great promise for nanostructuring due to its high-speed/large-scale imaging properties. Recently, a 2-D cantilever array consisting of 55 000 tips was reported by Mirkin *et al*. [130]. One other major disadvantage of SPM-based lithography is the limited layer thickness that can be patterned at high resolution. When compared to patterning techniques that use electron beams (when relatively thick layers of resist material can be created), SPL methods are unable to produce thick layers without a significant decrease in resolution. This leads to their applications being limited to selective etching processes, as the thin resist layers are often unable to withstand the extreme conditions of the etching method.

Possible resist layers can be based on SAMs. These are parts of two main categories of which have been used basically in recent research: (i) silane-based monolayers, which react with silicon, glass and activated metal surfaces such as Al; and (ii) thiol monolayers, which react on gold or silver substrates. Thiol-based monolayers were first utilized in 1983 by Nuzzo *et al*. [133], whereas silane-based monolayers were introduced slightly earlier, in 1980, by Sagiv [134]. Until now, the thiol/gold combination has received the most attention, mainly due its easy preparation. Overviews of thiol-based monolayers and their applications have been provided by Ulman *et al*. [135, 136], Everhart [137], Whitesides *et al*. [138], Woodruff [139], and Mutzutami [140].

Nevertheless, silane-based monolayers demonstrate certain advantages over thiols, notably the high stability of the monolayer which results from a covalent network formation, consisting of three bonds, between the surface and the silane molecules. In particular, silane-based SAMs are stable and closely packed, which allows them to serve as a good resist layer for chemical transformation, and also to provide insulating layers. Such high stability also allows further modification steps to be carried out, without affecting the monolayer, in particular at higher temperatures. Moreover, silane-based monolayers are also compatible with silicon technology; notably, the electronic properties of silicon can be influenced by the presence of SAMs, as reported by Peor *et al*. [141], whereby alkyl-, benzyl-, chloro-methylbenzyl-, chlorobenzyl-, bromobenzyl- and iodobenzyltrichlorosilanes were each self-assem-

bled and studied using Kelvin probe techniques. The two main parameters required to tune the electronic properties were coverage of the substrate and the molecular dipole moment of the molecules [141]. Subsequently, Rittner *et al*. investigated the electrical properties of SAMs on hydroxylated silicon surfaces by utilizing C_{18} alkyl chains bearing methyl, thiol, thiophene, phenoxy, and biphenyl end groups. Of particular interest here were the insulating properties of the monolayers and the breakdown voltage. For example, the fact that iodine doping led to an increase in conductivity suggested that it might be possible to build a nanomolecular transistor by using the functional end group as an active layer for the deposition of a conductive layer on the SAM dielectric layer [142]. In further studies, Li *et al*. self-assembled ferrocene-containing monolayers onto silicon and investigated both their capacity and conductance. Interestingly, because they are reversibly chargeable, such monolayers might not only find potential applications in memory devices [143] but also permit the use of optical techniques (e.g., fluorescence spectroscopy) in their investigation. For example, Lee *et al*. described the preparation of spot arrays for protein synthesis by patterning through a photoresist, followed by perfluorination and finally amination with various silane monolayers. In order to achieve this, Fmoc (9-fluorenylmethyl chloroformate) amine acid was first coupled onto the glass surface, followed by fluorescent labeling that allowed the reaction to be monitored via fluorescence imaging. This allowed the creation of a model library of amino acids, with the α-chymotrypsin subsite specificities being replicated by coupling Cy5–streptavidin to the remaining biotin, following enzymatic digestion [144].

These monolayers, and their effective implementation into structuring and nanofabricaton schemes, achieved significance when it was first realized that SAMs and their surface reactions could prove valuable in SPM-based lithography for tuning the surface properties towards specific applications. In particular, based on their inherent hydrophobicity, monolayers showed a potential to improve the resolution of electro-oxidative nanolithography methods, and it soon became clear that the nature of the terminal end groups played an important role in reducing the dimensions of the water meniscus required for the electro-oxidation process. The SAM/ceramic bilayer coatings have also been shown to play important roles in the protection of silicon devices or other electronic applications. For example, Salami *et al*. have described a series of phosphonato-based triethoxysilanes used to manipulate the growth of zirconium oxide [145].

The thermal stability of silane-based SAMs renders them compatible with chemical transformations at higher temperatures, and also allows a wide variety of chemical surface reactions to be conducted, including surface-initiated atom transfer radical polymerization (ATRP) at 100 °C [146], surface-initiated reversible addition fragmentation chain transfer (RAFT) radical polymerizations, which were carried out at up to 90 °C [147], and/or the formation of an imide bond between an amine acid salt bilayer at temperatures up to 210 °C [60]. These examples indicate the versatility of functional or reactive SPM-modified substrates to tune the surface's properties, to render its functionality, or simply to attach objects to that surface. Thus, the implementation of such substrates into nanometric frameworks represents a

promising approach to the creation of nanometer-sized functional structures by means of SAM-based lithographic techniques.

This brief overview of SPM-based lithography, and its possible combination with SAMs, is indicative of the high potential of this method; hence, extensive investigations have been conducted in this area during the past decade.

During the past decade, the integration of SAMs into SPM lithographic approaches has fuelled sustained research activities, such that a variety of modification schemes has been developed. In the following sections, a brief overview will be provided, highlighting (with selected examples) the potential that derives from the implementation of monolayers in nanofabrication processes. In the past, although mechanical, thermal, electrostatic and chemical interactions have each been used for the inscription of structures, the following relates to two main approaches. The first approach is focused on the patterning methods relating to structuring with SAMs, with the DPN method being used as an example. Next, SPM lithographic methods will be discussed, by which patterned features can be obtained on SAMs on different materials. In addition to nanografting and nanoshaving, the mechanical fabrication routes used to develop nanometer-sized features – notably electro-oxidative nanolithography and the chemical activation of SAMs by means of tip-induced surface reactions – will each be discussed in greater detail. Following a description of the details of the ready-made silane-based monolayers and surface reactions of SAMs that can be implemented into the nanofabrication route. Discussion will be centered on the combination of patterning and surface chemistry of SAMs towards multifunctional surfaces.

2.2
Structuring with Self-Assembled Monolayers

2.2.1
Dip-Pen Nanolithography

Since its development in 1999, the fabrication of nanometer-sized features by DPN has attracted significant attention in the field of nanotechnology [114, 148–155]. This technique, which was introduced by Mirkin *et al.*, uses an AFM tip to directly transfer an ink onto a substrate via capillary transport, so as to create patterned surfaces [156]. Such transport is possible due to the formation of a water meniscus between the AFM tip and the substrate, while the driving force behind the transfer of the ink to the surface is the chemisorption between the ink molecules and the surface [148]. The chemisorption of these molecules leads to the fabrication of stable surface structures. A schematic representation of the DPN technique is illustrated in Figure 2.3.

Although Mirkin and coworkers were the first to introduce the concept of DPN, the initial experiments for depositing ink from an AFM tip were described by Jaschke and Butt back in 1995 [157], with the deposition of 1-octadecanethiol (ODT) onto mica via an AFM tip. In this case, two main results were noted in terms of the ODT depositions. In some experiments, transfer of the ODT onto the surface occurred

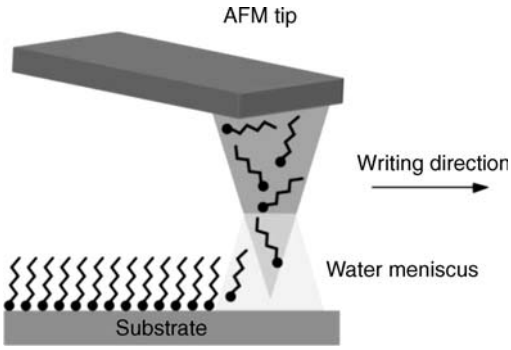


Figure 2.3 Schematic representation of the dip-pen nanolithography (DPN) technique.

immediately after contact of the tip with the mica; however, in other cases a slow growth of small and randomly placed structures was observed, although the inscribed structures were homogeneous in terms of their height and also stable. Four years later, Mirkin and colleagues carried out a further series of experiments on the deposition of thiols onto Au with ODT as the ink [156], in which case the formation of stable surface structures was achieved by the covalent linkage of −SH moieties to the Au surface. After having dipped an AFM tip into a solution of ODT in acetonitrile for 1 min, lateral force microscopy (LFM) measurements were carried out to investigate, in direct manner, the successful deposition of the transferred ink to the surface (Figure 2.4). This was verified by a lower frictional contrast of the inscribed features compared to the bare gold substrate, as illustrated in Figure 2.4a. The lattice-resolved LFM image (Figure 2.4b) confirmed the formation of a highly ordered, densely packed monolayer of thiol molecules.

In further experiments, ODT and 16-mercaptohexadecanoic acid (MHA) were combined as an ink for the fabrication of multicomponent nanostructures by a stepwise writing process [158]. In this way, two differently coated AFM tips were applied for the fabrication of parallel lines consisting of ODT and MHA, while subsequent LFM investigations revealed the precise inscription of six parallel lines of the two inks. These results showed that it was possible not only to create accurate nanostructures with multiple inks, but also to align them with a precision better than 5 nm. In a further experiment, where a pattern of MHA was "overwritten" with ODT, the previously inscribed structures were not influenced by any subsequent writing process, due to the fact that the ink had become linked exclusively to the surface areas, where it had a chemical affinity.

The method of DPN depends on several different aspects, each of which must be taken into account. The resolution of the patterns relies on various parameters, including the grain size of the substrate, the chemical affinity of the ink molecules to the substrate, the contact time between the tip and the substrate, the tip radius and the material of the tip, the writing speed, and the relative humidity [156]. Whereas, a grainy surface leads to interrupted lines, smooth substrates are suitable for the writing of long lines with nanometer-sized widths and high quality. As the formation of stable structures relies mainly on the chemisorption of the ink molecules to the

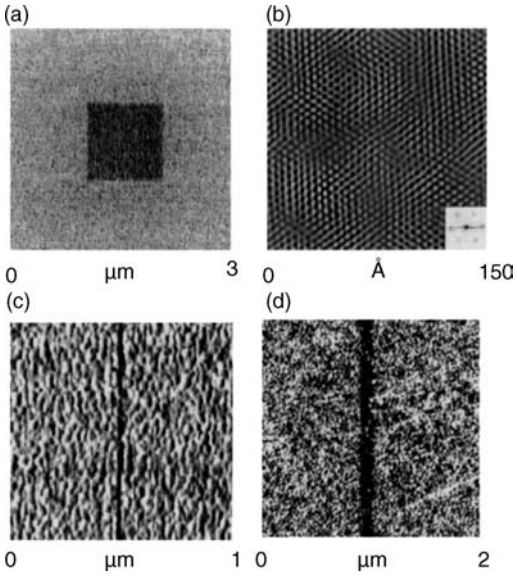


Figure 2.4 (a) Lateral force microscopy image of a square of 1-octadecanethiol inscribed on Au (1 µm × 1 µm); (b) Lattice-resolved, lateral force microscopy image of an 1-octadecanethiol self-assembled monolayer on Au(111)/mica; (c) Lateral force microscopy image of a 30 nm-wide line by dip-pen nanolithography (DPN); (d) Lateral force microscopy image of a 100 nm line by DPN. The darker regions correspond to areas of relatively lower friction. Reproduced with permission from Ref. [156].

substrate, the choice of ink is an important parameter for the fabrication of well-ordered features. In general, it is possible to use different types of ink on various substrates. Notably, the relative humidity controls the size of the water meniscus formed between the tip and the substrate, and this, in turn, offers the possibility to control the ink transport rate, the feature size, and therefore also the line widths [158]. Recently, some critical questions arose regarding the role of the water meniscus, which was proposed to mediate the transport of ink to the surface [159]. However, the question remained as to how water-insoluble inks could be transferred to the surface. In an attempt to resolve this problem, Sheehan and Whitman conducted a study regarding the role of the relative humidity and thiol diffusion on DPN [160], and showed that such ink deposition could be observed after 24 h, even under dry air or in a N_2 atmosphere. These results suggested that the water meniscus was not necessary for the transfer process. In a later study, Schwartz also reported on the molecular transport from the AFM tip [159], where the patterning was investigated under various parameters including temperature, relative humidity, and an ethanol vapor, while the patterning was carried out using different tip coatings. The study results indicated that the writing process was not necessarily dependent on the water meniscus, since both ODT and MHA could also be patterned under 0% humidity. However, the resolution of the inscribed features was seen to depend on the relative humidity during the structuring process. A higher resolution of the inscribed

patterns could also be obtained by using sharper tips. The amount of material transferred to the surface was also seen to be a crucial parameter for the writing time and the number of features which can be inscribed, as well as for the size of the inscribed structures. Consequently, several different methods have been reported to increase the amount of ink that is adsorbed on the AFM tip [149, 161, 162]. The use of tips that are made from, or are coated with, polydimethylsiloxane (PDMS), represents one possibility of enhancing the amount of ink molecules deposited [152, 163, 164]. Other favorable characteristics of PDMS include the wide variety of inks that can be coated onto the tip, the reduced evaporation of the adsorbed ink, and the utilization for the patterning of rough surfaces due to its elastic properties [165]. The modification of the tip with a layer of, for example, 1-dodecylamine, alters the surface property to hydrophilic, and this leads to an improved quality of the LFM measurements due to a reduction in the capillary force and a higher resolution of soft-inked materials [166]. The transport of the ink to the surface, which in turn influences the quality of the patterns, is a complex process that is influenced by different parameters [151, 152, 167], notably the purity of the ink and the surface, the shape of the tip, and also the material, the relative humidity, the temperature during the writing process, and so on.

2.2.1.1 **Thermal Dip-Pen Nanolithography**

In an attempt to expand the number of usable inks for patterning, a variation of the normal DPN approach was developed, termed thermal dip-pen nanolithography (tDPN) [167–170]. In this process, the tip is first coated with a material that is solid at room temperature, and then brought into contact with the surface, such that the ink is transferred to the substrate when the cantilever is heated. Not only can tDPN be used to control the rate of deposition onto a localized area, but the writing can also be switched on and off. Furthermore, when compared to the normal DPN process, tDPN excludes any contamination of the inscribed structures during the imaging step, due to the on/off deposition of the ink. Reported materials that have been inscribed via tDPN range from organic molecules (e.g., octadecylphosphonic acid), metals (e.g., indium) and the polymers poly(*N*-isopropylacrylamide) (PNIPAAM) and poly(3-dodecylthiophene). In seeking a system that did not require a thermal cantilever, Mirkin and coworkers described a DPN technique that used high-melting temperature molecules, but did not need tDPN [171]. On investigating the patterning of various inks with melting points between 99 and 231 °C, Mirkin's group showed that if the conditions of the writing process were carefully optimized, then no heatable tips would be required for the inscription. However, for any inks that were poorly water-soluble, tDPN represented an important improvement compared to "normal" DPN.

In comparison to other nanofabrication techniques, DPN demonstrates several clear advantages. One important characteristic of the DPN method is its ability to use a wide range of different substrates, including metals, insulating, and semiconducting surfaces [149], as well as a wide variety of inks (including diverse chemical moieties) that may lead to functional surface patterns. Examples of these substrates include alkanethiols [156, 158, 172], silanes [173, 174], polymers [175–177], metal ions [178], and biomolecules [179–181] such as DNA [182–185], proteins [186–192],

and peptides [193]. The wide variety of materials that can be used as inks allows the possibility of investigating a number of different processes, such as biorecognition, the control of single virus particles on a surface, virus–cell infectivity processes, cell–cell adhesion and the mechanism of cell migration, as well as studies related to nanoscale phenomena and the monitoring of molecular processes *in-situ*, such as monolayer nucleation and growth [149, 194, 195]. In particular, the inscription process is not limited to one ink; in fact, it is possible to write with several different inks on the same substrate. The preparation of pristine multiple ink nanostructures can be used in the study of molecule-based electronics, catalysis and molecular diagnostics [158]. Compared to other nanofabrication techniques, DPN requires only small amounts of material to create nanometric features [156], and neither does it depend on the use of harsh conditions such as electron-beam, UV-light and/or development steps [149] to fabricate nanometer-sized features, and this helps to prevent contamination and destruction of the substrate. The resolution of the inscribed structures is less than 50 nm, which in turn leads to line widths in the range of 15 nm being achieved, and an alignment resolution of approximately 5 nm [150, 196]. Compared to other patterning techniques, DPN is a direct-write technique [151], with no need for any preparation of resist layers, stamps (as in micro-contact printing) or masks, nor the use of commercially non-available equipment [150, 154, 156, 158]. On a practical point, as the inscription process can be operated under either ambient or inert conditions, no ultrahigh-vacuum techniques are required [150]. Moreover, the fact that both the writing and imaging of the obtained features can be carried out with a single tip represents an important point for efficient writing and scanning processes [148]. Notably, the tips most frequently used for DPN are commercially available, while conductive tips are not required at all.

Two important disadvantages of DPN when used to fabricate large patterns are the speed of the writing process, and the size of the obtained features [154]. When implementing into high-throughput systems, the technique must be expanded from a serial to a parallel set-up, and various attempts have been made to overcome this limitation. Two strategies are available to achieve this: (i) the fabrication of a passive-pen array, where each tip duplicates the desirable structure [197]; and (ii) the use of an active array with individually addressable tips. Of these two methods, the passive array is the most applicable because the ink transfer appears to be force-independent, while the implementation of a nanoplotter with an array of probes for applications in parallel DPN has been introduced and further improved [123, 198, 199]. Others have reported the writing of large structures with three different probe arrays (with the number of tips ranging from 26 to 250) and a deposition speed of 0.935 cm min^{-1}. Currently, for DPN the largest number of tips used in an array has been 55 000 [130, 200]; when in 2-D format, such an array can be used for patterning over an area of several square centimeters, and with a resolution of less than 100 nm. As an example, when the successful writing of the image of Thomas Jefferson from a US five cent coin was achieved by the patterning of ODT on gold, 55 000 duplicates of the cover of the coin were written with high precision. The patterning of phospholipids has also been demonstrated using this array, with each tip writing the letter combination "INT" three times within only 12 s, over an area of 1 cm^2.

The second strategy involves the addressability of the individual tips. Actuation of the tip can be achieved in different ways, including thermal [132, 201, 202], piezoelectric, or electrostatic [203]. The *thermal actuation* of a tip array has been used most frequently, due to the easy fabrication process, the use of simple materials, and the large displacement of the tip array and performance at low voltages [201]. The main disadvantage of this method is the thermal crosstalk that occurs between neighboring tips, due to heat transfer between the probes; the technique is also unsuitable for use with temperature-sensitive materials [203]. Mirkin *et al.* demonstrated a thermally actuated array of 10 tips for the inscription of ODT on Au with sub-50 nm line widths [132, 201, 202]. An alternative approach to actuate the individual tips, *electrostatic actuation*, occurs due to the generation of electrostatic attraction forces [201] and is produced by the presence of two oppositely charged electrodes. In contrast to thermal actuation, the electrostatic method depends on a complex fabrication process, as the probes are not heated during the patterning process it is possible to inscribe temperature-sensitive materials [203]. The actuator crosstalk is also reduced in the case of electrostatic actuation. Recently, Bullen and Liu described the successful patterning of ODT using electrostatic-actuated tips with line widths down to 25 nm, which was comparable to the commercially available silicon nitride cantilevers that are used for ink transfer.

Since its introduction, the use of DPN has been reported using a wide variety of different inks and substrates. A selection of experiments and interesting examples from various research areas are outlined in the following sections.

2.2.1.2 **DPN with Biomolecules**

The construction of protein arrays represents an important area in the field of proteomics, cell research and diagnostics, among others [186]. For example, Mirkin and coworkers described the fabrication of arrays of MHA features on a background of passivating 11-mercaptoundecyl-tri(ethylene glycol). When the substrates were subsequently immersed in solutions of different proteins to test their adsorption behavior, the proteins assembled selectively on the MHA features, but no nonspecific adsorption occurred on the background layer. Most importantly, the proteins demonstrated biological activity after the adsorption process. Subsequently, cell-adhesion tests were carried out on patterns of Retronectin, adsorbed onto MHA, where the cells were attached selectively only onto the patterned areas. A direct approach for writing patterns of proteins was reported later by the same group [204]. When both rabbit immunoglobulin G (IgG) and anti-rabbit IgG nanostructures were inscribed on negatively charged and aldehyde-terminated substrates, fluorescence imaging was used to reveal the chemical identity of the fluorophore-labeled anti-rabbit IgG protein. Additional studies were carried out on the high-throughput production of large protein patterns [205], in which features of *N*-hydroxysuccinimide (NHS) were inscribed on gold for selective reactions with a variety of proteins; the proteins were later labeled with Alexa Fluor 594 to investigate the biological activity of the antibodies when located on the protein structures (Figure 2.5). The tapping mode height and fluorescence images confirmed the antibody adsorption onto the protein-array templates.

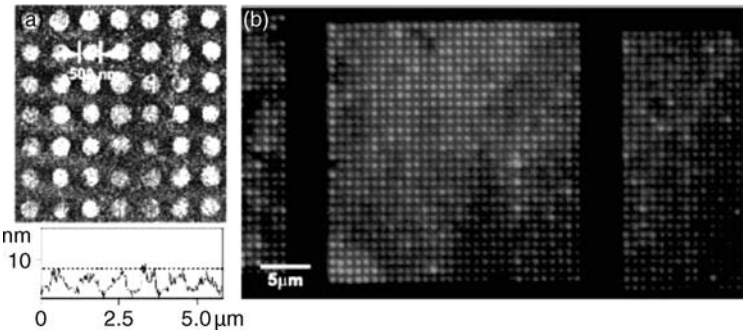


Figure 2.5 (a) Tapping mode height image and height profile of fluorescein isothiocyanate Alexa Fluor 594-labeled human immunoglobulin G (IgG) nanoarrays immobilized onto protein A/G templates; (b) Fluorescence microscopy image of Alexa Fluor 594-labeled antibody nanoarray patterns. Reproduced with permission from Ref. [205].

Further studies in the field of biomolecules were conducted by Li *et al.*, who reported the fabrication of nanopatterns on individual, stretched DNA molecules [206]. Nanostructures of gold, created by using DPN, can be used to assemble thiol-terminated DNA molecules [207], after which the DNA structures can further react with complementary DNA, or with particles modified with complementary DNA. This process was used also by Chung and coworkers to assemble single DNA-functionalized nanoparticles into the gap regions of single-electrode junctions [185]. The method described might be useful for the development of biosensors and to investigate electrical transport through such features. Other experiments related to the fabrication of functional electrical gaps for the detection of DNA have been recently reported by Li *et al.* [208]. In this case, DPN was used to pattern chip DNAs into micrometer-sized electrical gap structures, after which the DNA assemblies were reacted with target single-stranded DNA and DNA-functionalized nanoparticles to form structures capable of conducting an electrical current. The patterning of enzymes onto DNA-terminated monolayers leads to the possibility of performing nanoscale enzymology [209]. Hyun *et al.* demonstrated the writing of DNase 1 onto oligonucleotide SAMs, with a subsequent treatment of the surface with Mg^{2+} ions so as to create hole structures, caused by digestion of the surface-bound substrate by the enzyme.

2.2.1.3 DPN with Polymers

The patterning of polymers represents an interesting area of research, based on the possibility of using these structures to fabricate sensors, in catalysis, and for optical devices. In particular, the patterning of conductive polymers has attracted significant attention, as they might be used for the fabrication of nanodevices and nanosensors. Lim *et al.* described the deposition of nanometer-sized structures of self-doped sulfonated polyaniline and doped polypyrrole onto positively and negatively charged surfaces via electrostatic interactions [210], with the obtained features being characterized by LFM and electrochemical measurements. The writing of polythiophene

nanowires on semiconducting and insulating substrates was demonstrated by Maynor and coworkers [211], who used a variation of the normal DPN process, the so-called "electrochemical DPN." For this, the monomer units were polymerized at the tip–substrate interface, which led to the creation of conducting polymeric nanowires with a resolution of more than 100 nm. Other examples included the combination of DPN and ring-opening metathesis polymerization (ROMP) to create combinatorial libraries of functional polymer features [212], the guided pattern formation in spin-coated polymer blend films from DPN-inscribed surface templates [177], and the creation of nanostructures in poly(4-vinylpyridine) by local protonation with a pH 4 buffer solution used as ink [213].

2.2.1.4 **DPN with Fluorescent Dyes**

Patterned areas of fluorescent dyes have the potential for application in high-density optical information storage, optoelectronic devices, and biological staining [214, 215]. When Su and Dravid demonstrated the inscription of different organic dyes on both bare and modified silicon substrates [214], characterization of the structures by fluorescence microscopy showed the emission of, for example, eight parallel lines of rhodamine 6G (R6G) on negatively charged silicon. Notably, the line widths could be controlled by changing the scanning speed of the tip. Other examples included the fabrication of luminescent patterns of R6G and their characterization, using scanning confocal microscopy, down to the single-molecule level [215], the placement of fluorescent-labeled silazanes [216], and the deposition of fluorescent adamantyl-functionalized molecules on β-cyclodextrin monolayers [217].

2.2.1.5 **DPN in the Field of Electrolytes**

The fabrication of polyelectrolyte structures via layer-by-layer formation might be incorporated into applications such as nanoelectronics, and also in the study of cell-adhesion characteristics. Yu and coworkers described the fabrication of polyelectrolyte structures on different surfaces [218], whereby poly(diallyldimethylammonium) chloride (PDDA) and poly(styrenesulfonate) (PSS) were used as inks to obtain positively and negatively charged nanofeatures. The method demonstrated a potential for implementation into other surface-engineering applications, such as directed cell growth and surface-mediated molecular assemblies. Lee *et al*. created polyelectrolyte multilayers of PDDA and PSS on structured MHA patterns on gold [219], after which the background was filled with different molecules, such as ODT or poly(ethylene glycol) (PEG), to prevent any nonspecific adsorption of the polyelectrolytes. Fabrication of the multilayer was achieved by alternating the assembly of PDDA and PSS onto the MHA structures. The features were then characterized using fluorescence microscopy, after having labeled the multilayers with a fluorescein solution to reveal uniform fluorescein structures on the surface.

2.2.1.6 **DPN with Nanomaterials**

The use of nanomaterials in the process of DPN offers the possibility to guide, for example, the assembly of individual particles on a surface to study quantum phenomena or particle–particle and particle–substrate interactions; moreover,

nanoelectronic structures and devices could be designed by the patterning of metallic materials. The reported examples of different nanomaterials used in DPN range from gold [220–222], platinum [223] and magnetic nanoparticles [224–227], cadmium selenide nanostructures [228] and SnO_2 [229], positively charged modified polystyrene spheres [230], nanowires [231, 232] to single-wall carbon nanotubes (SWCNTs) [233]. The first report of the formation of metallic nanostructures by an electroless metal deposition process was reported by Maynor *et al.* [222]. In these studies, the water meniscus between the AFM tip and the surface acted as a reactor vessel, causing the metal ions to be reduced to metal atoms with subsequent deposition of the nanofeatures. The structures showed a high stability against several washing steps, as well as thermal stability up to 300 °C. Porter and coworkers also demonstrated the writing of gold and palladium lines via electroless deposition [234], obtaining line widths of 30 nm with a height of 10 nm. The generation of arrays of magnetic particles was demonstrated by Liu [224], where a pattern of MHA and ODT, fabricated by DPN, served as a template for the selective assembly of pre-prepared magnetic Fe particles. Further experiments of the fabrication of "hard" magnetic nanostructures were performed with barium hexaferrite [225], in which the writing of an ink containing iron nitrate and barium carbonate yielded BaFe particles with sub-100 nm diameters. The magnetic properties of the inscribed features were characterized by magnetic force microscopy (MFM), which revealed the magnetic nature of the particles. Basner and coworkers described an interesting approach for the generation of metallic nanowires [232] in which enzymes modified with Au nanoparticles were used as biocatalytic inks for the inscription of lines of different metals. This method proved to be useful for the generation of complex nanocircuitry.

2.2.1.7 **Chemical DPN**

Finally, DPN has been used in the field of chemical surface reactions, with Degenhart and coworkers reporting the fabrication of robust micrometer- and nanometer-sized reaction areas on surfaces [235]. For this purpose, NHS-terminated monolayers on gold were used for patterning with $-NH_2$-functionalized polyamidoamine dendrimers, which then underwent a chemical surface reaction (amide linkage formation), with the formation of covalently attached structures. As these structures are highly stable, they could be used in a variety of applications, including chemical sensors. Additional nanometric surface reactions were demonstrated by Chi and Choi [236], in which an interchain carboxylic anhydride (ICA)-terminated monolayer was used for patterning with alkylamines to form stable surface structures via amide bond formation. The ICA-terminated SAM was prepared by the treatment of a carboxylic acid-functionalized SAM on Au with trifluoroacetic anhydride and triethylamine. Subsequent characterization of the inscribed features, using LFM, revealed a lower friction of the structures compared to the background. Long and coworkers demonstrated the localized click chemistry by using DPN [237], where acetylene-terminated silicon substrates served as surface templates for the patterning of azide-modified dendrimers. The $-C\equiv CH$ end groups were obtained via a two-step synthesis, starting from an amino-terminated SAM, which was treated with an acetylene-functionalized carboxylic acid under peptide bond formation conditions. Coupling of the azide

dendrimer was achieved via a Cu(I)-catalyzed cycloaddition, while LFM was carried out to detect any changes in composition after the DPN process.

These selected examples of nanostructures that can be obtained by using DPN demonstrate the versatility of the method. Due to the fact that a wide range of materials can be patterned on different substrates, the procedure has many potential applications, including the fabrication of sensors, as electronic devices, and in studies of biorecognition processes. Moreover, its implementation in high-throughput experimentation has demonstrated the possibility of patterning large surface areas and large amounts of structures, with high quality.

2.3 Structuring of Self-Assembled Monolayers

Compared to the above-described DPN approach, the following structuring methods employ SAMs that have been self-assembled on various substrates and subsequently patterned. The main advantages of using SAMs are the thermal, chemical and physical stability of the system and the densely packed nature of the monolayers. Structuring with an AFM tip by employing mechanical forces (e.g., nanoshaving and nanografting) and by applying a bias voltage between the substrate and the tip – that is, with LAO and a constructive nanolithography approach – are detailed in the following sections.

2.3.1 Nanoshaving

The first stages of fabricating the nanometer-sized features of SAMs by mechanical forces with the aid of an AFM tip were reported by Liu *et al.* in 1994 [238]. Here, the structure and stability of $CH_3(CH_2)_9SH$ ($C_{10}SH$) and ODT, self-assembled on Au (111), were investigated by using AFM. Under sufficiently high loads of the AFM tip, the molecules could be removed from the surface; moreover, the process was seen to be reversible and, by applying a decreased load, the thiol molecules were able to diffuse back to the surface. This area of research was extended in 1995 with studies of alkylsilanes on mica [239], examining the displacement of monolayers of octadecyltriethoxysilane on mica and characterizing their stability by means of AFM. A comparison with their thiol analogues showed that a much greater force was needed to displace the silane-based SAMs. When compared to thiols on gold, the process was seen to be irreversible, due to a reduced diffusion and mechanical strength.

This effect is used in the so-called "nanoshaving" approach to create negative structures by the displacement of a SAM with an AFM tip under a high local pressure [115, 119, 153, 240]. This process can be divided into three steps: (i) characterization of the surface, using AFM operated at low forces; (ii) removal of the SAM to inscribe the nanofeatures (this causes the AFM tip to be scanned at a high local pressure over the surface, resulting in a high shear force on the contact areas and subsequent displacement of the SAM) and (iii) imaging and visualization of the

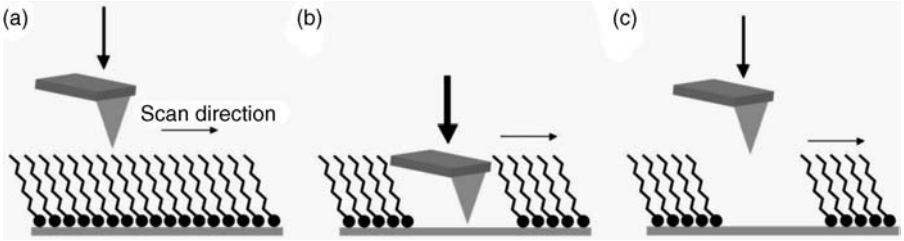


Figure 2.6 Schematic representation of the nanoshaving process. (a) Imaging of the surface; (b) Patterning of the self-assembled monolayer (SAM) under high local pressure; (c) Imaging of the surface.

inscribed structures, again under reduced loads. The manipulation via nanoshaving is shown schematically in Figure 2.6.

This process is highly dependent on the force applied to the surface [119, 240]. Whilst a very high force of the AFM tip can lead to plastic deformation or displacement of the substrate, low forces will result in an incomplete removal of the SAM; therefore, it is required that each system is investigated carefully and independently. The fabrication of high-resolution patterns depends on different parameters, including a molecule-by-molecule displacement, an immediate removal of the SAM, and a slow readsorption rate in order to prevent any backfilling of the structures with the self-assembled molecules that have been removed. Readsorption of the adsorbates depends on the environment of the fabrication route [119]. During the patterning of thiols on gold in air or water, the readsorption rate is higher than in the case of, for example, ethanol or 2-butanol; hence, the solubility of the thiols in the solvent used has a clear influence on the readsorption rate. Whilst thiols are not (or are only poorly) soluble in water, the removed adsorbates will remain weakly bound to the gold surface, and consequently a reversible displacement of the thiols will lead to low-resolution patterns. As the solubility of thiols is greater for ethanol and 2-butanol, however, the readsorption will be decreased and sharp patterns are obtained. The structuring of siloxane monolayers on mica can be achieved under different conditions. Due to the presence of only a few covalent bonds between the mica substrate and the siloxanes, the system has no long-range order. Furthermore, the siloxane molecules are connected via Si−O−Si bonds, which are responsible for the formation of a stable network. Following the nanoshaving process, the siloxane molecules that have been removed will show a low reactivity towards the mica substrate, and it is for this reason that the displacement of the molecules is irreversible and the structures obtained will have sharp features with a high resolution. Several examples have been reported where nanoshaving has been used not only to create structures with a high resolution, but also on a variety of substrates to demonstrate the potential of the technique in the fabrication of nanoelectronics. The production of functional semiconducting wires from sexithiophene, using an AFM tip, was described by Chwang *et al.* [241], who created wires between 300 and 70 nm wide via this approach. In order to investigate the electrical properties of these products, both photoconductivity and temperature-dependent transport measurements were carried out to

compare them with single grains of sexithiophene. This study also demonstrated the possibility of using other organic semiconductors to create nanometer-sized structures for applications in nanoelectronics. Liu and coworkers described the precise positioning of gold nanoparticles, surrounded by a shell of alkanethiol and alkanedithiol molecules [242]. In this case, an alkylthiol monolayer was used as a background for the nanostructuring by a sharp AFM tip. The nanoparticles adsorbed only onto the inscribed structures, such that the alkenedithiols served as anchoring groups for attachment to the surface. The construction of three-dimensional (3-D) protein–DNA features on gold substrates was demonstrated by Zhou *et al*. [243], whereby a thiolated DNA as monolayer was used as the base for patterning to obtain hole structures of $400 \times 400\,nm^2$. As this background layer proved to be resistant to the nonspecific adsorption of DNA–streptavidin assemblies, 3-D structures of DNA and streptavidin could be created via a step-by-step growing procedure. Zauscher *et al*. described the fabrication of stimuli-responsive nanopatterned polymer brushes of PNIPAAM onto gold surfaces [244, 245]. Here, a monolayer of ODT was formed as a resist layer onto Au, to serve as a template for the structuring by nanoshaving. Subsequent backfilling of the nanometer-sized features was achieved using a thiol-terminated initiator for the polymerization of NIPAAM. These findings might be important for the fabrication of silicon-based devices, where nanometer-sized polymer brushes could serve as barriers to different wet chemical etchants. The patterning of 1-alkenes, self-assembled onto hydrogen-passivated silicon surfaces, was described by Berrie and coworkers [246]. These authors showed that, depending on the applied load and the number of etching scans under high loads, the depths of the structures in the alkyl monolayer could be varied between 2 and 15 nm. Furthermore, these results were compared to the patterning of alkyl siloxane monolayers on silicon and mica. The reversible, templated nanostructure electrodeposition on a Au(111) surface by means of a "write, read and erase" nanolithographic approach was demonstrated by Borguet *et al*. [247], where a monolayer of ODT on gold was used for the inscription of nanometer-sized patterns of Ag. Deposition of the Ag structures was achieved by using electrochemistry, and was also reversible, depending on the applied voltage. Thus, this process would be suitable for the *in situ* deposition of metal structures for the fabrication of complex nanostructures. Cremer and coworkers reported on the fabrication of supported phospholipid bilayers with a resolution of less than 100 nm [248]. In this case, a bovine serum albumin (BSA) monolayer was patterned by an ultrasharp AFM tip under an applied force of approximately 300 nN. The inscribed structures were then backfilled with a vesicle solution which consisted of 1-palmitoyl-2-oleoyl-*sn*-glycero-3-phosphocholine and a dye-labeled lipid. After washing with a phosphate-buffered saline solution, investigations using fluorescence microscopy revealed a uniform fluorescence from the 55 nm-wide lines.

To summarize, nanoshaving represents a simple method for the fabrication of nanometric features via the mechanical removal of SAMs with an AFM tip. Under ambient conditions, nanoshaving enables patterning with high resolution, as well as an immediate characterization of the structures obtained. Various reports have outlined the possible use of this approach to fabricate complex 3-D biomolecular

structures, to create organic semi-conductor nanowires for transport studies, or in the "grafting from" technique of functional monomers. Nevertheless, the writing process must be carefully optimized to ensure not only that no damage of the underlying substrate occurs, but also that there is a complete removal of the SAM. either without or with less readsorption of the molecules onto the surface.

2.3.2 Nanografting

Nanografting represents an alternative patterning method, where thiol chemistry on gold is combined with AFM to create nanometer-sized features [249]. This technique is similar to nanoshaving, but includes an additional step. The procedure begins with the imaging of a SAM on the surface, under low pressure, in a liquid medium containing a second reactive adsorbate [115, 119, 240]. The AFM tip is then scanned with higher forces over the surface so as locally to remove any self-assembled molecules, which are then transported into the liquid. In the meantime, the second adsorbate molecules are adsorbed onto the freshly created structures, after which the nanostructures may be imaged under low force with the AFM tip to investigate the inscribed structures. A schematic overview of the fabrication of nanostructures in this way is shown in Figure 2.7.

Both, nanografting and nanoshaving, are heavily influenced by the applied force of the AFM tip onto the surface [119, 240]. As noted above, an excessively high load of the tip can results in a plastic deformation or displacement of the underlying substrate. although if the force is too low the SAM cannot be completely removed.

The term "nanografting" was first coined in 1997 by Xu and Liu [250], who used a $C_{10}SH$ monolayer on gold as a resist for patterning the surface with the AFM tip. In this case, ODT was chosen as the reactive second thiol compound for backfilling of the nanometer-sized features. The subsequent characterization using AFM revealed a height difference of 8.8 Å, which could be correlated to a crystalline-phase SAM (Figure 2.8); furthermore, no exchange between the ODT and $C_{10}S$–Au was observed in the remaining background layer. In addition, the ability to obtain multiple nanostructures by altering the thiol compound before each fabrication step opened

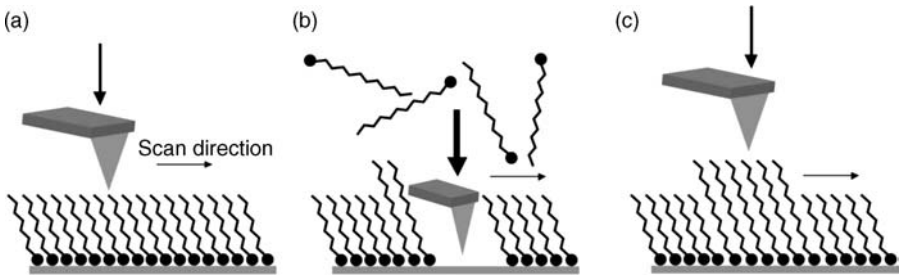


Figure 2.7 Schematic representation of the nanografting process. (a) Imaging of the surface; (b) Patterning of the self-assembled monolayer (SAM) under high local pressure with simultaneous adsorption of a second adsorbate molecule; (c) Imaging of the surface.

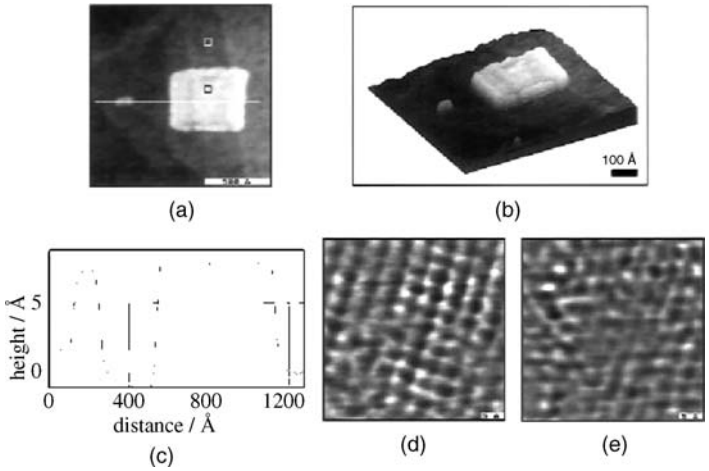


Figure 2.8 (a, b) Topographic images of the created square structures of the $CH_3(CH_2)_9SH$–Au resist (bright square = 1-octadecanethiol–Au; dark area = $CH_3(CH_2)_9SH$–Au); (c) Height profile of the 1-octadecanethiol–Au square; (d, e) Molecular-resolution images of the $CH_3(CH_2)_9SH$–Au and 1-octadecanethiol–Au areas. Reproduced with permission from Ref. [250].

the possibility of using the technique to fabricate, for example, nanoelectronic devices.

Further studies of the nanografting technique were performed and expanded by the group of Liu [251–259]; notably, investigations were also conducted into the kinetics of the self-assembly process of thiols on bare gold surfaces, compared to a spatially confined area [251]. In this way, an acceleration of the kinetics was observed for the nanometer-sized features, due to the generation of a transient reaction environment. Additional studies have been related to using an inscribed structure, backfilled with chemically active thiols, for the selective immobilization of proteins via electrostatic interactions or covalent binding [253]. Moreover, 3-D nanostructures could be obtained via selective surface reactions [255]. Depending on the resist layer on gold (such as ODT or 11-mercapto-1-undecanol), either positive or negative nanopatterns could be created, with the obtained features being backfilled with either an alkylthiol or an −OH-terminated thiol compound. In a third fabrication step, the −OH functionalities were reacted with OTS to construct structures in the third dimension.

Biorelated research in the field of nanografting – for example, concerning the fabrication of protein or enzyme patterns – has been conducted by several groups [256, 258, 260–268]. One such example included the incorporation of parallel, three-helix bundle metalloproteins on a gold surface via nanografting [260], while another described a double-cysteine-terminated maltose-binding protein that could be immobilized onto Au substrates at well-defined locations, with subsequent investigations being carried out using *in situ* AFM friction measurements to characterize the bioactivity of the protein products [262, 263]. Further investigations

were conducted, using AFM force–compression measurements, to probe the ligand-induced changes in the mechanical properties of the maltose-binding proteins. In this case, both positive and negative patterns were created containing –OH functionalities on gold. These hydroxyl groups would react with trichlorosilanes to build bilayer systems. Thiolated, single-stranded DNA could also be patterned in this way to provide biostructures with nanometer resolution, that might potentially be of interest for fabricating DNA biosensors and biochips [256, 258]. Further studies addressing the construction of 3-D protein nanostructures were detailed by Abell *et al*. [269], whereby an alkanethiol, terminated with a hexa(ethyleneglycol) group and self-assembled onto Au, was used as a base for the nanopatterning of three differently charged thiol compounds in the construction of multifunctional, nanometer-sized features. For this, various proteins were immobilized onto the surface to investigate the pH-dependency of protein adsorption. DeYoreo and coworkers reported on the structuring of virus particles on gold surfaces [270] which could be modified either genetically or chemically and attached onto the SAM. Additional studies were carried out on the covalent linkage of oligonucleotides on nanometer patterns for the formation of Pd crystals. In the field of nanochemistry, further investigations were conducted by using nanografting to fabricate the nanometer-sized features of maleimide [271]. Such nanoscale structures could then be coupled via a Michael addition with *p*-xylylenediamine to obtain free amine functionalities, which would further react with 11,11′-dimaleimidoundecyldisulfide. This step-by-step surface chemistry led to the build-up of 3-D features. A modified variant of nanografting – the so-called "nano-pen reader and writer process" – which was developed by Liu and coworkers [272] combined nanografting on a thiol-terminated gold surface with DPN. The tip to be used for patterning of the SAM was precoated with another thiol compound to create structures with multiple components. The construction of large patterned areas was also described by the groups of both Liu and Garno, using an automated nanografting approach [120, 273]. Notably, the group of Liu described computer-assisted design and automated vector SPL, while Garno's group focused their attention on the mechanics of automated nanografting, and demonstrated results for the different writing strategies.

The technique of nanografting is not limited to thiols on gold surfaces, but can also be applied to silicon substrates and self-assembled silane monolayers [274, 275]. Linford *et al*. described the successful patterning of monolayers of octadecyl- and octyldimethylmonochlorosilane on thin and thick silicon substrates, using an AFM tip, with the inscribed structures being backfilled with perfluorinated silanes and aminosilanes. The $-NH_2$ functionalities introduced were then used for the attachment of DNA strands and Pd cations, which could in turn be applied to testing of the electrical properties of nanoscale objects.

To summarize, nanografting represents a powerful technique for the fabrication of nanometer-sized features, with high spatial resolution. One advantage of nanografting in comparison to nanoshaving is the *in situ* backfilling of the created structures with chemically active self-assembling molecules. Consequently, this approach can be used for biorecognition and protein immobilization approaches, as well as for the investigation of reaction kinetics and mechanisms of surface reactions. Moreover,

the technique has the potential to create a variety of different structured monolayers with various functional elements. Yet, the disadvantages of nanografting are similar to those of nanoshaving, as it is equally necessary carefully to adjust the force applied to remove the SAMs, so as to prevent destruction of the substrate and of any remaining material. It is also important that the molecules being used to backfill the structures show a higher adsorption rate than the removed molecules, in order to prevent the formation of mixed monolayers, or the adsorption of the initial SAM molecules.

2.3.3 Electro-Oxidative Lithography

In electro-oxidative lithography processes, a bias voltage is applied between a substrate and the AFM tip, and this results in the creation of a localized electric field. This field can lead to physical and/or chemical modifications of the substrate or the SAM (as outlined in the introduction), which can be used to structure surfaces in the nanometer range. Thus, electro-oxidation has been defined as two discrete areas: (i) LAO, which, as described above, provides the possibility of generating oxide patterns on different substrates; and (ii) electro-oxidation, which focuses on the chemical modification of terminal end groups on a SAM.

2.3.3.1 Local Anodic Oxidation

At this point, electro-oxidative patterning will be reviewed with special emphasis on the possibility to structure and/or locally modify SAMs. Next to the anodic oxidation of semiconductors and metals, the method can also be applied to pattern molecularly functionalized surfaces. This area of LAO includes SAMs on silicon and gold, Langmuir–Blodgett (LB), and polymeric thin films. The use and patterning of these functionalized surfaces permits the introduction of various chemical functionalities in the patterned areas, so as to fabricate multifunctional surfaces. A brief overview of LAO on various modified substrates is provided in the following subsections.

LAO on SAMs on Silicon One of the first reports on the oxidation of organosilane-terminated monolayers was provided by Sugimura [276, 277], where a trimethylsilyl (TMS) SAM on Si was used to fabricate silicon oxide patterns. These oxide features were later etched in a mixture of $NH_4F/H_2O_2/H_2O$, which resulted in the formation of nanometer-sized grooves. The patterning was carried out with a conductive AFM tip with positive as well as negative bias voltages. Further studies were performed to investigate the effect of humidity, bias voltage, and probe scan rate on the degradation of the TMS monolayer. Further oxidation experiments included the use of octadecyldimethylmethoxysilane [278, 279], octadecyltrimethoxysilane [280], 1-dodecene [281, 282], 1-octadecene [283] and OTS [284, 285] monolayers to investigate the growth of silicon oxide features on various SAMs. Besides the oxidation of these $-CH_3$-terminated monolayers, which are chemically inert, other monolayers with functional end groups were also applied to LAO. For example, Zheng and coworkers reported on the anodic oxidation of thiol-terminated

monolayers on Si [286]; this involved the use of Au nanoparticles being attached to the −SH functionalities, as a lithographic mask, and preventing oxidation of the covered areas. This approach offered the possibility of controlling the size of the inscribed feature, simply by changing the size of the nanoparticles. Later, the group of Fréchet demonstrated the high-resolution anodic oxidation of SAMs of dendrimers on silicon and titanium [287, 288]. The dendrimers were modified either with a chlorosilane group or with a triethoxysilane group to enable self-assembly onto the substrates, and this allowed features with dimensions less than 60 nm to be fabricated on silicon. TiO_2 patterns were also created with line widths of 25 nm, heights of 12 nm, and spacing between individual lines of 50 nm; these monolayers could be used as both positive and negative tone resists in SPL. Other examples have demonstrated the oxidation of ester-terminated SAMs, for example, with methyl 10-undecenoate [283], amine-functionalized monolayers [289–291] and bromine-modified SAMs (C. Haensch *et al.*, unpublished results). Together, these examples not only demonstrate the variety of surface moieties that can be patterned, but also represent interesting candidates for the fabrication of bifunctional and multifunctional surfaces. Shortly after the first anodization of SAMs on silicon, some interesting post-modifications of the inscribed structures were reported which demonstrated the potential of this technique for the fabrication of molecular assemblies. As an example, Sugimura *et al.* demonstrated the combination of LAO with the self-assembly of organosilane molecules [292]. In this case, following the oxidation of a TMS monolayer, the created patterns were modified with a layer of (3-aminopropyl)triethoxysilane (APTES) to introduce chemical functionalities. These amino moieties were labeled with aldehyde-modified fluorescent latex nanoparticles, and investigated using fluorescence optical microscopy to reveal the selective attachment of particles on the $-NH_2$ groups. Sugimura and coworkers also described the fabrication of a coplanar nanostructure which consisted of a surface pattern of octadecyltrimethoxysilane and fluoroalkylsilane [293]. These structured features were characterized using Kelvin probe force microscopy to investigate their surface potential properties. When the fabrication of positive and negative patterns was demonstrated by Graaf *et al.* [294], a pattern of dodecyl and silicon oxide features was used for the self-assembly of different molecules and nanomaterials, namely R6G molecules and CdSe/ZnS nanocrystals. The R6G was selectively self-assembled on the oxide structures due to electrostatic interactions, whereas the nanoparticles (which were surrounded by a hydrophobic shell) self-assemble only on the alkyl-terminated layer. The nanoscale deposition of manganese single-molecule magnets on silicon oxide features was presented by Martínez and coworkers [291]. Here, the magnets were deposited site-selectively onto the SiO_2 features due to the positive charges of the magnets, and no self-assembly onto the $-NH_2$ structures was observed. The site-selective self-assembly of gold nanoparticles on the pattern of OTS and amine-terminated oxide features was described by Li and coworkers [285], while an interesting method for detecting metal ions by using LAO was provided by Kim and coworkers [295]. A pattern of (3-mercaptopropyl)trimethoxysilane (MPTMS) and APTES was used for the self-assembly of Au nanoparticles, which self-assembled only on the −SH functionalities. This part of the surface served as the fixed electrode,

whereas the amine groups were used for the self-assembly of other metal ions (i.e., Cu^{2+}). Notably, the conductance of the resultant structures depended on the concentration of the metal ions. When He *et al.* reported on the site-specific growth of SWCNTs on Si [290], the iron nanoparticles were selectively assembled onto oxide features due to electrostatic interactions, whereas the amine-terminated background layer was modified with sodium dodecyl sulfate to prevent any unspecific self-assembly of the Fe particles onto the $-NH_2$ groups. These particles subsequently acted as a catalyst for the growth of SWCNTs via chemical vapor deposition (CVD). These findings might be important for the development of SWCNT-based electronic devices. Other interesting research studies have focused on the alignment and stretching of λ-DNA wires into parallel patterns of OTS and $-NH_2$ functionalities [296, 297]. For this, the DNA wires were self-assembled selectively onto the inscribed features, due to coulombic interactions between the amine groups of the substrate and the phosphate backbone of the DNA. Other biorelated examples were demonstrated by Yoshinobu and coworkers [289], who utilized the anodic oxidation of OTS and APTES monolayers for the selective patterning of proteins in the fabrication of positive- and negative-tone structures. Lee *et al.* employed the oxidation of a monolayer of octadecyldimethylmethoxysilane fabricating nanopatterns for the modification with polymer brushes [298]. In this case, a ruthenium-based metathesis catalyst was assembled onto the silicon oxide features to serve as an initiator for the surface-initiated ROMP. Two different monomer units were tested for the site-selective growth of polymer brushes, and the obtained features were characterized using electric force microscopy.

Further interesting non-organosilane molecules used in the anodization process have included metal phosphate monolayers, fabricated via a two-step procedure [279, 299, 300]. This involved the reaction of an –OH-terminated silicon substrate with $POCl_3$ and subsequent reaction with, for example, Zr^{4+}, Hf^{4+}, Ca^{2+}, and Mg^{2+} ions. Such reaction with tetravalent metal ions led to the fabrication of positively charged phosphate monolayers, whereas with divalent metal ions the surface appeared to be neutral. A comparison of the two surfaces revealed a need for a lower threshold voltage for oxidation of the positively charged surfaces than for the neutral substrates.

The group of Lee investigated the anodization of SAMs of 1,12-diaminododecane·dihydrochloride (DAD × 2HCl) and *n*-tridecylamine·hydrochloride (TDA × 2 HCl) [122, 279, 301–304] and, in particular, the influence of the surface functionalities on the anodization process. When mixed monolayers of DAD × 2HCl and TDA × HCl were tested for this purpose, the DAD × 2HCl monolayer led to a positively charged surface due to the presence of the ammonium chloride moieties, whereas the TDA × HCl SAM was uncharged, because of the terminal $-CH_3$ groups. The terminal end group of DAD × 2HCl was found to lead to an enhancement of the oxidation [301]. Moreover, the positively charged surface of the DAD × 2HCl SAM was responsible for an increase in the line widths and the heights of the inscribed structures, an effect which could be explained by an enlargement of the water meniscus between the AFM tip and the substrate, due to surface charges [122, 303].

LAO on SAMs on Gold Jang and coworkers described the characterization of the desorption and oxidation of thiol-terminated gold substrates [305], having first prepared the gold surface by depositing a layer of Au onto silicon. Depending on the monolayer, two processes can occur and result in the formation of two different patterns: (i) removal of the SAM with the formation of recessed structures; and (ii) the formation of silicon oxide structures. Jang *et al.* investigated eleven different monolayers on Au, including $-CH_3$-, $-COOH$-, $-PO_3H_2$-, $-OH$-, $-NH_2$-, $-CF_3$-, and $-(OCH_2CH_2)_3OH$-terminated SAMs. The parameters that controlled the mode of patterning were the SAM chain length, the functional end group, the bias voltage, the local pH value, and the hydroxide anion accessibility.

LAO on LB Films Investigations conducted by Bourgoin and coworkers highlighted the potential to oxidize not only SAMs on Si or Au, but also organic films prepared by the LB technique [306]; indeed, phthalocyanine LB films could be patterned with line widths down to 50 nm. Further investigations were carried out on LB films of palmitic acid to study the bias dependence of the LAO process [307] and, by changing the applied voltage, it was possible to create both positive and negative patterns. Mixed LB films of hexadecylamine and palmitic acid were also studied to analyze the mixing and charge effects on the anodization, and to investigate the mechanism of the patterning process [81, 308].

LAO on Polymeric Films The anodic oxidation of polymeric films offers the possibility to construct patterns that can be used for the site-selective self-assembly of protein molecules. This was shown by Yam *et al.*, who oxidized oligo(ethylene glycol)-terminated films on silicon to test the specific adsorption of fibrinogen, avidin and BSA onto silicon oxide patterns [309]. Choi and coworkers described the *in situ* observation of biomolecules [310], whereby a methoxy-PEG-terminated monolayer was oxidized and served as a passivation background, whilst the patterns were used for the site-selective immobilization of streptavidin, labeled with Au particles, and pure streptavidin. The nonlabeled streptavidin patterns were further investigated for the detection of biotinylated materials. These findings demonstrated the potential of polymeric films for applications in biosensing devices.

The anodic oxidation process represents an interesting method for the construction of nanometer-sized features on a large variety of different substrates. In addition to the widely investigated silicon substrates, other semiconductors and metals formed the focus of these studies. In order to introduce chemical functionalities, LAO was applied to organic layers of various materials on silicon surfaces, which in turn opened the possibility of functionalizing the surface with multiple chemically active groups. Potential applications of these patterns may be found in the fabrication of metal-oxide-semiconductor transistors and biosensor devices, and also for investigating chemical processes at the nanometer scale. One disadvantage of this patterning technique, however, is the possible destruction of the underlying substrate, and the process must be very carefully tuned. The technique is also rather slow, such that its use in high-throughput strategies would not easily be realized.

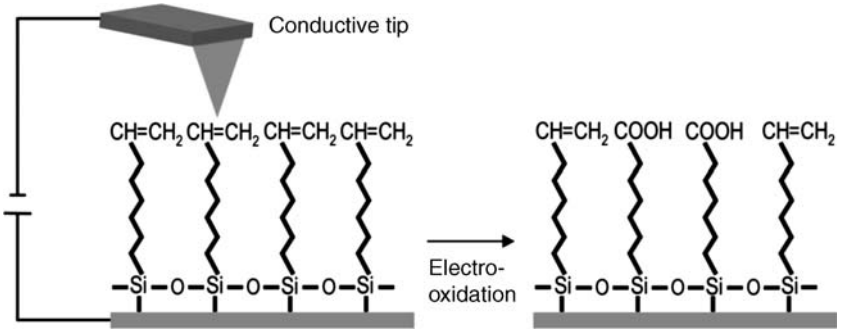


Figure 2.9 Schematic representation of the electro-oxidation on self-assembled monolayers of 18-nonadecenyltrichlorosilane on silicon.

2.3.3.2 **Chemical Activation of Self-Assembled Monolayers**

In 1999, Sagiv introduced another electro-oxidation process of SAMs, namely "constructive lithography" or the electro-oxidation process [311]. In contrast to the above-described anodization process, the monolayer will not be degraded while the features are inscribed, but the terminal end groups will be chemically activated. In the first of these experiments, which was carried out with a monolayer of 18-nonadecenyltrichlorosilane (NTS), a conductive AFM tip was used to pattern the SAM, and the terminal $-CH{=}CH_2$ groups were converted to carboxylic acid functionalities (Figure 2.9). A bilayer of OTS was then self-assembled onto the $-COOH$ groups to create multilayer structures. Depending on the voltage applied, the patterning process would induce an electro-chemical surface transformation of the terminal end group of the monolayer, although the overall structure of the monolayer would, ideally, not be influenced.

This process was later applied also to OTS monolayers on silicon [312], where the terminal $-CH_3$ groups were converted to $-COOH$ moieties. Transformation to the carboxylic acid functionalities was followed by AFM and Fourier transform infrared (FT-IR) spectroscopy. Subsequent contact-mode AFM measurements allowed characterization of both the topographic and (especially) the frictional changes of the oxidized monolayers. FT-IR spectroscopy on a macroscale structured surface showed the $-CH_2$ vibrations to be preserved, but the terminal $-CH_3$ vibration to be greatly reduced. A further indication of carboxylic acids formation was the appearance of an absorption peak for $-C{=}O$ at $1713\,cm^{-1}$. As an additional surface-sensitive technique for investigating chemical transformations, time-of-flight secondary ion mass spectrometry (TOF-SIMS) represents an interesting alternative [313]. For this, Pignataro and coworkers used patterns of a 1-octadecene monolayer, self-assembled on hydrogen-terminated silicon, to analyze the transformation process [314], and obtained both elemental and molecular information concerning the chemical features after modification. The modified areas showed the presence of C_xH_yO- and C_xH_yN-type peaks, both of which increased with in line with higher bias voltages and were related to the formation of organic polar moieties. In contrast, a reduction in the SiC_xH_y signal was observed. The characterization of micrometer-sized features

can also be addressed by using XPS to analyze the chemical structure. For example, Andruzzi *et al.* used a micrometer pattern of OTS and PEG for their XPS analysis [315].

Nonetheless, the characterization of nanometer-sized structures is, in general, a difficult process, and very few techniques allow investigations to be made of the chemical state of the features. Furthermore, the sensitivity and resolution of the different techniques are frequently insufficient. Consequently, the most frequently used technique for the indirect characterization of nanometric features is that of AFM, which demonstrates mainly the changes in frictional and topographic properties following the oxidation process. When Wouters *et al.* used AFM to investigate the influence of applied voltage and pulse duration on the electro-oxidation process of OTS on silicon [316], the formation of silicon oxide structures resulted in an increase of the topographic image. Moreover, the carboxylic acid-terminated structures showed a change in height, depending on the direction of the scan (this was in fact correlated to a crosscoupling of the friction and the height signal). The change in height was used as an indication for the formation of −COOH moieties, whilst degradation of the monolayer under harsher conditions was associated with a detectable positive change in the height images, independent of the scan direction. The dependence of oxidation time versus the bias voltage is shown in Figure 2.10, where a small window for the oxidation of OTS was observed. Above a certain threshold – that is, at high pulse duration and bias – the monolayer was degraded and silicone oxide formed, but below the threshold line no oxidation had occurred, due to short pulses and low bias voltages.

Hoeppener *et al.* described a series of AFM studies to investigate surface properties during the oxidation process, and at the transition state of monolayer oxidation and degradation [318]. In general, the writing process itself was seen to depend heavily on

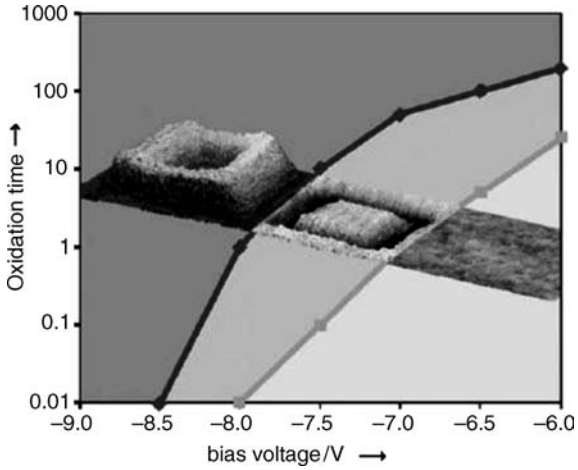


Figure 2.10 Dependence of oxidation time on bias voltage. Reproduced with permission from Ref. [317].

a number of parameters, while the choice of conducting AFM tip and its geometry had a critical influence on the required oxidation voltages and times, and also affected the size of the inscribed structures. In order to reduce the size of the patterned areas, the tip diameter could be reduced by using highly doped and/or noncoated tips. The humidity of the environment during the writing process was identified as another important issue; whilst a low humidity resulted in longer writing times and incomplete pattern formation, a too-high humidity resulted in wider line widths.

One limiting factor when implementing electro-oxidation into high-throughput strategies is that the actual process is rather slow, though this might be overcome by using an automated writing process with a software-driven AFM set-up. The use of such a system with conductive parallel cantilever arrays was demonstrated by Wouters *et al.* for fabricating large areas of oxidized surface patterns [127]. By using an automated oxidation set-up, 1000 circles could be inscribed onto an OTS monolayer on silicon. Following the *in situ* imaging of the features, it was clear that the conductive coating of the tip had not been degraded during the patterning, as indicated by a constant line width and the intensities of the friction signals measured on the circles. Moreover, these structures could be used for the self-assembly of nanoparticles, which occurred exclusively onto the oxidized areas, rather than on the OTS background. Another way to increase the patterning speed would be to use an array of four cantilevers for the oxidation process; in this way, a successful oxidation of square structures was demonstrated on OTS, with the subsequent self-assembly of CdSe/ZnS particles. Following this, Cai and coworkers introduced an alternative method of reducing the writing and fabrication times of nanometer-sized patterns [319], when they described the process of electro-pen nanolithography (EPN). This was a combination of electro-oxidation with an AFM tip and DPN, in which an ink-coated conducting AFM tip was used to oxidize an OTS monolayer. In this way, patterns of $-$COOH could be created, with the ink being transferred directly onto the patterned features due to the higher surface energy of the OTS_{ox} areas. The structures obtained revealed line widths of approximately 50 nm, and were inscribed at a writing speed of 10 μm s^{-1}. In addition, different inks could be used to introduce chemical functionalities onto the patterns, and in particular trialkoxysilanes and quaternary ammonium salts. The self-assembly of MPTMS molecules resulted in an availability of $-$SH moieties, and to prove a successful transfer of the thiol molecules, functionalized gold nanoparticles were self-assembled onto the surface. Subsequent AFM investigations revealed the self-assembly of particles only on the nanopatterns, as well as an increase in height of 2 nm, which was in good agreement with the diameter of the Au particles. This process might have a future role in the fabrication of 3-D structures and, depending on the ink used, also for the direct writing of biological patterns. In order to obtain areas in the range of square micrometers, the AFM tip could be exchanged with a conductive metal stamp, which enabled the pattern of the grid to be transferred onto the SAM [320]. Hoeppener and coworkers later described an oxidation process which employed a metal stamp, composed of a Cu transmission electron microscopy (TEM) grid, onto SAMs of OTS on silicon. Following its exposure to a saturated water vapor

atmosphere, the copper grid was pressed against the monolayer-terminated surface. Compared to the oxidation process using an AFM tip, the inscription with a TEM grid required longer oxidations times of approximately 30 s, and a higher voltage of 30–35 V. The need for longer patterning times and higher voltages might be due to the larger distance between the TEM grid and the surface, and also to the presence of a thicker water layer. An alternative approach towards patterning SAMs utilized a patterned monolayer on silicon; this could first be pressed onto another SAM, after which a bias voltage would be applied to transfer the structures onto the unmodified surface [321].

The inscribed structures could, furthermore, be used after the electro-oxidation process for a large variety of different post-modification steps, including the self-assembly of different nanomaterials (e.g., nanoparticles, nanowires), the formation of multilayers, and also for the application of nanometric surface chemistry. An overview of the different possibilities for post-modification of the inscribed features is shown in Figure 2.11, where different driving forces have been used to add nanomaterials or functional molecules to the structures.

Self-Assembly of Additional Silane Molecules onto the Nanopatterns In 2000, Sagiv and colleagues had already suggested many different possible post-modification reactions of the oxidized areas, via a variety of chemical transformations based on pattern functionalization [312]. The $-$COOH functionalities, created during the oxidation process, are suitable for the self-assembly of a second layer of reactive trichlorosilanes, due to the reaction of $-SiCl_3$ groups with carboxylic acid moieties. Examples of this include the self-assembly of NTS [312], 11-bromoundecyltrichlorosilane (BTS) [322], 11-undecyltrichlorosilane (UTS) [325], and others. Wouters and coworkers also demonstrated the self-assembly of quaternary ammonium salts (e.g., trimethyloctadecylammonium bromide) onto the acid structures [325]. The formation of a bilayer can lead to the introduction of different chemical active functionalities. Moreover, these moieties can be used for other modification sequences, such as nanometer-sized surface chemistry and the site-selective self-assembly of nanomaterials (e.g., nanoparticles). Some interesting examples of this are described in the following subsections.

Surface Chemistry on the Nanopatterns One important application of the oxidized areas is their use in surface chemistry, when it is necessary to utilize surface reactions that have high yields and can be carried out under mild reaction conditions, with readily available starting materials. Examples include the reaction of the terminal end groups of NTS monolayers (ethylenic functionalities) with H_2S and $BH_3 \times THF$ (tetrahydrofuran), which leads to $-$SH moieties, whereas oxidation of the ethylenic groups with $KMnO_4$ and KIO_4 creates $-$COOH end groups [312, 327]. The formation of $-NH_2$ functionalities can be obtained by a photoreaction with formamide; this results in the creation of amide moieties, while a subsequent reduction with $BH_3 \times THF$ leads to amine functionalities [328]. One concept that fulfils the criteria for surface reactions is the so-called "click chemistry" approach, as introduced by Sharpless in 2001 [329]. Until now, one of the most important click

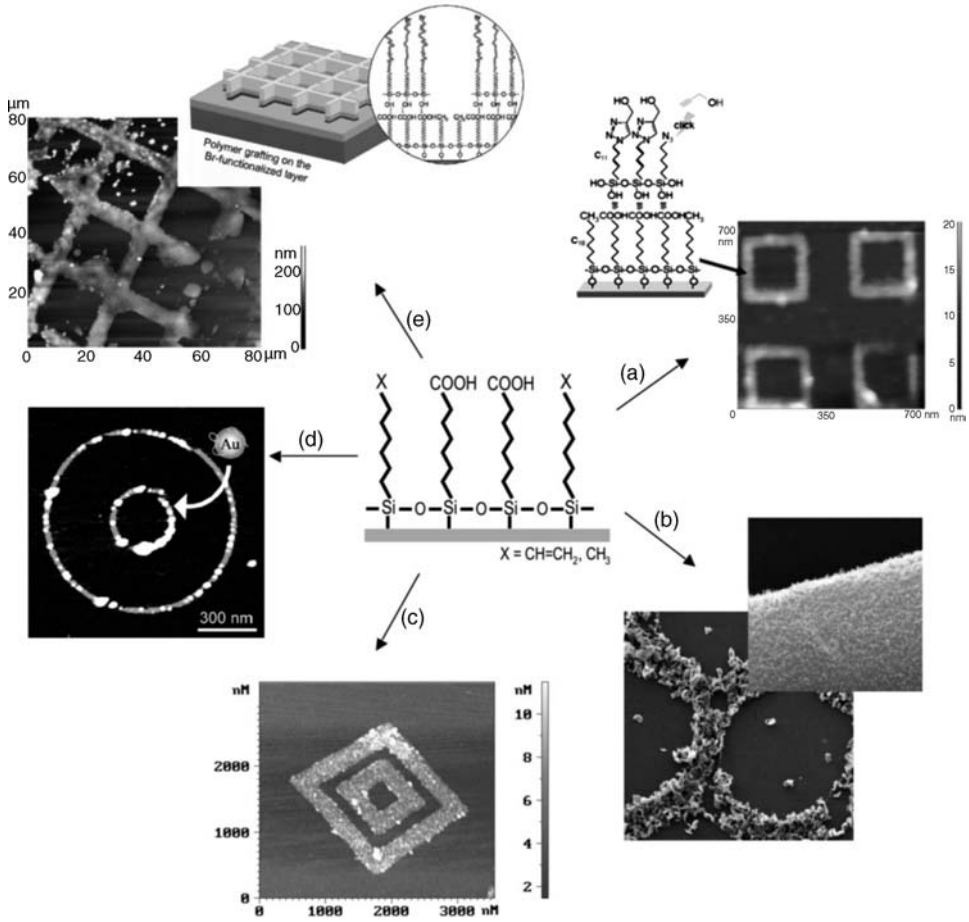


Figure 2.11 Examples of the post-modification of oxidized nanometer-sized structures on SAMs on silicon. (a) Clicking on the nanometer-scale (reproduced with permission from Ref. [322])); (b) Patterned growth of carbon nanofibers [323]; (c) Site-selective assembly of Fe(II) salt (reproduced with permission from Ref. [324]); (d) Self-assembly of Au nanoparticles onto the nanostructures (reproduced with permission from Ref. [325]); (e) Polymer brushes (reproduced with permission from Ref. [326]).

reactions is the 1,3-dipolar cycloaddition of terminal acetylenes and organic azides, which results in the regioselective formation of 1,4-disubstituted triazoles [330]. Haensch *et al.* reported the details of a 1,3-dipolar cycloaddition of nanometer-sized azide structures with propargyl alcohol with line widths of 50 nm on a background layer of OTS [322]. A subsequent characterization of the obtained features was carried out using AFM, and the results were confirmed on nonstructured functional substrates using FT-IR spectroscopy and XPS. The major advantage of this reaction scheme is the wide diversity of clickable moieties, including phosphorescent iridium complexes [331], gold nanoparticles [332], CNTs [333], dendritic systems [334], and

others [335]. Willner and coworkers used the tip-mediated oxidation of monolayers of OTS, and activated the surface templates via a secondary enzymatic reaction for further purpose [336]. The −COOH functionalities formed during the oxidation process were reacted with tyramine, which itself can be oxidized biocatalytically (with the enzyme, tyrosinase) to catechol moieties that control the self-assembly of magnetic nanoparticles and boronic acid-terminated gold nanoparticles. This approach permits the possible fabrication of nanobiosensors and/or nanocircuitry. Andruzzi *et al.* combined the electro-oxidation of OTS with a conductive stamp with NHS chemistry to obtain bioselective patterns of PEG and OTS [315]. Here, carbodiimide NHS chemistry was used to react NHS-terminated micrometer-sized patterns with an amino-terminated PEG, and the PEG/OTS patterns were later applied to protein adsorption studies with fluorescently labeled BSA. The characterization of the features, using fluorescence microscopy, showed a reduced adsorption on the chemically modified lines. Furthermore, although a significant inhibition of cell adhesion onto the PEG patterns was noted, cell growth was maintained on the functionalized areas. Wouters *et al.* reported successive functionalization reactions on patterns of OTS with 40 nm resolution [325], where UTS was used for bilayer formation on the structures, so as to fabricate a pattern for the radical polymerization of styrene. Yet, only a partial increase in pattern height was observed, this being related to the UTS structure possibly undergoing horizontal polymerization. The results presented by these authors might lead potentially to applications in electronics, or perhaps for DNA and protein sensors. The fabrication of polymer brushes on micrometer-sized surface areas was demonstrated by Becer *et al.*, who attached BTS as a chemically active molecule onto the patterns [326]. In this way, the bromine functionality could be used as an initiator for the ATRP of, for example, styrene. This "grafting-from" approach led to the creation of polymer brushes of various heights, depending on the reaction time. Notably, the height of the polymer brush was seen to increase linearly with the polymerization time, and was indicative of a controlled polymerization process. In an additional test to determine whether the polymer terminated with bromine end groups, a second ATRP polymerization with *tert*-butyl acrylate was performed, and this revealed an increase in polymer brush height of about 20–40 nm. Clearly, the functionalization of patterned substrates with defined polymer block systems promises much with regards to biomedical applications, and/or for the creation of responsive brush systems.

Self-Assembly of Different Nanomaterials onto Nanopatterns The self-assembly of nanomaterials onto the structured features can be achieved via two different strategies. For the first strategy, carboxylic acid functionalities were used directly for the attachment of nanoparticles. For example. Hoeppener *et al.* demonstrated the site-selective binding of magnetic Fe_3O_4 particles onto predefined surface areas of −COOH moieties [337]. In this case, the particles self-assembled onto the −COOH features due to hydrophilic interactions of the acid functionalities with the ligand shell. Subsequent treatment of the surface with conventional adhesion tape led to the removal of unspecifically bonded material, but without destroying the created structures. The same method was also used to create nanostructures of Fe particles

by the subsequent reduction of Fe(II) ions assembled onto the oxidized areas [324]; here, the typical particle size was 6–7 nm, with a high degree of uniformity. Following this, magnetic force microscopy (MFM) measurements were conducted to investigate the magnetic properties of the nanoparticles, thus revealing their magnetic origin. The fabrication of such nanosystems permits the possible creation of magnetic structures with decreasing device dimensions. For example, Wouters and colleagues described the self-assembly of positively charged gold nanoparticles onto −COOH patterns [325], after which tapping mode AFM studies revealed an increase in height of 18–20 nm, in good agreement with the diameter of the Au particles. The subsequent successful self-assembly of two differently sized gold nanoparticles onto oxidized nanopatterns was demonstrated by the same group [338], when they investigated the sequential oxidation steps performed on the OTS and the subsequent self-assembly of various nanoparticles. The Au particles, which were self-assembled in an initial step, had to be stabilized (e.g., by thermal annealing at 90 °C for 6 h) to prevent their exchange during a second self-assembly step. Druzhinina and coworkers described the growth of carbon nanofibers and nanotube patterns on OTS/-COOH structures [323]; in this case, iron acetate was assembled onto the acid moieties and subsequently reduced to metallic Fe particles that acted as a catalyst for carbon nanofiber growth under microwave irradiation. Today, applications of these structures can be found in electronic devices.

The second strategy involved the self-assembly of a second layer on top of the −COOH features, before the selective self-assembly process can be performed Previously, examples of surface chemistry conducted on nanopatterns were given to demonstrate the potential of chemical modification schemes for introducing a wide variety of chemical functionalities. Such chemical end groups are suitable for the selective self-assembly of different nanomaterials; for example, thiol and carboxylic acid groups can be used to deposit various materials such as gold, silver, or cadmium selenide and $[Au]_{55}$ clusters [312, 327, 339]. Hoeppener *et al.* demonstrated the self-assembly of Cd^{2+} ions on functionalized −SH nanopatterns [339], where cadmium cations were self-assembled onto thiol groups and were then reacted with H_2S to create CdS particles. The latter particles could then be metalized by treatment with an aqueous solution of $HAuCl_4$, with the formation of Au structures being confirmed by subsequent silver deposition from a silver enhancer solution on the gold patterns. The same authors also demonstrated the creation of millimeter-sized silver electrodes via a monolayer photodesorption with gallium onto OTS monolayers, creating molecularly sharp boundaries in the process. Surface chemistry was then used to functionalize fabricated surface structures also with thiol moieties, so as to assemble silver onto the oxidized structures. Liu *et al.* demonstrated the site-selective deposition of $[Au_{55}(Ph_2PC_6H_4SO_3Na)_{12}Cl_6]$ clusters onto −SH nanopatterns [327]; these features were stable against thermal treatment and structurally robust (e.g., cleaning with "Scotch tape" did not disturb the structures, but any unspecifically bound material was removed). Liu *et al.* reported on the hierarchical self-assembly of colloidal gold particles on silicon [328], where nanopatterns of $-NH_2$ functionalities could be created via the chemical

reaction of ethylenic groups with formamide and $BH_3 \times THF$. Protonation of the amine moieties led to positively charged surface features that were suitable for the attachment of negatively charged particles via electrostatic interactions [340–344]. The latter interactions were responsible for the spontaneous self-assembly of [Au-citrate] particles (which were negatively charged) onto the nanostructures. Such defined molecular templates could be used to fine-tune the distances between nanomaterials which are anchored to the surface, and might also be used in the advancement of 3-D nanofabrication techniques. The construction of hydride metal–organic surface nanostructures was demonstrated by Maoz *et al.* [345], when a monolayer of a thiol-functionalized silane was loaded with silver ions and used to create nanometer-sized structures on specific surface areas via either a chemical or a tip-induced reduction of the Ag ions to metallic nanoparticles. Hoeppener *et al.* demonstrated the preparation of amine-terminated nanostructures for the selective binding of CNTs [346]; in this case, the $-NH_2$ moieties were obtained by a vapor-phase self-assembly process of 3-aminopropyltrimethoxysilane (APTMS) onto the electro-oxidized surface areas, while the CNTs selectively self-assembled onto the amine features. These investigations might have important implications for the future, notably in the field of nanoelectronics.

Wetting-Driven Self-Assembly Concept Another versatile approach to the template-guided fabrication of metal nanopatterns is that of wetting-driven self-assembly, as recently introduced by Sagiv and coworkers [347]. This process utilizes the selective adhesion of nanosized volumes of wetting liquids to the lyophilic surface structures of a lyophilic/lyophobic substrate, where the lyophilic/lyophobic surface consists of a pattern of $-COOH$ areas versus OTS. The patterns were fabricated by retraction of the COOH/OTS surface from the melt of three different compounds, namely eicosene, ODT, and dodecanoic acid. This method has allowed the introduction of various chemical functionalities by the use of readily available and cheap functional alkanes. Post-modification of the obtained terminal surface groups led to templates for the site-specific self-assembly of metallic gold and silver. Checco *et al.* used the strong hydrophilic/hydrophobic contrast between oxidized structures and the OTS monolayer to study the wetting behavior of ethanol and octane on patterned line features [348, 349]. This allowed the precise control and stabilization of liquid objects in desired confinements, enabled studies of the wetting phenomenon to be conducted, and allowed the determination of liquid profile shapes with sub-100 nm resolution. Cai *et al.* studied the liquid-behavior at the nanometer level on iodine patterns which had been inscribed by the electro-oxidation process [350]. As iodine serves as a good tracing and visualizing agent in this type of study, such investigations may provide an understanding of the evaporation dynamics of liquid solvents on nanometer-sized structures. Depending on the deposition method used, the nanometer-sized structures can be either gel-like (solution–deposition method) or dendritic, snowflake-shaped polycrystalline iodine sheets (vapor-phase–condensation method).

These examples stress the versatility of the structuring approach, and are potentially compatible with other surface-chemistry schemes.

2.3.4
Catalytic Lithography

The fabrication of nanometer-sized structures is possible not only via the electro-oxidation or anodization of SAMs, but also by catalyzing a surface reaction with an AFM tip [351]. As the reaction takes only place in those areas where the tip is in contact with the surface, this approach will lead to a selective functionalization of the terminal end groups, without applying an electrical current and without destroying the underlying monolayer [352]. Hence, this technique is not limited to the use of conducting substrates; rather, a wide variety of different substrates can be used. Despite the fact that the technique represents a promising approach to performing nanometric surface chemistry, very few examples have been reported to date. Müller *et al*. were among the first to describe the use of an AFM tip coated with a catalyst to perform nanochemistry [353], and to modify the surface functionalities of spatially defined areas on a silicon substrate. In this case, the hydrogenation of an azide-terminated silicon surface with a platinum-coated AFM tip was investigated as a model reaction; the reaction scheme is depicted in Figure 2.12, where the terminal $-N_3$ groups were converted to amine functionalities that could be used for further modification sequences to yield more complex structures. Fluorescence labeling with fluorescein-labeled, aldehyde-modified latex beads or 3-(2-furoyl)quinoline-2-carboxaldehyde (ATTO-TAG) was chosen as post-modification reaction. When investigations of the modified areas were performed using confocal scanning laser microscopy, the measurements revealed brightly fluorescent squares that represented the reacted surface areas, but imaging of the azide-terminated surface, after derivatization with the fluorescent compounds, showed no signal. These studies led to the development of a general approach that used AFM coated tips to perform nanometric catalytic surface chemistry.

The catalysis of chemical surface reactions by a palladium-coated AFM tip for the fabrication of nanometer-sized features was demonstrated by Blackledge *et al*. [352].

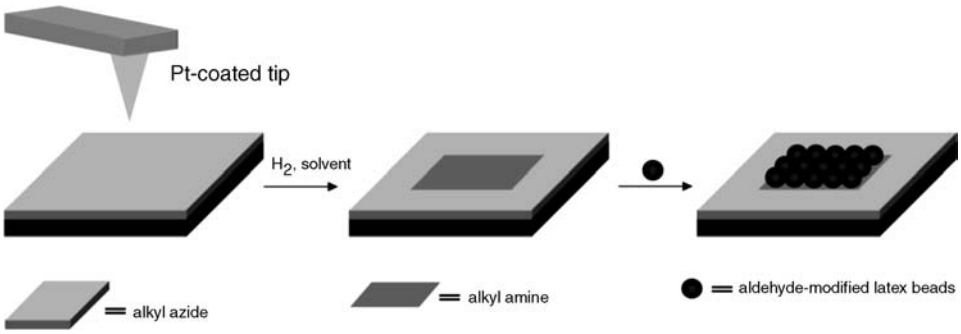


Figure 2.12 A Pt-coated AFM tip is scanned over an azide-terminated substrate in the presence of H_2, resulting in the formation of $-NH_2$ groups. The obtained amine functionalities were modified with fluorescein-labeled latex beads or ATTO-TAG to yield a site-selective self-assembly of fluorescent dyes on specific surface areas.

Two example reactions included the hydrogenation of azide groups and *N*-benzyloxycarbonyl-protected amines to $-NH_2$ functionalities, while a third reaction scheme showed the addition of aminobutyldimethylsilane (ASiH) to terminal carbon–carbon double bonds. The inscribed amine end groups were subsequently labeled with 5- and 6-carboxytetramethylrhodamine succinimidyl esters in order to conduct fluorescence measurements on the obtained structures. The reaction mechanism described proposed the formation of a reactive palladium–organosiloxane intermediate that could only be formed if the monolayers were to be deformed during the structuring process. The Langmuir–Hinshelwood mechanism, which was suggested as a possible reaction model, includes chemisorption of the terminal end group of the monolayer and H_2 or ASiH on the Pd-coated tip, and a subsequent reaction. Davis and colleagues described surface-confined Suzuki and Heck carbon–carbon coupling reactions under Pd catalysis [354, 355]; the schematic outline of the surface reactions is depicted in Figure 2.13a and b. In this case, 4-bromo-*N*-(3-(methylthio)-propyl)-4-vinylbenzamides and a styrene (*N*-3(methylthio)-propyl)-4-vinylbenzamide, self-assembled on gold, were used as functional monolayers for the spatially controlled surface modification sequences. The Suzuki reaction was performed with a polyvinylpyrollidone (PVP)-Pd nanoparticle-functionalized AFM tip at a pressure of 15 to 25 nN, and in a reaction solution of methanol, sodium acetate and 3-aminophenylboronic acid or phenylboronic acid. The Heck reaction was performed in a solution of dimethylformamide (DMF) with sodium hydrogen carbonate and 4-iodobenzoic acid under a pressure of 25 to 40 nN, which led to line widths of 12 to 15 nm being achieved. The successful reaction sequences were proven by a combination of fluorescence tagging, frictional imaging, and labeling with appropriate nanoparticles. The fluorescence image of a Suzuki-catalyzed square functionalized with NHS–fluorescein revealed a brightly fluorescent rectangle (see Figure 2.14a), while the AFM height image of a Suzuki-catalyzed square functionalized with aldehyde-terminated nanospheres that were attached only to the inscribed structures, is shown in Figure 2.14b.

The reduction of a monolayer of imines, using an AFM tip coated with a reducing agent, was investigated by Blasdel *et al.* (Figure 2.13c) [356]. For this, sodium triacetoxyborohydride ($Na(OAc)_3BH_4$) was used as reducing agent to form the corresponding secondary amines, the generation of which was confirmed using a chloranil test. The colorless amine was then reacted with acetaldehyde and tetrachloro-*p*-benzoquinone to yield a bright blue tertiary amine. Visualization of the inscribed structures was achieved using inverted optical microscopy under bright-field illumination.

Bis(ω-*tert*-butyldimethyl-siloxyundecyl)disulfide (TBDMS) monolayers on gold could be hydrolyzed in the contact areas of an acidic tip to create nanopatterns of hydrolyzed TBDMS SAMs [357]. In this case, 2-mercapto-5-benzimidazole sulfonic acid was attached to the Au tips as the catalytic species, such that inscribed features with a line width of approximately 25 nm were obtained. Cleavage of the bulky terminal end group led to the formation of spaces between the residual monolayer that could be refilled with dendritic wedges. The corresponding AFM height images revealed an increase in height of 1.3 nm, indicating a successful filling with the

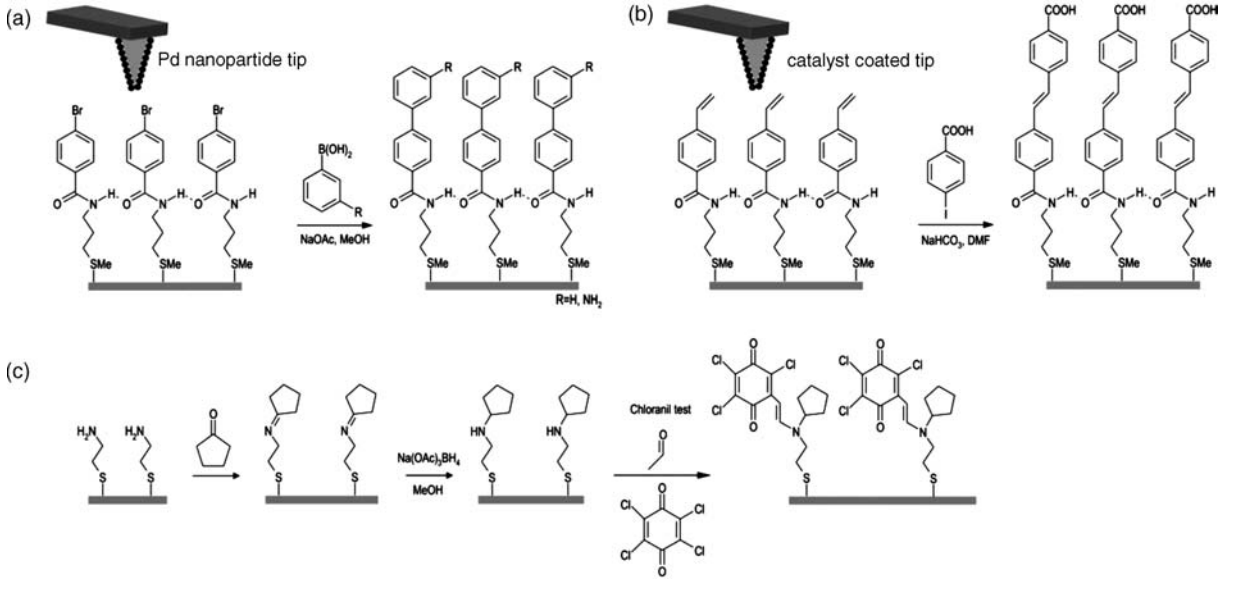


Figure 2.13 (a) Suzuki reaction between an aryl bromide monolayer and phenylboronic acid/3-aminophenylboronic acid in methanol and sodium acetate with a Pd nanoparticle-coated AFM tip; (b) Heck reaction between an aryl styrene monolayer and an aryl halide in DMF and sodium hydrogen carbonate with a catalyst-coated AFM tip; (c) An amine-terminated monolayer was reacted with cyclopentanone to yield an intermediate imine, which was scanned in methanol and $Na(OAc)_3BH_4$ with an AFM tip to react to secondary amines; the chloranil test revealed the formation of bright-blue tertiary amine functionalities.

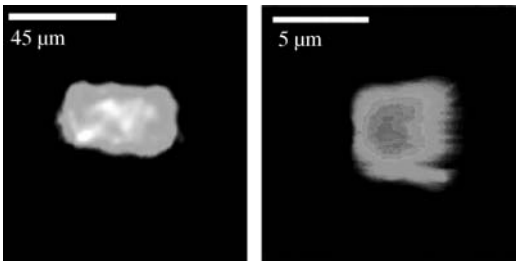


Figure 2.14 (a) Fluorescence image of a Suzuki-catalyzed square functionalized with NHS-fluorescein; (b) AFM height image of a Suzuki-catalyzed square functionalized with aldehyde-terminated nanospheres. (Reproduced with permission from Ref. [354]).

dendritic ligands. Zorbas and coworkers described the photochemical reaction of a dye layer with a chemically modified AFM tip [358]; here, the dye used was a commercially available Procion Red MX-5B that was oxidized under UV-mediated photocatalysis, while the catalytic species were TiO_2 particles attached to an AFM probe. Subsequent AFM, optical microscopy and mid-FT-IR investigations revealed the photocatalytic degradation of the dye molecules. The fabrication of nanometer-sized features by applying a Diels–Alder reaction onto silicon was shown by Matsubara *et al.* [359]. For this, a monolayer terminated with alkene functionalities was reacted with an AFM tip that had been coated with a solution of 2-(13-hydroxy-2-oxatridecanyl)furan. As a consequence, a force of 32 nN and a writing speed of $0.2\,\mu m\,s^{-1}$ were sufficient to successfully couple the furan molecules onto the terminal surface groups. Wang and coworkers demonstrated the tip-assisted hydrolysis of a monolayer of dithiobis(succinimidoundecanoate) on gold [360], where the base hydrolysis of the ester-terminated surface was accelerated after contact mode imaging, due to the implementation of a disorder into the structure. It appears that hydroxide ions have an easier access to the acyl carbon atoms of the monolayer, and this resulted in an accelerated hydrolysis. A localized click chemistry was reported by Long *et al.* [361], who used an acetylene-terminated surface as the template for a nanometer-scale cycloaddition reaction. This was achieved by using an AFM tip immersed in a solution of an azide reagent and the Cu catalyst. Later, a new and versatile patterning approach – the so-called thermochemical nanolithography (TCNL) – was introduced by Szoszkiewicz *et al.* [362]. In this case, a heatable AFM tip was used to inscribe features onto polymeric substrates with line widths in the region of 12 nm. The deprotection of an ester moiety to acid functionalities was tested as a possible model reaction, while the change in hydrophilicity was investigated using LFM. The parallelization of individually addressable tips was also demonstrated for the preparation of large-scale patterning.

The combination of catalytic chemistry with the application of bias voltages has led to some interesting methods for creating nanometer-sized surface patterns. As an example, Fresco and coworkers demonstrated the chemical activation of protected amine and thiol surfaces by applying a bias voltage between the substrate and the AFM tip [363, 364]. In this case, the α,α-dimethyl-3,5-dimethoxybenzyloxycarbonyl

(DZZ) group was selected as a protective moiety, due to the mechanism of cleavage of this functionality via an ionic intermediate, with 3,5-dimethoxy-α-styrene and carbon dioxide being released and primary amine and thiol functionalities being obtained. The inscribed amino groups were used in further modification steps for the reaction with a dendrimer and a dendronized polymer. The AFM images revealed increases in height of 1 nm and 4 nm, respectively, while the line widths of the structures were broadened due to the flexibility of the polymer chain. Furthermore, exposure of the thiol functionalities to a solution of gold nanoparticles resulted in the selective self-assembly of the particles onto the −SH groups. Consequently, the placement of single particles could be achieved by a programmed application of electrical pulses to the surface. The electro-oxidation of thiol-terminated monolayers on both nonstructured and nanometer-sized surfaces, by applying a positive voltage to the surface, was reported by Pavlovic *et al*. [365, 366]. This activation resulted in the formation of thiolsulfinates and thiolsulfonates, which could be used for the covalent immobilization of biomolecules; release of the biomolecules was detected by treating the surface with a disulfide-cleaving reagent. Sugimura and coworkers introduced the reversible nanochemical conversion of amino-terminated monolayers with an AFM tip [367, 368] when, by applying a positive bias voltage to the surface, the terminal $-NH_2$ groups could be converted to −NO functionalities. The nitroso-terminated SAM could then be reduced to $-NH_2$ moieties by changing the voltage to negative values, and surface potential measurements confirmed the chemical surface reaction. The fabrication of metallic structures on the nanometer scale via an AFM tip was achieved by Li *et al*. [223]. For this, a positive bias voltage was applied to a tip, which had been coated with H_2PtCl_6; the latter then dissolved in the water meniscus, such that the platinum(IV) was reduced to metallic Pt. This protocol demonstrated an interesting approach to using the water meniscus as a reaction vessel for a wide range of different surface reactions.

To summarize, these examples have demonstrated the potential of catalytic lithography for the chemical functionalization of nanometer-sized features. Moreover, a range of surface reactions can be applied to yield a wide variety of terminal end groups, while the modified functionalities can be used for fluorescent labeling, for the self-assembly of nanoparticles, and for the preparation of metallic nanostructures. Moreover, this approach enables a preservation of the underlying substrate and also of the monolayer, which is not in contact with the AFM tip.

2.4
Surface Chemical Reactions for Structured Surfaces

The key feature of the previously introduced patterning is the combination of surface structures with functional moieties. Clearly, the use of SAMs represents a versatile method for implementing a wide variety of chemical functions that provide access to surface reactions conducted on the nanometer scale. Moreover, a whole range of chemical interactions, including electrostatic and covalent binding, hydrogen bonding, hydrophilic/hydrophobic interactions and complex formation, are available to

attach and stabilize nanomaterials to the structures. Additional possibilities for expanding the capabilities of a combination of lithography and functional SAMs has emerged from the field of surface chemistry, which allows the implementation of many reaction schemes to obtain different surface functionalizations. To date, a wide range of synthetic routes that uses different precursor molecules to form dense monolayers have been described, and some examples are provided in the following sections. These reactions have been conducted on nonpatterned surfaces and on both micrometer and nanometer scales, in combination with alternative structuring approaches or by using techniques discussed earlier in the chapter. Nonetheless, each of the described reactions is suitable for implementation in a range of lithographic techniques. Due to the greater stability of silane-based monolayers compared to thiol monolayers on gold, attention will be focused at this point on the surface chemistry performed on SAMs on silicon-based substrates.

2.4.1
Molecular Overlayers – Functionalization – Precursors

The formation of SAMs on silicon surfaces can be carried out using either silane-based precursor molecules that self-assemble on oxidized silicon-based surfaces (such as glass or silicon [134]), or via activation by $SiCl_4$ and $HNEt_2$, followed by the addition of hydroxyl-functionalized molecules [369, 370]. Yet, a conceptually different approach for the formation of SAMs on silicon substrates is the hydrogenation of the silicon substrate and reaction with alkene-functionalized molecules [371].

Silane-based precursor molecules are mainly functionalized with a trichlorosilane, trimethoxysilane or triethoxysilane group that reacts with the surface to form a covalent network. In this respect, trichlorosilane is the most reactive precursor, and trimethoxysilane and triethoxysilane are less reactive. Although, a wide variety of functional moieties can be implemented into the silane-based monolayers by utilizing the terminal groups of the silane precursors, these functional groups must be compatible with certain criteria. For example, the introduced functional group should not interact with the surface, and/or should not react with the silane group in order to avoid the formation of multilayers, or destruction of the silane group. The spacer, which usually is an alkyl chain, also plays an important role, with longer alkyl chains resulting in the formation of more stable and densely packed SAMs [134]. Most patterning approaches concentrate on the use of commercially available silane-based molecules such as OTS, BTS, 1H, 1H, 2H, 2H-perfluorodecyltrichlorosilane (PFDTS), *N*-[3-(trimethoxysilyl)propyl]ethylenediamine (EDATMS), APTMS, APTES, *N*-(6-aminohexyl)-3-aminopropyltrimethoxysilane (AHAPS), MPTMS, and PEG silanes.

In general, OTS and PFDTS are used as passivation layers, due to the hydrophobic and chemically relative inert properties of the formed layers. BTS is applied mostly for surface reactions such as substitution reactions, or it may act as an initiator for polymerization reactions. Amino-terminated SAMs are used for the binding of negatively charged nano-objects as well as for surface reactions, such as Schiff base reactions or esterifications. MPTMS is used for the binding of nano-objects, while

PEG silanes have been extensively studied as bio-repellent materials that demonstrate their importance in biorelated systems and find applications in cell and protein micropatterning. These systems are also of significant importance in research related to the development of biosensors, lab-on-a-chip devices, tissue engineering, fundamental cell biology studies, drug screening, or medical diagnostics. Some examples of these precursor molecules in their fields of application are provided in the following subsections.

Typically, these materials are applied for purposes that include protein and cell-adhesion studies. Notably, OTS and PFDTS can each be self-assembled so as to obtain mixed monolayers that demonstrate varying protein-adsorption behaviors. For example, protein adsorption can be suppressed compared to the pure monolayer, with adsorption occurring preferentially on PFDTS [372]. When Hoffman *et al*. investigated the protein-repellent properties of mixed PEG silane-based monolayers and alkyl silane-based monolayers, the pure PEG silane monolayer was shown to be fully protein repellent, whereas on the mixed monolayer no protein absorption was observed above a PEG silane content of 90% [373]. Yap *et al*. created a micro pattern with topographical features for selective cell adhesion by modifying a silicon substrate with a PEG silane to create a nonfouling background. The PEG-modified surface was then patterned by photolithography, using a photoresist layer placed on top of the SAM, such that the free positions were subsequently functionalized with [3-(2-aminoethylamino)propyl]trimethoxysilane. When the polystyrene latex beads were later self-assembled on an amine-functionalized surface, HT-29 cancer cells were shown to adhere preferentially to the hydrophilic substrate areas, with the increasing distance between the adhesive sites promoting cell adhesion [374]. When gas-phase soft lithography was used to pattern APTMS and MPTMS, a PDMS could be applied to a master such that its shape became adapted. After curing, the PDMS mold was peeled off and placed in a reaction chamber with APTMS or MPTMS. The PDMS mold was then pressed onto the silicon substrate, and the ink was transferred by diffusion to react with the surface. The presence of amine groups was demonstrated by the adsorption of an oligonucleotide by the negatively charged phosphate groups [375], while the precursor molecules could be used to guide nano-objects onto certain positions. The fabrication of a gold nanoparticle array was achieved by first creating an OTS and an APTMS patterned surface; in this way, an OTS monolayer was self-assembled and patterned by AFM anodization (as described earlier). The oxidized areas were then used to self-assemble the APTMS, and this bifunctional template was subsequently used to self-assemble the gold nanoparticle array [285]. Guidance of the gold nanoparticle assembly into nanometer arrays could also be achieved with a hexadecene monolayer attached to a hydrogenated silicon surface. For this, the monolayer could be patterned by LAO and used to self-assemble APTMS that could, in turn, be used to guide the gold nanoparticles [376]. Moreover, APTMS can be used to self-assemble well-dispersed CNTs on the oxidized silicon surface [377]. Pang *et al*. demonstrated the improved patternability and adhesion of poly(3,4-ethylenedioxythiophene) (PEDOT) by the microcontact printing of OTS and subsequent filling of the nonfunctionalized areas with amine-functionalized SAMs. Following this, $FeCl_3$ could be selectively spincoated onto the amine-functionalized

areas, and PEDOT films were grown selectively on the amine-functionalized surface by vapor-phase polymerization. The amine-functionalized monolayer and $FeCl_3$ resulted in an improved adhesion of the PEDOT film [378].

SAMs are also frequently used to tune the surface properties, with the hydrophobic coatings playing an important role in the fluid–surface interactions of microfluids. Thus, Feng *et al.* investigated the influence of OTS on the fluid dynamics of chemically modified silicon dioxide and glass microchannels [379].

Whilst these few examples highlight the wide variety of applications able to benefit from surface modifications with commercially available precursor molecules, significant interest remains in widening the availability of the precursor molecules. Hence, two quite different strategies can be used to generate functionalized monolayers. The first strategy relates to the synthesis of new precursor molecules that provide tailor-made functional groups, while the second strategy involves the implementation of new functionalities and focuses on the chemical modification of conventional precursor layers.

Notably, three alternative approaches have been proposed that describe the route to synthesize trichlorosilane precursor molecules; this is illustrated schematically in Figure 2.15.

The first synthesis of a silane-based monolayer was described by Netzer *et al.*, who initially converted the hydroxyl group of a 10-undecenyl alcohol into a chloride group that was then activated by Mg; tetrachlorosilane was subsequently added to form the trichlorosilane group. An extension of the alkyl chain length can be achieved by adding oxirane or oxetane to introduce either two or three methylene groups to the Mg-activated molecule [380]. Other possible approaches to the synthesis of silane-based molecules include the synthesis of alkene-functionalized molecules, which can be converted by adding $HSiCl_3$ and H_2PtCl_6 as a catalyst, or by the use of bromine-functionalized molecules, which can be converted to silane molecules by adding Mg, followed by the addition of $SiCl_4$ (as in the above-described approach). Maoz *et al.* described the synthesis of *trans*-13-docosenyltrichlorosilane [381], while Wasserman reported the synthesis of methyl 11-(trichlorosilyl)decanoate via a platinum-catalyzed method and the synthesis of 16-hepatdecenyltrichlorosilane and 10-uncecenyltrichlorosilane via a magnesium-activated approach [382].

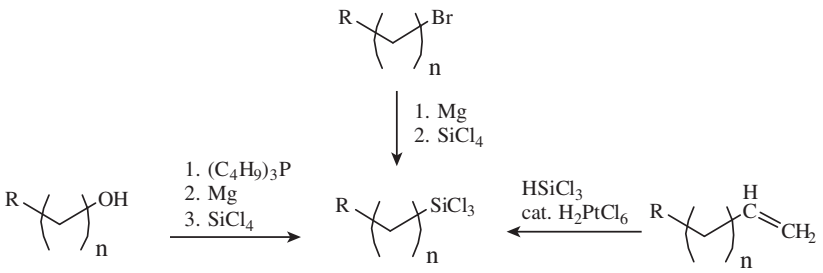


Figure 2.15 An overview of the synthesis routes to obtain trichlorosilane-based functional molecules.

Based on these general strategies, several functional precursor molecules have been synthesized, demonstrating the versatility of this approach. The synthesis of cyano-, bromo-, thiocyanato-, and thioaceto-terminated C_{16} trichlorosilanes was introduced by Balachander *et al.*, using the platinum catalyst-mediated process [383]. Additionally, the synthesis of iodo-, chloroacetate-, iodoacetate-, benzyl bromide-, and benzyl iodide-terminated C_{16} trichlorosilanes has been established using the same method [384]. Phthalocyanine molecules, which may have potential applications in display technology, as chemical sensors or as photoconducting devices [385], have been functionalized with an alkyltrichlorosilane group via the platinum catalyst pathway, and can be self-assembled on silicon. For this purpose, an 11-(3-thienyl) undecenyltrichlorosilane could be synthesized via a platinum catalyst method and then used for the self-assembly onto surfaces. These SAMs were proposed for use in thiophene polymerizations, to form conductive layers [386]. A maleiimido-terminated alkyl trichlorosilane molecule has also been synthesized using the platinum catalyst method. In this case, the maleimido group can be utilized for the covalent binding of nucleophilic heterocycles, alkylthiols or amines, thus making the attachment of a wide class of molecules to the surface possible [387]. The synthesis of $C_{10–12}$ alkyl chains terminated with a functional hydroxyl group and a PPh_2 group was reported for the self-assembly on silicon surfaces via activation of the surface by $SiCl_4$ and diethylamine, and the attachment of hydroxyalkylphosphine. Here, the PPh_2 end group could act as a ligand for Rh catalysts on the surface, or reaction of the Rh could be performed prior to the self-assembly process. The covalently attached Rh catalyst was tested for the hydrogenation of tolan [388]. When Zhang *et al.* demonstrated the formation of a PEG silane by the reaction of PEG with tetrachlorosilane, an effective depression of plasma protein adsorption and cell attachment was noted on these surfaces [389]. Subsequently, Sharma *et al.* described the development of ultrathin, uniform, stable *in-vivo*-like environments and conformal PEG films for silicon-based microdevices. For this, the surface-reactive PEG molecule was prepared by dissolving PEG in anhydrous toluene, followed by successive triethylamine addition. A tetrachlorosilane was then added so as to form a trichlorosilane group on the hydroxyl group. Different reactions, including the functionalization of PEG hydroxyl groups, as well as a reaction of tetrachlorosilane with several PEG chains up to cycle formation, might occur during this process. Nonetheless, the PEG silane was shown to form uniform films on silicon surfaces with a degree of roughness less than 1 nm by optimization of the self-assembly conditions [390]. Chi *et al.* synthesized an imidazolium chloride-functionalized triethoxy silane by the reaction of methylimidazole with a triethoxysilane-functionalized alkyl chloride. These authors investigated the use of such SAMs for controlling the wettability of silicon substrates by anion exchange, and showed that the water contact angle could be varied from 28 to 42° simply by exchanging the counterion, from choride to PF_6^- [391]. Each of these examples underlines the major impact of silane-based molecules on adding chemical functionality to SAMs, and the possibility of expanding the availability of tailor-made functional groups to bind and/or stabilize nano-objects, or for subsequent chemical modification schemes.

2.4.2 Surface Chemistry

Besides the (sometimes difficult to control) synthesis of functional silane precursor molecules – due to the high reactivity and water-sensitivity of the reactants – an alternative route has been proposed to acquire functional SAMs. This approach employs chemical reactions that are performed on SAMs that consist of silanes and which are known to form monolayers of reliable quality. The advantage of this technique is seen in the better quality of the starting monolayer, the possibility of avoiding undesired interactions between the silane and the functional group, and no need to optimize the self-assembly process for each individual precursor molecule. However, the subsequently performed surface reaction should be highly efficient in order to maximize the availability of the desired surface functionalities. Sagiv *et al*. reported the first chemical reaction on a covalently attached SAM, by demonstrating the conversion of a double bond to a hydroxyl function by treatment with B_2H_6 and successive reaction with H_2O_2/NaOH solution; the newly formed alcohol end groups were then used for the formation of multilayer systems [380]. Based on these initial studies, a substantial research activity on surface chemistry in the field of functional silane-based monolayers has subsequently emerged.

As a result, substitution, esterification, Schiff base reactions, the formation of thiourea and ureas, click chemistry, photochemistry, oxidation, and the growth of polymer brushes – among others – have each been introduced and are discussed with respect to their applications in surface chemistry in the following subsections. A schematic overview of these reaction schemes is provided in Figure 2.16.

2.4.2.1 Substitution

Substitution reactions utilize the displacement of one functional group by another, such as the displacement of bromide by a thiocyanate or an azide (Figure 2.16). The substitution is mainly divided into *nucleophilic substitution*, whereby an anion attacks the functional group, or an *electrophilic substitution*, where a cation replaces the functional group. For surface reactions, BTS monolayers are mainly used for nucleophilic substitution reactions. In particular, Balachander *et al*. self-assembled a BTS monolayer and demonstrated replacement of the bromine group by substitution with azide or thiocyanate [383]. Furthermore, bromine- or chlorine-terminated SAMs can be substituted with iodine, as demonstrated by Lee; the function can subsequently be replaced with decanethiol, *n*-decylamine, *p*-nitrothiolphenol, glutathione and lamini fragment peptides [384]. Additionally, Fryxell *et al*. demonstrated the substitution of a bromine SAM with molecules such as cysteine and amines [392]. Shuye *et al*. enlarged the available functional groups by conducting a nucleophilic substitution of a bromine-functionalized SAM with thioacetate, sulfonate, and nitrile [393], while Haensch *et al*. self-assembled BTS monolayers and showed the substitution of bromine with primary amines such as propargylamine and 5-(2,2:6,2″-terpyridin-4yloxy)pentylamine; here, the terpyridine moiety was further used for complexation with PEG_{70}–$RuCl_3$ [394]. Another possible means of introducing terpyridine moieties is via reaction with an acetylene-modified Fe(II) complex;

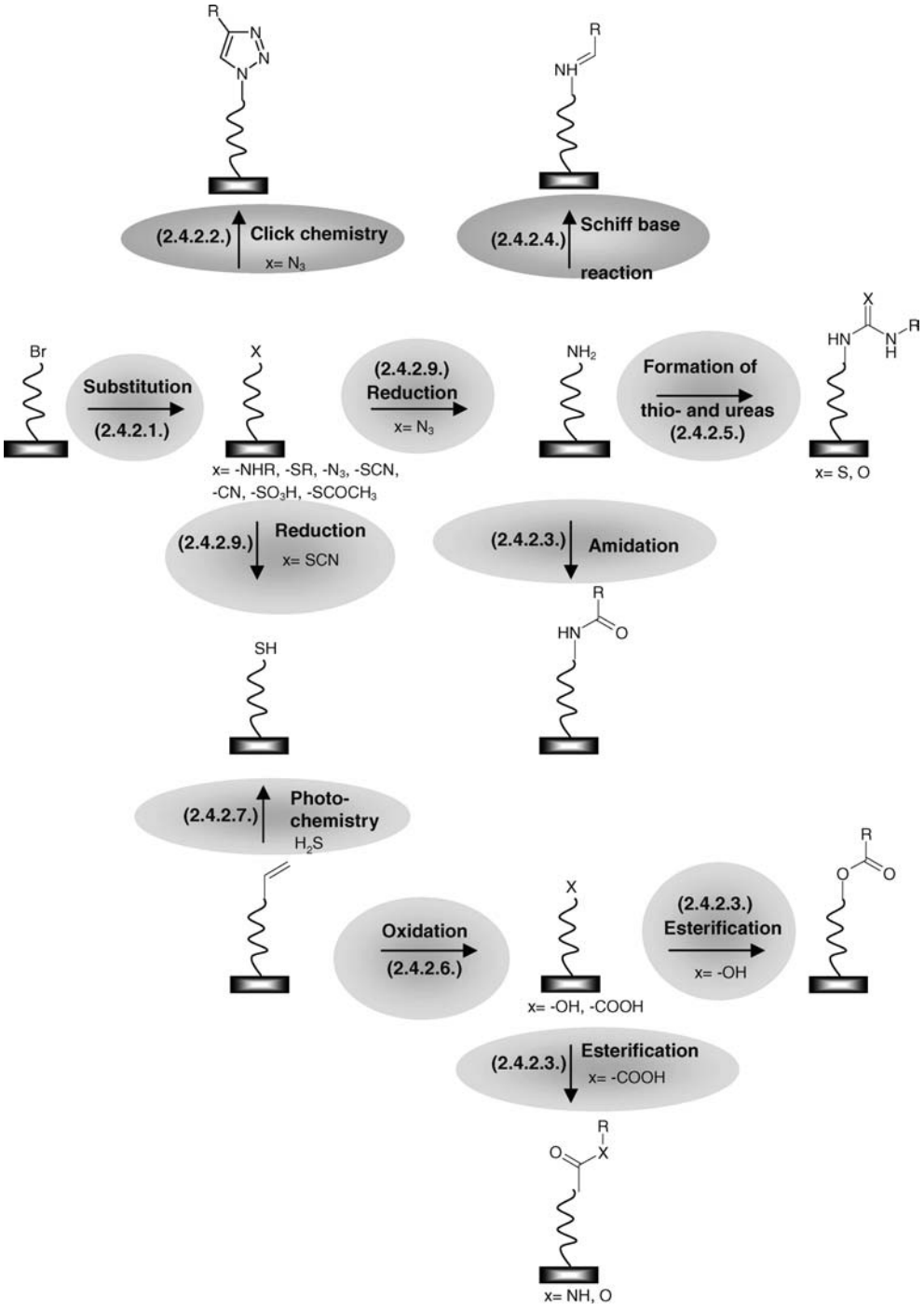


Figure 2.16 A schematic overview of possible modification schemes that can be implemented by surface reactions.

this can be initiated by treatment with HCl, with the free terpyridine moieties which can subsequently being closed with other metal ions such as Ir(III) or Zn(II) [395].

The nucleophilic substitution reaction scheme can be also used for structured surfaces. For example, Herzer *et al.* prepared chemical nanostructures by utilizing a high-resolution lithographic technique to form replaceable barrier nanostructures of triangular shape. For this purpose, an OTS monolayer was self-assembled onto a glass substrate patterned with the gold barrier structures. Then, after removal of the barrier structure, a BTS monolayer was self-assembled selectively onto the former positions of the barrier structures. The BTS monolayer was converted first, via substitution, into a thiocyanate monolayer, and subsequently to a thiol by reduction. Site-selective binding of gold nanoparticles could be demonstrated with high sensitivity on the thiol functional group [396]. Haensch *et al.* demonstrated the patterning of a BTS monolayer via a selective degradation of the BTS monolayer by LAO, with PFDTS being self-assembled in site-selective fashion as a passivation layer onto the oxidized areas. The BTS monolayer was subsequently substituted with azide and successively converted into an amine group by reduction, which could site-selectively bind silicon nanoparticles [397] (also C. Haensch *et al.*, unpublished results).

The substitution reaction represents a powerful tool for creating a large variety of versatile functionalized surfaces, and thus introducing a wide diversity of surface properties, such as charges or biofunctionalities. These properties can be utilized in a variety of applications, such as biosensors, for the growth or deposition of inorganic materials, in microfluidics, for the microengineering of smart surfaces for bioseparation or data storage, as sensors, or in the microfabrication of controlled-release devices.

2.4.2.2 **Click Chemistry**

The now widely used click chemistry was introduced in 2001 by Sharpless [329]. The main characteristics of a click reaction include its modularity, the wide scope, the high yield, and the lack of formation of any byproducts (or, at least, of byproducts that can be removed using nonchromatographic methods, such as crystallization or distillation). Furthermore, the reaction must be stereospecific, the reaction conditions mild, and the starting materials and reactants readily available. Furthermore, no solvents or easily removable solvent should form any part of the reaction, and the products should show a good stability under physiological conditions. The most commonly used click reaction on surfaces is the 1,3 dipolar cycloaddition of azide-functionalized surfaces with acetylene-functionalized molecules (see Figure 2.16). The first example of click chemistry to be conducted on a silica surface was described by Lummerstorfer *et al.* in 2004 [398], when BTS was self-assembled on a silicon wafer and the bromine functions were subsequently converted into azide groups via a substitution reaction. The azide functions were later used for the Huisgen 1,3-dipolar cycloaddition reaction [330] with three differently substituted acetylenes, for example, $R-C\equiv C-R'$ ($R,R' = C_6H_{13}$, H; $COOCH_3$, H; $COOC_2H_5$, $COOC_2H_5$). In this case, only ester-functionalized acetylenes showed a quantitative conversion, whilst for hexyl-substituted acetylene no reaction was observed (this was explained by the influence of electron-withdrawing ester groups). Rohde *et al.* described the activation

of a hydrogenated surface by chlorination to bind sodium acetylene. Here, the acetylene functionality was used to click the electroactive benzoquinone which, when covalently attached, was reduced to a primary amine group by the application of a voltage that was used to covalently bind a ferrocene complex via an amide linkage [399]. The covalently attached electro-active ferrocene molecules might find potential applications in charge–storage molecular devices. Ciampi *et al*. described the covalent immobilization of commercially available diacetylene compounds on hydrogenated silicon surfaces via a hydrosilylation procedure; after which the alkyne end group was used to click various azide compounds [400]. The clicking of molecules such as polymers [401] and dyes [402] has also been demonstrated.

Click reactions have also been demonstrated on structured surfaces. For example, an *n*-octyldimethylchlorosilane monolayer was gradually modified following exposure to UV light to generate ozone to form acid groups. The acid groups were then used to bind acetylene-terminated molecules, which were further used to click peptides via the 1,3-dipolar cycloaddition, via the formation of a triazole ring [403]. Click chemistry based on microcontact printing has also been demonstrated by Rozkiewicz *et al*. to obtain structured surfaces. These authors self-assembled BTS and substituted the bromine with azide, after which a PDMS stamp was inked with the acetylene-terminated molecules and pressed onto the azide-terminated surface [404]. Ravoo *et al*. demonstrated, moreover, the preparation of carbohydrate microarrays by using microcontact click chemistry. For this purpose, BTS was self-assembled on glass or silicon surfaces, and subsequently converted to an azide moiety by substitution. Previously, alkyne-functionalized carbohydrates have been synthesized and clicked onto the azide-functionalized surface by pressing the PDMS stamp, inked with the carbohydrates, $CuSO_4$ and ascorbic acid, onto the substrate. Lines with dimensions of down to 5 μm could be created in this way [405]. Click chemistry, directed using scanning electrochemical microscopy, was first introduced for the self-assembly of a BTS monolayer onto a glass slide, with subsequent conversion into an azide group by a substitution reaction. In this way, a gold microelectrode could be used to transfer the acetylene-functionalized fluorescent dye and to create Cu(I) ions locally between the tip and the substrate, to catalyze the click reaction. By using this method, features of about 500 μm were created [406]. Oxidative nanolithography has been also combined with the click chemistry approach, when an OTS monolayer was oxidized with a biased AFM tip and BTS was site-selectively self-assembled onto the activated region. The bromine end group was replaced with azide and used to click propargyl alcohol; in this way, surface reactions on feature sizes down to 50 nm could be performed [448].

These few examples of click reactions highlight the versatility of the process, with both biomolecules and electroactive molecules being introduced onto the surfaces to demonstrate potential future applications in areas such as electronics, sensors, or glycomics.

2.4.2.3 **Esterification/Amidation**

Another possible means for binding functional molecules onto a surface is that of esterification or amidation, which utilizes the reaction of an acid or ester group with a

hydroxyl or with an amine group, as depicted in Figure 2.16. In this way, either carboxylic acid- or ester-functionalized molecules can be introduced onto the surface; alternatively, amine- or hydroxyl-functionalized molecules can be used. Esterification and amidation reactions are important for biological applications, as they may be used to attach biomolecules such as DNA or biotin. These types of functionalized patterns are also important for the placement of cells, and to study cellular interactions.

Fryxell *et al.* demonstrated the formation of amide bonds, by the reaction of a trifluoroethylester SAM with primary amines [392], thus providing a possible means of covalently attaching amine molecules to the surface. Maoz *et al.* described the microwave-induced reaction of an acid-salt-terminated SAM and an amine-functionalized alkyl chain, which permitted amine-functionalized molecules to be bound onto the surface [407]. Flink *et al.* demonstrated the chemical functionalization of APTES with acid chloride-, thionylchloride- and carboxylic acid-functionalized molecules [408], thus opening the possibility to covalently attach carboxylic acid-functionalized molecules.

These reactions were also used to create functionalized patterns, and some representative processes are highlighted here. For example, Zhang *et al.* patterned a self-assembled PFDTS monolayer by using electron beam lithography (EBL), with APTMS subsequently being self-assembled on the irradiated spots and further functionalized with biotin by reaction of the amine group with an activated acid group attached to the biotin so as to form an amide bond. By using this method, nanostructures down to 250 nm could be created [409]. Normally, the direct attachment of amine-functionalized molecules to carboxylic acid-functionalized surfaces leads to salt formation, rather than to the creation of the desired amide bond. Hence, the attachment of an amine to carboxylic acid-functionalized surfaces must be carried out by activation with NHS (Figure 2.17).

In this reaction, NHS reacts with the carboxylic acid groups under the formation of a succinate ester, which presents a good leaving group and can subsequently be reacted with the desired amine molecules, forming of an amide bond. Previously, this method has been used by Fabre *et al.*, who self-assembled an ethyl undecylenate and a 1-decene monolayer on a hydrogenated silicon surface. The ethyl undecylenate SAM

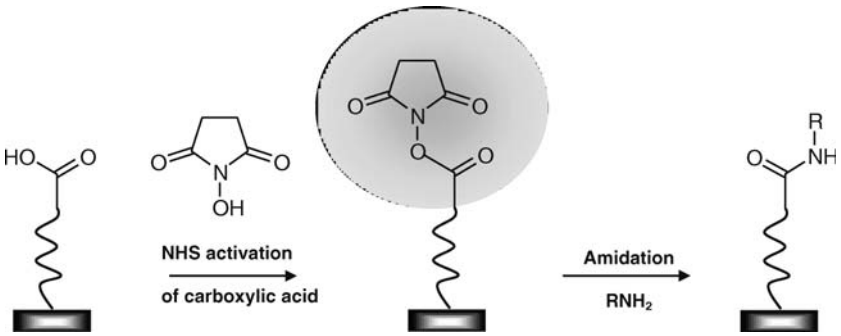


Figure 2.17 NHS activation of carboxylic acid for the binding of amine-functionalized molecules.

was activated with NHS by immersing the substrate into a solution of 1-(3-dimethylaminopropyl-3-ethylcarbodiimide) and NHS. A 2- aminoethylferrocenylmethylether was then covalently attached to the NHS-activated ethyl undecylate SAM, with formation of the amide bond [410]. The intact amine was used to bind DNA by first reacting with succinimidyl-4-[maleimidophenyl]butyrate to form an ester bond, after which the thiol-functionalized DNA was covalently attached [411]. An improved method for the reduction of nonspecific protein adsorption was introduced via the self-assembly of 7-octenyltrimethoxysilane and successive EBL, during which the irradiated areas were used for the self-assembly of APTES and the vinyl groups of the 7-octenyltrimethoxysilane were converted to hydroxyl groups. The binding of proteins onto amine groups was suitable to create patterns with a resolution down to 250 nm [412]. Crivillers *et al*. demonstrated the amidation of EDATMS by reaction with a carboxylic acid-terminated radical; in this case, the radical functionalized surface could be reversibly oxidized and showed different absorption spectra in the oxidized and reduced states. This approach proved to be highly attractive, as it provided a chemical redox-switchable surface that led to changes in the optical and magnetic responses of the system. It was also proposed that patterns of these radical molecules could be created by micro-contact printing, and this was subsequently demonstrated by inking the PDMS stamp with the radicals and pressing the stamp onto an amine-functionalized surface [413]. Duan *et al*. described the bifunctional chemical patterning of APTMS and PFDTS by micro-contact printing, whereby the APTMS was functionalized with a fluorescent dye by esterification, thus opening the possibility for further modification processes [414].

2.4.2.4 **Schiff Base Reactions**

Amine-terminated surfaces can be functionalized with aldehyde molecules via Schiff base reactions to form an imine bond, or vice versa (Figure 2.16). La *et al*. demonstrated the nanopatterning of SAMs via a selective chemical transformation induced by soft X-ray irradiation. For this purpose, (3-aminopropyl)diethoxymethylsilane was self-assembled on a silicon surface and functionalized with 4-nitrobenzaldehyde or 4-nitrocinnamaldehyde, under the formation of an imine bond. A nitrosubstituted phenyl-imine SAM was then converted into a secondary amine monolayer by selective X-ray irradiation, and the nonirradiated areas were later hydrolyzed to amines. Moreover, the amines could be converted with a Cy3-tagged oligonucleotide, with selective conversion of the amine by the Cy3-tagged oligonucleotide being confirmed by fluorescence imaging [415]. The Cy3-tagged oligonucleotides may also be attached covalently to the amine by using an esterification reaction, as previously introduced and described by Chen *et al*. [411]. Rozkiewicz also demonstrated the Schiff base reaction of an amine-functionalized monolayer, such that the EDATMS was self-assembled on silicon surfaces and the amine functionality was converted to an imine bond by reaction with different aldehydes. The imine bond can also be formed by microcontact printing of the aldehyde onto an amine-functionalized surface, as demonstrated by microcontact printing of the fluorescent dye Lucifer Yellow on amine-functionalized glass [416]. Schiff base reactions have also been shown capable of successfully immobilizing cytophilic proteins by microcontact

printing, utilizing a similar approach to that described above. The amine monolayers were first self-assembled and reacted with terephthalaldehyde; the proteins were then contact-printed onto these aldehydes with feature sizes down to 100 μm, and the remaining areas were filled with amine-functionalized PEG [417]. Moreover, a mixed aldehyde/alkyl organic monolayer that had been self-assembled on a hydrogenated silicon surface was selectively functionalized on the aldehyde group by amine-functionalized biomolecules, such as the lysine groups of proteins [418].

To summarize, Schiff base reactions have been applied in particular for the binding of biomolecules, providing the reaction scheme with possible future applications in bioassays and other cell-related studies.

2.4.2.5 **Formation of Thioureas and Ureas**

The formation of thioureas and ureas has also been used to introduce functional moieties into SAMs (see Figure 2.16). Flink *et al*. demonstrated the chemical functionalization of APTES with isothiocyanate- and isocyanate-functionalized molecules to form thiourea or urea groups on the SAM [408].

This chemical reaction can be shown as being compatible with patterning methods, with a self-assembled EDATMS monolayer being patterned using nanoimprint lithography, for example. In this case, the amine function was used in the reaction with 1,4-diphenylene diisothiocyanate, and then successively reacted with cyclodextrin-functionalized gold or silicon nanoparticles. The same template was further utilized to monitor host–guest interactions with suitable molecules, as well as for two types of layer-by-layer assembly [419]. Maury *et al*. used nanoimprint lithography to create nanopatterns of His-tagged proteins, and thus prepared EDATMS–PEG or –PFDTS patterns. Using these structures, protein adsorption experiments were conducted via electrostatic and supramolecular interactions. The amine was also reacted with a 1,4-phenylene diisothiocyanate, and *N*-nitrilotriacetic acid subsequently attached; a protein was then attached to the nitrilotriacetic acid after treatment with a Ni(II) solution. Such protein interactions have been shown to be reversible by exchanging the Ni(II) with a strong competing ligand [420].

The formation of thioureas and ureas represents a highly compatible reaction scheme that can be integrated into several different patterning approaches, notably in the efficient binding of biomolecules that could be employed for biological applications.

2.4.2.6 **Oxidation**

Oxidation has been used to introduce hydroxyl- or acid-functionalized moieties onto SAMs, which can in turn be used for the esterification reactions described above to introduce other molecules (as shown in Figure 2.16). Netzer *et al*. described the oxidation of alkene end-functionalized trichlorosilane SAMs with different alkyl chain lengths to hydroxyl groups by treatment with B_2H_6 and subsequent treatment with a mixture of H_2O_2 and NaOH; hence, hydroxyl-functionalized SAMs were used for the formation of multilayers [380, 421]. Maoz *et al*. also demonstrated the oxidation of unsaturated alkyl trichlorosilane SAMs by crown ether-solubilized $KMnO_4$ to cleave the alkene bond, under the formation of a covalently attached

alkyl silane-acid salt and an alkyl acid salt, that is removed from the surface. The length of the alkyl chain was seen to depend on the length and position of the starting unsaturated trichlorosilane [422]. Wasserman *et al*. demonstrated the oxidation of alkene or methyl functional groups of SAMs to carboxylic acid groups by $KMnO_4$, $NaIO_4$, and K_2CO_3 [382], while Shyue *et al*. reported the oxidation of nitrile-functionalized SAMs to carboxylic acid by stirring the substrates in a solution of sodium bicarbonate [393].

Oxidation can also be carried out on structured surfaces, and some examples of this are provided at this point. Maoz *et al*. described the oxidation of a methyl-terminated alkyl chain with an AFM tip (for details of this reaction, see Chapter 2.3.3.2), and also demonstrated the possibility of forming multilayer systems on the nanometer scale [311]. Miyaki *et al*. reported the details of the high-resolution EBL of octenyl-trimethoxysilane, where the vinyl groups were converted into hydroxyl groups by treatment with $BH_3 \times THF$ and a mixture of H_2O_2 and NaOH, following the previously introduced approach [380, 421]. Feature sizes down to 18 nm could be achieved using this method [423].

Oxidation reactions allow, in particular, the formation of multilayer systems, which in turn might allow the introduction of both ready-made and synthesized silane-based molecules, in addition to their combination with esterification reactions. Such a scheme would permit a wide variety of possible applications, depending on the nature of the introduced molecules as well as the formation of 3-D structures.

2.4.2.7 **Photochemistry**

Photochemical reactions represent an additional powerful means of functionalizing SAMs. Previously, Frydman described the photochemical conversion of a NTS SAM to a thiol-functionalized monolayer by treatment with H_2S and UV-irradiation at 254 nm (see Figure 2.16) [424], with the formed disulfide bonds being cleaved by treatment with either $NaBH_4$ [339] or $BH_3 \times THF$ [327]. A similar photochemical conversion into thiol-functionalized SAMs was also implemented into surface-structuring techniques, whereby the electro-oxidation of an OTS monolayer was applied and the NTS self-assembled onto locally formed acid groups. The double bond was subsequently utilized for the formation of thiol-functionalized groups, where both Ag and CdSe nanoparticles could be assembled [312] to form macrosopic electrodes on the surfaces. Such a modification scheme would be compatible with constructive electro-oxidation lithography, and might represent a technique for the creation of conductive metallic wires, with nanometer resolution [339].

Photochemistry can also be applied to photoisomerization reactions, with azo-silanes being synthesized and self-assembled on silicon oxide substrates. Following irradiation with UV light (360 nm), a *trans–cis* photoisomerization was observed, whereas UV-light irradiation at 450 nm triggered a *cis-trans* molecular isomerization. These changes in molecular photoisomerization were investigated using surface plasmon resonance spectroscopy (SPRS), a technique that allows the detection of very small changes in the thickness of these ultrathin layers [425]. These reversible transitions within the monolayers represent an example of the light-triggered reversible switching of the layer thickness.

Hozumi *et al*. described the photochemical conversion of an aldehyde-functionalized SAM to carboxylic acid-functionalized SAMs by irradiation using 172 nm vacuum UV light [426].

Brandow *et al*. reported on the self-assembly of *p*-chlorophenyltrichlorosilane which could be patterned through a photomask by irradiation with UV light (193 nm). In this way, the chlorine groups on the irradiated region were converted into aldehyde groups, which in turn were further converted to amines by treatment with NH_4OAc and $NaBH_3CN$. Subsequently, the amine moieties were used for the self-assembly of a Pd(II) catalyst, which catalyzed the electroless deposition of nickel [427]. Hong *et al*. demonstrated the photoreactivity of *n*-octadecyltrimethoxysilane by irradiation (through a photomask) with vacuum UV light of 172 nm, and observed the formation of carboxylic acid groups whilst the alkyl chain was degraded. The carboxylic acid moieties were further used for the binding of fluoro-functionalized silane molecules or of (*p*-chloromethyl)phenyltrimethoxysilane [428]. Chen *et al*. demonstrated the patterning by UV-light irradiation at 193 nm through a photomask of (aminoethylaminoethyl)phenylsiloxane (PEDA) SAMs. In this case, the benzylic C—N bond was cleaved on the irradiated regions, followed by the formation of an aldehyde. The intact amine was used to bind DNA by first reacting it with succinimidyl-4-[maleimidophenyl]butyrate to form an ester bond, after which the thiol-functionalized DNA became attached covalently [411].

Photochemistry represents a powerful tool for the creation of functional patterns in a one-step procedure, notably because of its similarity to the patterning of resist materials, and hence its compatibility with standard photolithography. In particular, the functional groups can be used for further chemical reactions, not only to introduce functional moieties but also for the preparation of light-induced switches.

2.4.2.8 **Polymer Brushes**

Polymers can be used to "tune" the properties of a surface, a procedure that is highly desirable in many areas of research, and especially in biotechnology and advanced microelectronics. Although, the primary method for attaching polymeric brushes to a surface is via the adsorption of block copolymers [429–431], it is the noncovalent nature of this procedure that becomes its major weakness. Yet, such limitations can be relatively easily overcome by providing a *covalent* attachment of the polymers onto the surface, and for this two different strategies are favored: (i) a "grafting- from" approach; and (ii) a "grafting-to" approach. Whilst both methods permit the surfaces to be functionalized with polymer chains, the "grafting-from" method involves a polymer being grown from the surface, whereas the "grafting-to" method involves the attachment of pre-synthesized polymeric chains.

In 1996, Chang *et al*. were the first to demonstrate different ways of constructing SAMs of poly(γ-benzyl-L-glutamate) on silicon substrates, using both "grafting-from" and "grafting-to" methods. In the first of these methods, the preformed polymer is "grafted-to" the surface via a chloroformat group, whereas in the second method poly(γ-benzyl-L-glutamate) is first functionalized with a triethoxysilane head group and then attached to the silicon surface, using the triethoxysilane groups. As an

alternative, the polymerization could be initiated by attaching APTES to the surface. It was concluded by these authors that "grafting-to" was more promising than "grafting-from," mainly because of the low degree of polymerization observed for the surface-initiated polymerization [432]. Based on these results, an extensive search was undertaken in the field of surface-initiated polymerizations, and a schematic overview of the various polymerization methods used on different surfaces is shown in Figure 2.18. The polymerization techniques are explained in detail, and representative examples discussed, in the following subsections.

"**Grafting-From**" **Approach** In the past, a number of different polymerization techniques have been used to produce the polymeric building blocks needed to carry out a "grafting-from" approach. The most notable of these were controlled radical polymerization techniques, including atom transfer radical polymerization (ATRP), reversible addition-fragmentation chain transfer (RAFT) polymerization and nitroxide-mediated polymerization (NMP). All three methods have been shown to provide well-defined polymer materials for surface-initiated polymerizations. Husseman *et al*. were the first to discuss the possible synthesis of a silane-based initiator for NMP and ATRP, whereby monosilane or trichlorosilane groups could be introduced to allow attachment of the initiators to the surface. Subsequently, the successful growth of different homo-, random, and block copolymers using these initiators was demonstrated [433]. Matyjaszewski *et al*. also described the ATRP of styrene and methyl acrylate from an initiator that had been attached covalently to the silicon surface. The initiator was first synthesized via an esterification of 10-undecen-1-ol and 2-bromoisobutyl bromide, respectively, and this reaction was followed by hydrosilylation, which led to a highly reactive trichlorosilane group that could be used to form a covalent bond to the substrate. The living character of the polymerization was demonstrated by a linear increase in layer thickness with increasing polymerization time; and the possibility of growing block copolymers was also demonstrated [146]. Rowe-Konopacki *et al*. demonstrated the surface-initiated RAFT polymerization of diblock copolymers, using the same trichlorosilane molecules as described by Matyjaszewski, and converted them with a dithiobenzoyl disulfide to obtain a RAFT initiator on the surface [147].

Each of these methods has also been shown as suitable for the generation of patterned polymer brushes. For example, Piech *et al*. described the surface-initiated ATRP polymerization of spirobenzopyran-*co*-methylmethacrylate from quartz slides, onto the surfaces of which an ATRP initiator was self-assembled and, subsequently, ATRP polymerization performed. The polymer film produced was selectively irradiated by a focused 366 or 780 nm pulsed laser beam, so as to pattern the polymer film by inducing a ring-opening isomerization of the spirobenzopyran to a zwitterionic merocyanine isomer. This reaction was demonstrated to be reversible by, for example, heating or irradiation with light at 585 nm, with a color change from light yellow to purple/red on the irradiated areas indicating successful conversion. Following treatment of the spirobenzopyran-*co*-methylmethacrylate copolymer, the modified colloids self-assembled preferentially onto the irradiated regions [434]. Becer *et al*. carried out the patterning of an OTS monolayer on silicon by using

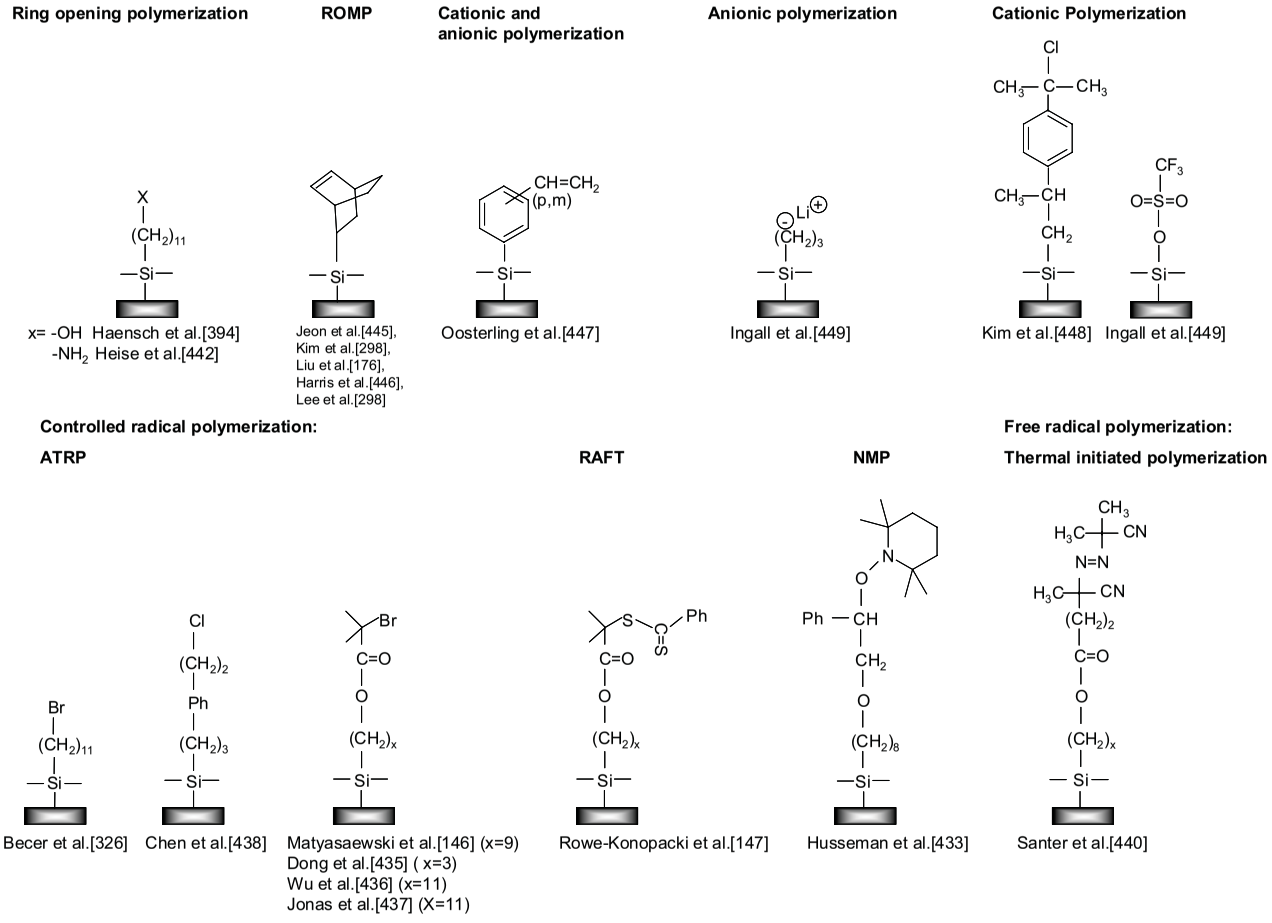


Figure 2.18 A schematic representation of surface-initiated polymerizations.

constructive electro-oxidative nanolithography, starting from a BTS monolayer which was self-assembled on the surface and utilized for the ATRP polymerization not only of styrene but also of block-copolymers. Polymer brushes with a line width of 70 nm could be obtained in this way [326, 397]. Later, Dong *et al.* showed that PEG/polyacrylic acid patterns obtained by photolithography could be used to obtain structured polymer brush surfaces. As an example, PEG trichlorosilane was self-assembled on a silicon wafer and coated with a photoresist that had been patterned through a photo mask; the PEG was later removed from the non-covered areas by O_2 plasma. An ATRP initiator with a surface-reactive chlorosilane group was then self-assembled onto the now free areas, and the polymerization of acrylic acid was carried out. In an attempt to attach proteins to the polyacrylic acid, the acid groups were activated by *N*-(3-dimethylaminopropyl)*N*′-ethylcarbodiimide, NHS, and 4-morpholine-ethanesulfonic acid, after which the amine-terminated proteins were attached covalently to the polymeric chains [435]. Wu *et al.* prepared a mixed OTS/polyacrylic acid pattern in which the base OTS monolayer was formed using a vapor diffusion technique, while the unmasked regions were used for the subsequent self-assembly of the ATRP initiator that was used to perform the polymerization of the acrylic acid [436]. Alternatively, patterning using either nanoimprint lithography or EBL was used to create nanometer-sized patterns of an ARTP initiator and PEG. After having first deposited a 100 nm-thick poly(methylmethacrylate) (PMMA) film on a silicon wafer, and creating a nanopattern, the ATRP initiator was self-assembled onto the exposed areas; following removal of the PMMA layer, the PEG monochlorosilane was self-assembled and the 2-(2-methoxyethoxy)ethyl methacrylate finally polymerized. In this way, polymer brush features with a lateral resolution down to 35 nm could be created [437]. Chen *et al.* described the formation of various polymer brush patterns prepared by the treatment of the hydroxylated silica surface, utilizing hexamethyldisilazane self-assembly and the subsequent activation of the $Si–CH_3$ groups by O_2-plasma and/or EBL. The locally formed, higher reactive Si–O species could be used to bind the ATRP initiator (4-chloromethyl)phenyltrichlorosilane. These surface-bonded initiators were used to polymerize the methylacrylate, such that features down to 5 μm in size could be obtained [438]. Brinks *et al.* also highlighted the possible use of NMP for surface structuring; in this case, the NMP-polymerized brushes were used on patterns formed by LB structuring, after which LB lithography was carried out utilizing L-α-dipalmitoyl-phosphatidylcholine (DPPC) and the triethoxysilane-functionalized NMP initiator; this resulted in a surface that was patterned with uniform DPPC lines. Ultimately, when the initiator was attached covalently and the DPPC removed from the substrate, monomers such as styrene and acrylate became polymerized [439].

Another possible method of implementing surface-initiated polymerization is that of free radical polymerization [440] (which is also compatible with several structuring techniques), or an electrochemical polymerization approach that permits the polymerization of monomers such as thiophene, which has an important role in electronic devices. Inaoka *et al.* demonstrated the electrochemical polymerization of different thiophene derivatives from a self-assembled oligothiophene-substituted alkylsilane [441].

Ring-opening polymerization (ROP) has also been used as a "grafting-from" technique. For example, Heise *et al.* prepared a mixed monolayer of BTS and 1-trichlorosilylundecane, whereby the bromine functional group was converted via an azide substitution and reduction to an amine functional group. The amine group was then used for the polymerization of benzyl γ-benzyl-L-glutamate *N*-carboxyanhydride [442]. Choi *et al.* demonstrated the surface-initiated ROP of ε-caprolactone from a *N*-(2-aminoethyl)-3-aminopropyltrimethoxy SAM [443], while Yoon *et al.* reported the ROP of poly(*p*-dioxanone). The same authors self-assembled (*N*-triethoxysilylpropyl)-*O*-poly(ethylene oxide)urethane on a silicon surface, and utilized the hydroxyl functional group for the $Sn(Oct)_2$-catalyzed ROP [444]. The possibility of introducing this polymerization approach to patterned substrates was shown by Haensch *et al.*, when the electro-oxidation of an OTS monolayer, the self-assembly of a BTS, and subsequent substitution to an azide was used to link propargyl alcohol to the surface structures, using the 1,3-dipolar cycloaddition. The propargyl alcohol was subsequently used to perform the ROP of L-lactide.

The ROMP technique was also used on silicon and patterned surfaces, with Jeon *et al.* demonstrating the patterning of OTS via micro-contact printing and a ROMP catalyst. In this case, the OTS was transferred via a PDMS stamp onto the substrate, and the non-covered areas were later filled with norbonyl trichlorosilane. The latter was activated with a ruthenium catalyst to surface-initiate the polymerization of derivatives of norbornene; use of this method led to smallest feature sizes of 2 μm being obtained [445]. Liu *et al.* described a combination of DPN and ROMP, and in particular demonstrated the successful transfer of 5-(bicycloheptenyl)trichlorosilane onto a Si substrate via an AFM tip, a subsequent activation of the structures with the first-generation Grubbs catalyst, and the subsequent polymerization of norbormene derivatives [176]. The photolithographic patterning of a ROMP catalyst was introduced as another method to obtain structured linear and crosslinked polymer brushes, and for this the inactivation of an Ru catalyst by exposure to UV-light was utilized. Consequently, the silicon wafer was functionalized with the initiator, which was in turn activated in a second step by a Ru catalyst. The substrates were illuminated through a mask that had been placed onto the substrate. Finally, the patterned catalyst layer was immersed in a solution containing the norbomyl monomer, and the polymer observed to grow only on the nonilluminated areas of the substrate. Features with a resolution of 4.3 μm could be created in this way [446]. Lee *et al.* used AFM anodization lithography to obtain nanometer-sized features of polymer brushes, with octadecylmethyldiethoxysilane being used as a resist layer. With a biased AFM tip the monolayer was locally removed, and SiO_2 was grown. Onto these features the 5-(bicycloheptenyl)triethoxysilane was self-assembled and, after activation, the polymers were grown. By using this method, feature sizes down to 75 nm could be prepared [298].

Other possible ways of performing surface-initiated polymerizations are anionic and cationic in nature. Oosterling *et al.* demonstrated the possibility of anionic polymerization from silicon surfaces by self-assembling *p*(*m*)-vinylbenzyltrichlorosilane on the surface, and then polymerizing monomers such as MMA [447]. Kim *et al.* prepared chlorosilyl-functionalized initiators and carried out a surface-initiated

living cationic polymerization of isobutylene [448]. Both, cationic and anionic, surface-initiated polymerization techniques have been shown suitable for the patterning, for example, by Ingall *et al*. These authors carried out the patterning of a phenylsilane monolayer by irradiation through a mask with 193 nm UV light. The nonirradiated areas were first functionalized by treating the monolayer with trifluoric acid to replace the phenyl ring, after which the triflat was substituted by nucleophils or monomers – that is, $-C\equiv CH$, $-OCH_2CF_3$, $-O(CH_2)_6NH_2$. The triflat function could be used for the cationic polymerization of MMA to form polymer films of PMMA or poly(propylene oxide) (PPO). The anionic polymerization of acrylonitrile has been performed from a bromine-functionalized silane monolayer, which could be activated using lithium-*tert*-butylphenyl [449].

In summary, almost all standard polymerization techniques have been applied to the fabrication of polymer brushes via the "grafting-from" method, and these examples highlight the versatility of this approach, whereby uniform brushes with various properties can be obtained. Nevertheless, characterization of the polymer brushes – in particular on the nanometer scale – is difficult, as the growth kinetics might be influenced by the surface and/or the dense organization of the individual polymer chains. The "grafting-to" method (see below) has the advantage that the polymeric material can be characterized prior to the polymers being attached to the substrate. Moreover, the polymer systems can be synthesized under optimal conditions, while conventional characterization and purification tools can be used to prepare well-defined polymer systems that subsequently will be linked to the surface.

"Grafting-To" Approach Among the "grafting-to" approaches used to covalently bind polymers onto silicon surfaces, the most obvious has been a functionalization of the polymer chain with surface-active silanes. Park *et al*. demonstrated the "grafting-to" of a rod–coil diblock copolymer to a silicon surface via immersion, casting, or contact printing. The rod–coil diblock copolymer consisted of a polystyrene part and a 3-(triethoxysilyl)-propylisocyanate, where the triethoxysilyl group could be attached covalently to the silicon surface. In this way, micropatterns of the polymer with 7.5 μm line width could be fabricated [450]. In particular, the grafting of PEG onto surfaces was of special interest, on the basis of its repellent properties towards proteins and cells. A PEG alkyltrichlorosilane pattern was fabricated by using a five-step procedure where first, a PMMA film was structured by EBL to selectively remove the PMMA from the silicon. The alkyltrichlorosilane was then self-assembled from the gas phase on the irradiated areas, the PMMA mask was removed, and the PEG trimethoxysilane was grafted onto the remaining silicon surface. These substrates were utilized so as to selectively adsorb collagen onto the hydrophobic alkyl tracks, which had a line width of between 30 and 90 nm [451]. Alternatively, a PEG silane could be self-assembled and spin-coated. As an example, Brough *et al*. used an eight-arm amine-terminated PEG to crosslink PEG by EBL in solution, after which biotin was immobilized on the amine-terminated PEG spots by the formation of an amide bond. The biotin was then used successively to immobilize streptavidin, and this was suitable to initiate the polymerization of actin [452]. Gaubert *et al*. demonstrated the

preparation of biologically active, large-scale nanopatterns by utilizing nanoimprint lithography as a patterning technique, such that a gold silicon pattern was obtained. In particular, the self-assembly of a commercially available PEG silane onto the silicon oxide regions was used, and passivation of the noncovered gold areas with hexadecanthiol was subsequently performed. Cell-adhesion experiments demonstrated PEG's cell-adhesive/cell-repellent properties, with patterns down to 60 nm being created on a polyethylene background [453]. Dekeyser *et al.* investigated the "grafting-to" of PEG-trimethoxysilanes and trichlorosilanes with different chain lengths under different conditions, in order to develop procedures for the preparation of nanostructured surfaces for biomaterial applications using EBL. For this, although the layer thickness was about 1–2 nm, no significant difference was identified for the use of hexane or toluene as solvent; however, the grafted silane was not stable under standard incubation conditions (37 °C, 24 h, phosphate-buffered saline). As a consequence, the grafted PEG could not be used as a repellent polymer for proteins, though this problem was not observed at room temperature [454]. Whilst the bulk of the investigations was based on the self-assembly of silane-based PEG, the reaction of poly(ethyleneglycol methylether) with a hydrogenated surface has also been reported [455]. An additional method that can be used for the "grafting-to" approach is that of click chemistry (see above). As an example, Ostaci *et al.* demonstrated the self-assembly of ethynylenedimethylchlorosilane onto silicon substrates fabricated by the 1,3-dipolar cycloaddition of different polymers such as PEG-N_3 or PMMA-N_3 [401]. LeMieux and coworkers described the "grafting-to" of carboxylic-terminated polystyrene and poly(butyl acrylate) on epoxysilane SAMs on silicon, where the total layer thickness was 1–3 nm. However, by varying the concentration of each polymer and the molar mass, very finely defined structured surfaces could be obtained with approximately 10 nm phase domains and less than 0.5 nm roughness. Due to the immiscibility of the two polymers, switching of the surface nanomechanical properties could be observed. The surface wettability was also shown to be affected by hydrolysis of the poly(butyl acrylate) to the corresponding acid [456].

In conclusion, the possible use of "grafting-from" and "grafting-to" methods to create polymer brushes on surfaces was achieved by using different strategies that enabled the properties of surfaces to be engineered; this represents an important step in the development of "smart" materials. It also provides access to a large number of potential applications in coatings, microfluidics devices, and other systems. Clearly, polymerization techniques that include controlled living radical polymerization, ROMP and cationic and anionic polymerization, have been shown suitable for the creation of nanopatterned polymer brushes.

2.4.2.9 **Others**

The wide variety of chemical modification schemes used to introduce functionality to monolayers and to tune the surface properties, highlights the versatility of using functional SAMs. In addition to the examples discussed in this chapter, a number of other chemical reactions have been used to modify surfaces. These include the Sonogashira coupling, as performed by Qu, where the authors self-assembled

a 1-(allyoxy)-4-iodobenzene on a hydrogenated silicon surface, after which the iodine functional group was coupled to an 1-ethynyl-4-fluorobenzene or 1-chloro-4-ethynylbenzene by a palladium-catalyzed Sonogashia reaction [457]. Shuye *et al*. demonstrated the hydrolysis of a thioester-functionalized SAM into a thiol by immersion in hydrochloric acid solution [393], while Balachander *et al*. described the formation of amino-terminated substrates, by the reduction of either azide- or nitrile-functionalized SAMs. Thiol-terminated substrates may be obtained via the reduction of thiocyanate- or thioester-functionalized SAMs (see Figure 2.16) [383]. Wasserman *et al*. introduced the bromination of alkene-functionalized SAMs by reaction with elemental bromine [382], whereby a mixed SAM of 1-octadecene and 11,11′-oxybis-1-undecene was prepared on a hydrogenated silicon substrate. The alkene functional groups of 11,11′-oxybis-1-undecene were successfully reacted with a first-generation Grubbs catalyst and different substituted alkenes [458].

It has been shown that most of these reactions may also be carried out on structured surfaces, so as to create tailor-made surface properties and binding sites in confined surface areas. Yet, surprisingly few reports are available concerning any structuring approaches that permit the introduction of more than two functional groups. Although, without doubt, these investigations have had a major impact on many fields of research, and also on a wide variety of applications (e.g., as sensor devices and in diagnostics), the issue of preparing such multifunctional structures will place additional demands on the fabrication process, a point which is discussed in the following subsections.

2.4.3
Multifunctionality

Today, the preparation of patterned multifunctional surfaces remains a challenge because compatible functionalization approaches must be used. In this respect, it will become necessary to identify surface reactions and structuring methods that are not exclusive of one another, nor do they destroy the integrity of the individual functional groups. To date, very few research groups have succeeded in preparing such multifunctional surfaces. Indeed, the methods applied have used either photochemical reactions, whereby irradiation is performed through a mask, or the selective deposition of molecules onto a certain spot, whether by pipetting, micro-contact printing, or DPN.

Two research groups in particular have reported the fabrication of multifunctional surface by using chemical photolithography, such that different molecules were self-assembled which provided photocleavable groups required different wavelengths for their cleavage. When irradiation of the first wavelength was applied through a photomask, the first photolabile group was cleaved; after moving the photomask to a different area of the surface, a second wavelength was applied that cleaved the second photolabile group. Ryan *et al*. prepared photosensitive thiol chains that contained two different photocleavable bonds: an *ortho*-nitrobenzyl amine-protecting group that cleaved at 365 nm; and a thiolate bond that cleaved at 220 nm. These thiols were self-assembled onto a gold surface and the substrate was illuminated through a

photomask that permitted illumination only at specific wavelengths, and in specific areas. By using this method, it was possible to introduce three different functional groups onto one substrate [459].

The second method, as introduced by del Campo *et al.*, employed photosensitive triethoxysilanes that had been self-assembled onto quartz or silicon substrates. In a first step, only those molecules with one photocleavable group were assembled; illumination of the substrates through a photomask at a certain wavelength was sufficient to cleave the photosensitive group resulting in the formation of a bifunctional substrate. Moreover, if two different photo-cleavable molecules were introduced, this approach proved to be viable for creating patterns with four different functional regions [460].

Two other groups have reported an ability to prepare multifunctional surfaces by applying micro-contact printing. Renault and coworkers described the formation of patterns with several binding sites for proteins on a surface, by using affinity contact printing, and proceeded to functionalize a PDMS stamp with a protein that acted as antigen when the stamp was immersed in a solution containing different antibodies. The result was that each antigen was bound specifically to a complementary antibody such that, when the stamp was brought into contact with a substrate, the antibodies would be transferred to the surface, creating a microarray of printed proteins [461]. An alternative approach, reported by Geissler *et al.*, involved edge-spreading lithography. This concept was based on the fact that alkanethiol molecules are delivered from a PDMS stamp onto a coinage-metal substrate by a relief structure. From a practical standpoint, silicon beads were used as guides and a PDMS stamp inked with an alkanethiol was pressed onto the gold surface covered with the silicon beads. The ink was transferred to the contact areas of the stamp and the silica beads, and formed a ring around the silica beads on the gold surface. In a further step, a second thiol was delivered by a PDMS stamp onto the gold substrate and formed another circle that nucleated at the edges of the first monolayer. By using this method, concentric rings consisting of different monolayers could be obtained [462].

Multifunctional surfaces were investigated for different biological assays as shown by, for example, Zammatteo *et al.* These authors investigated the coupling of DNA onto glass for creating DNA microarrays by comparing amino-, acid- and aldehyde-functionalized surfaces, and coupling the DNA via either acid- or amine-functional groups. In this way, DNA spots of 400 μm diameter were created on the amino-functionalized surfaces [463]. Beyer *et al.* described the preparation of multifunctional PEG-based arrays, whereby 7-octenyl trichlorosilane and OTS were self-assembled on a glass slide and activated by UV-ozone; subsequently, a UV-induced graft polymerization of PEGMA was performed. On the amine-terminated PEGMA functionalized glass slides, different peptides could be coupled by placing spots of 200 μm diameter, with the spot arrays being utilized as immunoassays [464]. Kim *et al.* prepared spot arrays on glass surfaces to measure the coupling competition of Fmoc amino acids; after optimization, the synthesis of model libraries of biotin-Gly-Ala-P_1-Gly (P_1: one of 19 amino acids) could be performed. These spot arrays were prepared by patterning with a photoresist and perfluorination, after which the

photoresist was washed off and amination performed on the free areas, using various silanes and polymers [144].

2.5 Summary and Outlook

In this chapter, an overview has been presented with regards to the fabrication schemes used to create structured surfaces. The aim was to introduce SPM-based structuring techniques that are especially powerful when combined with self-assembly techniques, as well as functional molecular layers that can be further modified in a chemical sense. The wide diversity of applications that might profit from this concept highlights the importance of further developments in this field. Notably, the special demands of this research requires a strong interaction of different disciplines; indeed, only a combination of different fields of science can provide solutions to the major problems that presently prevent these techniques from being implemented into "real" fabrication processes. Whilst engineering will, to some degree, support the development of instrumental implementations, chemists and materials scientists alike must contribute towards significant improvements of the modification schemes, as highlighted in this chapter. Although, to date, many developments have been made, only selected examples have been included here; however, the wide diversity of chemical reactions and molecular building blocks should enable the incorporation of chemical surface reactions into new fabrication concepts. Furthermore, the availability of a plethora of assembly schemes, confined locally by different structuring techniques and combined with the rules of chemical interactions, may be seen as the major strengths of this nanofabrication strategy.

Overall, the combination of chemically addressable surface templates that can be formed via lithographic techniques, together with reliable modification routines, holds great promise for the realization of a wide variety of structural features that can be implemented to create functional device structures, microfluidic devices, sensors, and diagnostic arrays in the future.

References

1 (a) Binnig, G. and Rohrer, H. (1984) European patent 27517-B1; (b) Binnig, G. and Rohrer, H. (1984) European patent 27517-A1; (c) Binnig, G. and Rohrer, H. (1984) Swiss patent 643397-A5; (d) Binnig, G. and Rohrer, H. (1984) European patent 27517-A; (e) Binnig, G. and Rohrer, H. (1984) US patent 4343993-A; (f) Binnig, G. and Rohrer, H. (1984) European patent 27517-B; (g) Binnig, G. and Rohrer, H. (1984) German patent 3066598-G; (h) Binnig, G. and Rohrer, H. (1984) Swiss patent 643397-A.

2 Binnig, G., Quate, C.F., and Gerber, C. (1986) *Phys. Rev. Lett.*, **56**, 930–933.

3 Dagata, J., Schneir, J., Harary, H.H., Evans, C.J., Postek, M.T., and Bennett, J. (1990) *Appl. Phys. Lett.*, **56**, 2001–2003.

4 Day, H.C. and Allee, D.R. (1993) *Appl. Phys. Lett.*, **62**, 2691–2693.

5 Yasutake, M., Ejiri, Y., and Hattori, T. (1993) *Jpn. J. Appl. Phys.*, **32**, L1021–L1023.

6 Tseng, A.A., Notargiacomo, A., and Chen, T.P. (2005) *J. Vac. Sci. Technol. B*, **23**, 877–894.
7 Garcia, R., Martinez, R.V., and Martinez, J. (2006) *Chem. Soc. Rev.*, **35**, 29–38.
8 Stiévenard, D. and Legrand, B. (2006) *Prog. Surf. Sci.*, **81**, 112–140.
9 Sugimura, H. and Nakagiri, N. (1995) *Jpn. J. Appl. Phys.*, **34**, 3406–3411.
10 Xie, X.N., Chung, H.J., Sow, C.H., and Wee, A.T.S. (2006) *Mater. Sci. Eng. R-Rep.*, **54**, 1–48.
11 Weeks, B.L., Vaughn, M.W., and DeYoreo, J.J. (2005) *Langmuir*, **21**, 8096-2098.
12 Tello, M. and Garcia, R. (2003) *Appl. Phys. Lett.*, **83**, 2339–2341.
13 Tello, M., Garcia, R., Martin-Gago, J.A., Martinez, N.F., Martin-González, M.S., Aballe, L., Baranov, A., and Gregoratti, L. (2005) *Adv. Mater.*, **17**, 1480–1483.
14 Martinez, R.V. and Garcia, R. (2005) *Nano Lett.*, **5**, 1161–1164.
15 Kinser, C.R., Schmitz, M.J., and Hersam, M.C. (2005) *Nano Lett.*, **5**, 91–95.
16 Suez, I., Backer, S.A., and Fréchet, J.M. (2005) *Nano Lett.*, **5**, 312–324.
17 Kim, Y., Kang, S.K., Choi, I., Lee, J., and Yi, J. (2005) *J. Am. Chem. Soc.*, **127**, 9380–9381.
18 Wang, D., Tsau, L., and Wang, K.L. (1994) *Appl. Phys. Lett.*, **65**, 1415–1417.
19 Snow, E.S., Juan, W.H., Pang, S.W., and Campbell, P.M. (1995) *Appl. Phys. Lett.*, **66**, 1729–1731.
20 Snow, E.S., Campbell, P.M., Twigg, M., and Perkins, F.K. (2001) *Appl. Phys. Lett.*, **79**, 1109–1111.
21 Chien, F.S.-S., Hsieh, W.-F., Gwo, S., Vladar, A.E., and Dagata, J.A. (2002) *J. Appl. Phys.*, **91**, 10044–10050.
22 Garcia, R., Calleja, M., and Pérez-Murano, F. (1998) *Appl. Phys. Lett.*, **72**, 2295–2297.
23 Garcia, R. and Calleja, M. (2000) *Appl. Phys. Lett.*, **76**, 3427–3429.
24 Pérez-Murano, F., Abadal, G., Barniol, N., Aymerich, X., Servat, J., Gorostiza, P., and Sanz, F. (1995) *J. Appl. Phys.*, **78**, 6797–6801.
25 Clément, N., Tonneau, D., Dallaporta, H., Bouchiat, V., Fraboulet, D., Mariole, D., Gautier, J., and Safarov, V. (2002) *Physica E*, **13**, 999–1002.
26 Losilla, N.S., Oxtoby, N.S., Martinez, J., Garcia, F., Garcia, R., Mas-Torrent, M., Veciana, J., and Rovira, C. (2008) *Nanotechnology*, **19**, 455308/1–455308/6.
27 Ma, Y.-R., Yu, C., Yao, Y.-D., Liou, Y., and Lee, S.-F. (2001) *Phys. Rev. B*, **64**, 195324/1–195324/5.
28 Vijaykumar, T., Raina, G., Heun, S., and Kulkarni, G.U. (2008) *J. Phys. Chem. C*, **112**, 13311–13316.
29 Mo, Y., Wang, Y., and Bai, M. (2008) *Physica E*, **41**, 146–149.
30 Mori, G., Layyarino, D., Ercolani, L., Sorba, L., Heuen, S., and Locatelli, A. (2005) *J. Appl. Phys.*, **97**, 114324/1–114324/8.
31 Martin-Sánchez, J., Gonzáles, Y., Gonzáles, L., Tello, M., Garcia, R., Ranados, D., Garcia, J.M., and Briones, F. (2005) *J. Cryst. Growth*, **284**, 313–318.
32 Cambel, V. and Soltys, J. (2007) *J. Appl. Phys.*, **102**, 74315/1–74315/7.
33 Lazzarino, M., Padovani, M., Mori, G., Sorba, L., Fanetti, M., and Sancrotti, M. (2005) *Chem. Phys. Lett.*, **402**, 155–159.
34 Okada, Y., Iuchi, Y., Kawabe, M., and Harris, J.S.Jr (2000) *J. Appl. Phys.*, **88**, 1336-1140.
35 Tsai, C.-H., Jian, S.-R., and Wen, H.-C. (2007) *Appl. Surf. Sci.*, **254**, 1357–1362.
36 Jian, S.-R., Fang, T.-H., and Chuu, D.-S. (2005) *J. Phys. D*, **38**, 2424–2432.
37 Lu, Y.F., Mai, Z.H., Qui, G., and Chim, W.K. (1999) *Appl. Phys. Lett.*, **75**, 2359–2361.
38 Hanke, M., Boeck, T., and Gerlitzke, A.-K. (2006) *Appl. Phys. Lett.*, **88**, 173106/1–173106/6.
39 Bo, X.-Z., Rokhinson, L.P., Yin, H., Tsui, D.C., and Strum, J.C. (2002) *Appl. Phys. Lett.*, **81**, 3263–3265.
40 Gwo, S. (2001) *J. Phys. Chem. Solids*, **62**, 1673–1687.
41 Chien, F.S.-S., Chan, J.-W., Lin, S.-W., Chou, Y.-C., Chen, T.T., Gwo, S., Chao, T.-S., and Hsieh, W.-F. (2000) *Appl. Phys. Lett.*, **76**, 360–362.

42 Chien, F.S.-S., Chou, Y.-C., Chen, T.T., Hsieh, W.-F., Chao, T.-S., and Gwo, S. (2001) *J. Appl. Phys.*, **89**, 2465–2472.

43 Fernandez-Cuesta, I., Borrisé, X., and Pérez-Murano, F. (2006) *J. Vac. Sci. Technol. B*, **24**, 2988–2992.

44 Hsu, H.-F. and Lee, C.-W. (2008) *Ultramicroscopy*, **108**, 1076–1080.

45 Choi, I., Yang, Y.I., Kim, Y.-J., Kim, Y., Hahn, J.-S., Choi, K., and Yi, J. (2008) *Langmuir*, **24**, 2597–2602.

46 Unal, K., Aronsson, B.-O., Mugnier, Y., and Descouts, P. (2002) *Surf. Interface Anal.*, **34**, 490–493.

47 Kim, T.Y., Di Zitti, E., Ricci, D., and Cincotti, S. (2008) *Physica E*, **40**, 1941–1943.

48 Kim, T.Y., Di Zitti, E., Ricci, D., and Cincotti, S. (2004) *J. Phys. D*, **37**, 1357–1361.

49 Cooper, E.B., Manalis, S.R., Fang, H., Dai, H., Matsumoto, K., Minne, S.C., Hunt, T., and Quate, C.F. (1999) *Appl. Phys. Lett.*, **75**, 3566–3568.

50 Matsumoto, K., Gotoh, Y., Maeda, T., Dagata, J.A., and Harris, J.S. (2000) *Appl. Phys. Lett.*, **76**, 239–241.

51 Kim, J., Kim, J., Song, K.-B., Lee, S.-Q., Kim, E.-K., and Park, K.-H. (2003) *Jpn. J. Appl. Phys.*, **42**, 7635–7639.

52 Takemura, Y., Kidaka, S., Watanabe, K., Nasu, Y., Yamada, T., and Shrakashi, J.-I. (2003) *J. Appl. Phys.*, **93**, 7346–7348.

53 Lin, H.-N., Chang, Y.-H., Yen, J.-H., Hsu, J.-H., Leu, I.-C., and Hon, M.-H. (2004) *Chem. Phys. Lett.*, **399**, 422–425.

54 Hsu, J.-H., La, H.-W., Lin, H.-N., Chuang, C.-C., and Huang, J.-H. (2003) *J. Vac. Sci. Technol. B*, **21**, 2599–2601.

55 Takemura, Y., Hayashi, S., Okazaki, F., Yamada, T., and Shirakashi, J.-I. (2005) *Jpn. J. Appl. Phys.*, **44**, L285–L287.

56 Bouchiat, V., Faucher, M., Thirion, C., Wernsdorfer, W., Fournier, T., and Pannetier, B. (2001) *Appl. Phys. Lett.*, **79**, 123–125.

57 Snow, E.S., Campbell, P.M., Rendell, R.W., Buot, F.A., Park, D., Marrian, C.R.K., and Magno, R. (1998) *Appl. Phys. Lett.*, **72**, 3071–3073.

58 Amato, J.C. (2004) *Appl. Phys. Lett.*, **85**, 103–105.

59 Archanjo, B.S., Silveira, G.V., Gocalves, A.-M.B., Alves, D.C.B., Ferlauto, A.S., Lacerda, R.G., and Neves, B.R.A. (2009) *Langmuir*, **25**, 602–605.

60 Rolandi, M., Quate, C.F., and Dai, H. (2002) *Adv. Mater.*, **14**, 191–194.

61 Davis, Z.F., Abadal, G., Hansen, O., Borisé, X., Barniol, N., Pérez-Murano, F., and Boisen, A. (2003) *Ultramicroscopy*, **97**, 467–472.

62 Snow, E.S., Park, D., and Campbell, P.M. (1996) *Appl. Phys. Lett.*, **69**, 269–271.

63 Farkas, N., Zhang, G., Evans, E.A., Ramsier, R.D., and Dagata, J.A. (2003) *J. Vac. Sci. Technol. A*, **21**, 1188–1193.

64 Tachiki, M., Fukuda, T., Sugata, K., Seo, H., Umezawa, H., and Kawarada, H. (2000) *Appl. Surf. Sci.*, **159–160**, 578–582.

65 Tachiki, M., Seo, H., Banno, T., Sumikawa, Y., Umezawa, H., and Kawarada, H. (2002) *Appl. Phys. Lett.*, **81**, 2854–2856.

66 Loh, K.P., Xie, X.N., Lim, Y.H., Teo, E.J., Zheng, J.C., and Ando, T. (2002) *Surf. Sci.*, **505**, 93–114.

67 Masubuchi, S., Ono, M., Yoshida, K., Hirakawa, K., and Machida, T. (2008) *Cond. Matter.*, arXiv:0812.0048v1.

68 Weng, L., Zhang, L., Chen, Y.P., and Rokhinson, L.P. (2008) *Appl. Phys. Lett.*, **93**, 93207/1–93207/3.

69 Gordon, A.E., Fayfield, R.T., Litfin, D.D., and Higman, T.K. (1995) *J. Vac. Sci. Technol. B*, **13**, 2805–2808.

70 Stiévenard, D., Fontaine, P.A., and Duboirs, E. (1997) *Appl. Phys. Lett.*, **70**, 3272–3274.

71 Avouris, P., Hertel, T., and Martel, R. (1997) *Appl. Phys. Lett.*, **71**, 285–287.

72 Dagata, J.A., Inoue, T., Itoh, J., and Yokoyama, H. (1998) *Appl. Phys. Lett.*, **73**, 271–273.

73 Dagata, J.A., Inoue, T., Itoh, J., Matsumoto, K., and Yokoyama, H. (1998) *J. Appl. Phys.*, **84**, 6891–6899.

74 Marchi, F., Bouchiat, V., Dallaporta, H., Safarov, V., Tonneau, D., and Doppelt, P. (1998) *J. Vac. Sci. Technol. B*, **16**, 2952–2956.

75 Garcia, R., Calleja, M., and Rohrer, H. (1999) *J. Appl. Phys.*, **86**, 1898–1902.

76 Dagata, J.A., Perez-Murano, F., Abadal, G., Morimoto, K., Inoue, T., Itoh, J., and Yokoyama, H. (2000) *Appl. Phys. Lett.*, **76**, 2710–2712.

77 Dubois, E. and Bubendorff, J.-L. (2000) *J. Appl. Phys.*, **87**, 8148–8154.

78 Snow, E.S., Jernigan, G.G., and Campbell, P.M. (2000) *Appl. Phys. Lett.*, **76**, 1782–1784.

79 Jungblut, H., Wille, D., and Lewerenz, H.J. (2001) *Appl. Phys. Lett.*, **78**, 168–170.

80 Tello, M. and Garcia, R. (2001) *Appl. Phys. Lett.*, **79**, 424–426.

81 Ahn, S.J., Jang, Y.K., Lee, H., and Lee, H. (2002) *Appl. Phys. Lett.*, **80**, 2592–2594.

82 Kuramochi, H., Ando, K., and Yokoyama, H. (2003) *Surf. Sci.*, **542**, 56–63.

83 Dagata, J.A., Perez-Murano, F., Martin, C., Kuramochi, H., and Yokoyama, H. (2004) *J. Appl. Phys.*, **96**, 2386–2392.

84 Dagata, J.A., Perez-Murano, F., Martin, C., Kuramochi, H., and Yokoyama, H. (2004) *J. Appl. Phys.*, **96**, 2393–2399.

85 Lee, S., Pyo, E.K., Kim, J.O., Noh, J., Lee, H., and Ahn, J. (2007) *J. Appl. Phys.*, **101**, 44905/1–44905/5.

86 Daia, H., Frankin, N., and Han, J. (1998) *Appl. Phys. Lett.*, **73**, 1508–1510.

87 Kuramochi, H., Tokizaki, T., Yokoyama, H., and Dagata, J.A. (2007) *Nanotechnology*, **18**, 135703/1–135703/6.

88 Kuramochi, H., Tokizaki, T., Ando, K., Yokoyama, H., and Dagata, J.A. (2007) *Nanotechnology*, **18**, 135704/1–135704/7.

89 Choi, J.S., Bae, S., Ahn, S.J., Kim, D.H., Jung, K.Y., Han, C., Chung, C.C., and Lee, H. (2007) *Ultramicroscopy*, **107**, 1091–1094.

90 Kuramochi, H., Ando, K., Shikakura, Y., Yasutake, M., Tokizaki, T.K., and Yokoyama, H. (2004) *Nanotechnology*, **15**, 1126–1130.

91 Sugimura, H., Uchida, T., Kitamura, N., and Masuhara, H. (1993) *Jpn. J. Appl. Phys.*, **32**, L553.

92 Campbell, P.M. and Snow, E.S. (1998) *Mater. Sci. Eng. B*, **51**, 173–177.

93 Zhang, Y.Y., Zhang, J., Luo, G., Zhou, X., Xie, G.Y., Zhu, T., and Liu, Z.F. (2005) *Nanotechnology*, **16**, 422–428.

94 Minne, S.C., Flueckiger, P., Soh, H.T., and Quate, C.F. (1995) *J. Vac. Sci. Technol. B*, **13**, 1380–1385.

95 Cavallini, M., Mei, P., Biscarini, F., and Garcia, R. (2003) *Appl. Phys. Lett.*, **83**, 5286–5288.

96 Minne, S.C., Manalis, S.R., Atalar, A., and Quate, C.F. (1996) *J. Vac. Sci. Technol. B*, **14**, 2456–2461.

97 Minne, S.C., Adams, J.D., Yaralioglu, G., Manalis, S.R., Atalar, A., and Quate, C.F. (1998) *Appl. Phys. Lett.*, **73**, 1742–1744.

98 Albonetti, C., Martinez, J., Losilla, N.S., Greco, P., Cavallini, M., Borgatti, F., Montecchi, M., Pasquali, L., Garcia, R., and Biscarini, F. (2008) *Nanotechnology*, **19**, 435303/1–435303/9.

99 Farkas, N., Comer, J.R., Zhang, G., Evans, E.A., Ramsier, R.D., Wight, S., and Dagata, J.A. (2004) *Appl. Phys. Lett.*, **85**, 5691–5693.

100 Martinez, R.V., Losilla, N.S., Martinez, J., Tello, M., and Garcia, R. (2008) *Nanotechnology*, **18**, 84021/1–84021/6.

101 Martinez, J., Losilla, N.S., Biscarini, F., Schmidt, G., Borzenko, T., Molenkamp, L.W., and Garcia, R. (2006) *Rev. Sci. Instrum.*, **77**, 86106/1–86106/3.

102 Minne, S.C., Soh, H.T., Flueckiger, P., and Quate, C.F. (1995) *Appl. Phys. Lett.*, **66**, 703–705.

103 Snow, E.S. and Campbell, P.M. (1995) *Science*, **270**, 1639–1641.

104 Ishii, M. and Matsumoto, K. (1995) *Jpn. J. Appl. Phys.*, **34**, 1329–1331.

105 Heinzel, T., Held, R., Lüscher, S., Ensslin, K., Wegschneider, W., and Bichler, M. (2001) *Physica E*, **9**, 84–93.

106 Lüscher, S., Held, R., Fuhrer, A., Heinzel, T., Ensslin, K., Bichler, M., and Wegschneider, W. (2001) *Mater. Sci. Eng. C*, **15**, 153–157.

107 Sigrist, M., Fuhrer, A., Ihn, T., Ensslin, K., Driscol, D.C., and Gossard, A.C. (2004) *Appl. Phys. Lett.*, **85**, 3558–3560.

108 Lüscher, S., Fuhrer, A., Held, R., Heinzel, T., and Ensslin, K. (1999) *Appl. Phys. Lett.*, **75**, 2452–2454.

109 Lüscher, S., Fuhrer, A., Held, R., Heinzel, T., Ensslin, K., Bichler, M., and Wegschneider, W. (2002) *Microelectron. J.*, **33**, 319–321.

110 Fuhrer, A., Lüscher, S., Ihn, T., Heinzel, T., Ensslin, K., Wegschneider, W., and Bichler, M. (2002) *Microelectron. Eng.*, **63**, 47–52.

111 Dorna, A., Sigrist, M., Fuhrer, A., Ihn, T., Heinzel, T., Ensslin, K., Wegschneider, W., and Bichler, M. (2002) *Physica E*, **13**, 719–722.

112 Grbic, B., Leturcq, R., Ihn, T., Ensslin, K., Reuter, D., and Wieck, A.D. (2008) *Physica E*, **40**, 1273–1275.

113 Grbic, B., Letrucq, R., Ensslin, K., Reuter, D., and Wieck, A.-D. (2005) *Appl. Phys. Lett.*, **87**, 232108/1–232108/3.

114 Wouters, D. and Schubert, U.S. (2004) *Angew. Chem., Int. Ed.*, **43**, 2480–2495.

115 Krämer, S., Fuierer, R.R., and Gorman, C.B. (2003) *Chem. Rev.*, **103**, 4367–4418.

116 Rolandi, M., Suez, I., Scholl, A., and Fréchet, J.M.J. (2007) *Angew. Chem., Int. Ed.*, **46**, 7477–7480.

117 Zauscher, S. (2004) Polymeric and biomolecular nanostructures: fabrication by scanning probe lithography, in *Dekker Encyclopedia of Nanoscience and Nanotechnology* (eds J.A. Schwarz, I. Contescu, and K. Putyera), Marcel Dekker, New York.

118 Tang, Q., Shi, S.-Q., and Zhou, L. (2004) *J. Nanosci. Nanotechnol.*, **4**, 948–963.

119 Liu, G.-Y., Xu, S., and Qian, Y. (2000) *Acc. Chem. Res.*, **33**, 457–466.

120 Cruchon-Dupeyrat, S., Porthun, S., and Liu, G.-Y. (2001) *Appl. Surf. Sci.*, **175–176**, 636–642.

121 Wu, C.-H., Sheu, J.-T., Chen, C.H., and Chao, T.-S. (2007) *Jpn. J. Appl. Phys.*, **46**, 6272–6276.

122 Lee, S., Kim, J., Lee, W.S.K., Shin, H.-J., Koo, S., and Lee, H. (2004) *Mater. Sci. Eng.*, **C24**, 3–9.

123 Zhang, M., Bullen, D., Chung, S.-W., Hong, S., Ryu, K.S., Fan, Z., Mirkin, C.A., and Liu, C. (2002) *Nanotechnology*, **13**, 212–217.

124 Wilder, K., Soh, H.T., Atalar, A., and Quate, C.F. (1999) *Rev. Sci. Instrum.*, **70**, 2822–2827.

125 Vettiger, P., Despont, M., Drechsler, U., Dürig, U., Häberle, W., Lutwyche, M.I., Rothuizen, H.E., Stutz, R., Widmer, R., and Binnig, G.K. (2000) *IBM J. Res. Develop.*, **44**, 323–340.

126 Minne, S.C., Adams, J.D., Yaralioglu, G., Manalis, S.R., Atalar, A., and Quate, C.F. (1998) *Appl. Phys. Lett.*, **73**, 1742–1744.

127 Wouters, D. and Schubert, U.S. (2007) *Nanotechnology*, **18**, 485306/1–485306/7.

128 Kakushima, K., Watanabe, T., Shimamoto, K., Gouda, T., Ataka, M., Mamura, H., Isono, Y., Hashiguchi, G., Mihara, Y., and Fujita, H. (2004) *Jpn. J. Appl. Phys.*, **43**, 4041–4144.

129 Watanabe, F., Arita, M., Motooka, T., Okano, K., and Yamada, T. (1998) *Jpn. J. Appl. Phys.*, **37**, L562–L564.

130 Lenhert, S., Sun, P., Wang, Y., Fuchs, H., and Mirkin, C.A. (2007) *Small*, **3**, 71–75.

131 Bullen, D., Wang, X., Zou, J., Chung, S.-W., Mirkin, C.A., and Liu, C. (2004) *Micromech. Syst.*, **13**, 594–602.

132 Bullen, D., Chung, S.-W., Wang, X., Zou, J., Mirkin, C.A., and Liu, C. (2004) *Appl. Phys. Lett.*, **84**, 788–791.

133 Nuzzo, R.G. and Allara, D.L. (1983) *J. Am. Chem. Soc.*, **105**, 4481–4483.

134 Sagiv, J. (1980) *J. Am. Chem. Soc.*, **102**, 92–98.

135 Ulman, A. (1996) *Chem. Rev.*, **96**, 1533–1554.

136 Ulman, A., Kang, J.F., Shnidman, Y., Liao, S., Jordan, R., Choi, G.-Y., Zaccaro, J., Myerson, A.S., Rafailovich, M., Sokolov, J., and Fleischer, C. (2000) *Rev. Mol. Biotechnol.*, **74**, 175–188.

137 Everhart, D.S. (2002) Self-assembling monolayers: alkaline thiols on gold, in *Handbook of Applied Surface and Colloid Chemistry*, vol. 2 (ed. K. Holmberg), Wiley, pp. 99–116.

138 Love, J.C., Estroff, L.A., Kriebel, J.K., Nuzzo, R.G., and Whitesides, G.M. (2005) *Chem. Rev.*, **105**, 1103–1169.

139 Woodruff, D.P. (2007) *Appl. Surf. Sci.*, **254**, 76–81.

140 Mizutani, F. (2008) *Sens. Actuators, B*, **130**, 14–20.

141 Peor, N., Sfez, R., and Yitzchaik, S. (2008) *J. Am. Chem. Soc.*, **130**, 4158–4165.

142 Rittner, M., Martin-Gonzalez, M.S., Flores, A., Schweizer, H., Effenberger, F., and Pilkuhn, M.H. (2005) *J. Appl. Phys.*, **98**, 54312/1–54312/7.

143 Li, Q., Mathur, G., Homsi, M., Surthi, S., Misra, V., Malinovskii, V., Schweikart, K.-H., Yu, L., Lindsey, J.S., Liu, Z., Dabke, R.B., Yasseri, A., Bocian, D.F., and Kuhr, W.G. (2002) *Appl. Phys. Lett.*, **81**, 1494–1496.

144 Kim, D.-H., Shin, D.-S., and Lee, Y.-S. (2007) *J. Pept. Sci.*, **13**, 625–633.

145 Salami, T.O., Yang, Q., Chitre, K., Zarembo, S., Cho, J., and Oliver, S.R.J. (2005) *J. Electron. Mater.*, **34**, 534–540.

146 Matyjaszewski, K., Miller, P.J., Shukla, N., Immaraporn, B., Gelman, A., Luokala, B.B., Siclovan, T.M., Kickelbick, G., Vallant, T., Hoffmann, H., and Pakula, T. (1999) *Macromolecules*, **32**, 8716–8724.

147 Rowe-Konopacki, M.D. and Boyes, S.G. (2007) *Macromolecules*, **40**, 879–888.

148 Mirkin, C.A., Hong, S., and Demers, L. (2001) *ChemPhysChem*, **2**, 37–39.

149 Salaita, K., Wang, Y., and Mirkin, C.A. (2007) *Nat. Nanotechnol.*, **2**, 145–155.

150 Mirkin, C.A. (2000) *Inorg. Chem.*, **39**, 2258–2272.

151 Ginger, D.S., Zhang, H., and Mirkin, C.A. (2004) *Angew. Chem., Int. Ed.*, **43**, 30–45.

152 Haaheim, J. and Nafday, O.A. (2008) *Scanning*, **30**, 137–150.

153 Leggett, G.J. (2005) *Analyst*, **130**, 259–264.

154 Huck, W.T.S. (2007) *Angew. Chem., Int. Ed.*, **46**, 2754–2757.

155 Li, X.-M., Huskens, J., and Reinhoudt, D.N. (2004) *J. Mater. Chem.*, **14**, 2954–2971.

156 Piner, R.D., Zhu, J., Xu, F., Hong, S., and Mirkin, C.A. (1999) *Science*, **283**, 661–663.

157 Jaschke, M. and Butt, H.-J. (1995) *Langmuir*, **11**, 1061–1064.

158 Hong, S., Zhu, J., and Mirkin, C.A. (1999) *Science*, **286**, 523–525.

159 Schwartz, P.V. (2002) *Langmuir*, **18**, 4041–4046.

160 Sheehan, P.E. and Whitman, L.J. (2002) *Phys. Rev. Lett.*, **88**, 156104/1–156104/4.

161 Kim, K.-H., Moldovan, N., and Espinosa, H.D. (2005) *Small*, **1**, 632–635.

162 Moldovan, N., Kim, K.-H., and Espinosa, H.D. (2006) *J. Microelectromech. Syst.*, **15**, 204–213.

163 Zhang, H., Elghanian, R., Disawal, N.A., Amro, S., and Eby, R. (2004) *Nano Lett.*, **4**, 1649–1655.

164 Zou, J., Wang, X., Bullen, D., Ryu, K., Liu, C., and Mirkin, C.A. (2004) *J. Micromech. Microeng.*, **14**, 204–211.

165 Wang, X., Ryu, K.S., Bullen, D.A., Zou, J., Zhang, H., and Mirkin, C.A. (2003) *Langmuir*, **19**, 895–8955.

166 Piner, R.D., Hong, S., and Mirkin, C.A. (1999) *Langmuir*, **15**, 5457–5460.

167 Nelson, B.A., King, W.P., Laracuente, A.R., Sheehan, P.E., and Whitman, L.J. (2006) *Appl. Phys. Lett.*, **88**, 33104/1–33104/3.

168 Sheehan, P.E., Whitman, L.J., King, W.P., and Nelson, B.A. (2004) *Appl. Phys. Lett.*, **85**, 1589–1591.

169 Lee, W.-K., Whitman, L.J., Lee, J., King, W.P., and Sheehan, P.E. (2008) *Soft Matter*, **4**, 1844–1847.

170 Yang, M., Sheehan, P.E., King, W.P., and Whitman, L.J. (2006) *J. Am. Chem. Soc.*, **128**, 6774–6775.

171 Hunag, L., Chang, Y.-H., Kakkassery, J.J., and Mirkin, C.A. (2006) *J. Phys. Chem. Lett. B*, **110**, 20756–20758.

172 Weinberger, D.A., Hong, S., Mirkin, C.A., Wessels, B.W., and Higgins, T.B. (2000) *Adv. Mater.*, **12**, 1600–1603.

173 Jung, H., Kulkarni, R., and Collier, C.P. (2003) *J. Am. Chem. Soc.*, **125**, 12096–12097.

174 Kooi, S.E., Baker, L.A., Sheehan, P.E., and Whitman, L.J. (2004) *Adv. Mater.*, **16**, 1012–1016.

175 Maynor, B.W., Filocamo, S.F., Grinstaff, M.W., and Liu, J. (2002) *J. Am. Chem. Soc.*, **124**, 522–523.

176 Liu, X., Guo, S., and Mirkin, C.A. (2003) *Angew. Chem., Int. Ed.*, **42**, 4785–4789.

177 Wei, J.H., Coffey, D.C., and Ginger, D.S. (2006) *J. Phys. Chem. B*, **110**, 24324–24330.

178 Basabe-Desmonts, L., Wu, C.-C., van der Werf, K.O., Peter, M., Bennink, M., Otto, C., Velders, A.H., Reinhout, D.N., Subramaniam, V., and Crego-Calama, M. (2008) *ChemPhysChem*, **9**, 1680–1687.

179 Willner, I., Baron, R., and Willner, B. (2007) *Biosens. Bioelectron.*, **22**, 1841–1852.

180 Zhang, H., Lee, K.-B., Li, Z., and Mirkin, C.A. (2003) *Nanotechnology*, **14**, 1113–1117.

181 Zhou, H., Li, Z., Wu, A., Wei, G., and Liu, Z. (2004) *Appl. Surf. Sci.*, **236**, 18–24.

182 Demmers, L.M., Ginger, D.S., Park, S.-J., Li, Z., Chung, S.-W., and Mirkin, C.A. (2002) *Science*, **296**, 1836–1838.

183 Nyamjav, D. and Ivanisevic, A. (2003) *Adv. Mater.*, **15**, 1805–1809.

184 Nyamjav, D. and Ivanisevic, A. (2005) *Biomaterials*, **26**, 2749–2757.

185 Cung, S.-W., Ginger, D.S., Morales, M.W., Zhang, Z., Chandrasekhar, V., Ratner, M.A., and Mirkin, C.A. (2005) *Small*, **1**, 64–69.

186 Lee, K.-B., Park, S.-J., Mirkin, C.A., Smith, J.C., and Mrksich, M. (2002) *Science*, **295**, 1702–1705.

187 Agarwal, G., Naik, R.R., and Stone, M.O. (2003) *J. Am. Chem. Soc.*, **125**, 7408–7412.

188 Lee, K.-B., Lim, J.-H., and Mirkin, C.A. (2003) *J. Am. Chem. Soc.*, **125**, 5588–5589.

189 Valiokas, R., Vaitekonis, S., Klenkar, G., Trinkunas, G., and Liedberg, B. (2006) *Langmuir*, **22**, 3456–3460.

190 Lee, M., Kang, D.-K., Yang, H.-K., Park, K.-H., Choe, S.Y., Kang, C., Chang, S.-I., Han, M.H., and Kang, I.-C. (2006) *Proteomics*, **6**, 1094–1103.

191 Kim, J.D., Ahn, D.-G., Oh, J.-W., Park, W., and Jung, H. (2008) *Adv. Mater.*, **20**, 3349–3353.

192 Lee, K.H., Kim, J.D., Kim, Y.J., Kang, S.H., Yung, S.Y., and Jung, H. (2008) *Small*, **4**, 1089–1094.

193 Jinag, H. and Stupp, S.I. (2005) *Langmuir*, **21**, 5242–5246.

194 Hong, S., Zhu, J., and Mirkin, C.A. (1999) *Langmuir*, **15**, 7897–7900.

195 Rosi, N.L. and Mirkin, C.A. (2005) *Chem. Rev.*, **105**, 1547–1562.

196 Zhang, Y., Salaita, K., Lim, J.-H., Lee, K.-B., and Mirkin, C.A. (2004) *Langmuir*, **20**, 962–968.

197 Mirkin, C.A. (2007) *ACS Nano*, **1**, 79–83.

198 Hong, S. and Mirkin, C.A. (2000) *Science*, **288**, 1808–1811.

199 Salaita, K., Lee, S.W., Wang, X., Huang, L., Dellinger, T.M., Liu, C., and Mirkin, C.A. (2005) *Small*, **1**, 940–945.

200 Salaita, K., Wang, X., Fragala, J., Vega, R.A., Liu, C., and Mirkin, C.A. (2006) *Angew. Chem., Int. Ed.*, **45**, 7220–7223.

201 Wang, X., Bullen, D.A., Zou, J., Liu, C., and Mirkin, C.A. (2004) *J. Vac. Sci. Technol. B*, **22**, 2563–2567.

202 Bullen, D., Wang, X., Zou, J., Chung, S.-W., Mirkin, C.A., and Liu, C. (2004) *J. Microelectromech. Syst.*, **13**, 594–601.

203 Bullen, D. and Liu, C. (2006) *Sens. Actuators A*, **125**, 504–511.

204 Lim, J.-H., Ginger, D.S., Lee, K.-B., Heo, J., Nam, J.-M., and Mirkin, C.A. (2003) *Angew. Chem., Int. Ed.*, **42**, 2309–2321.

205 Lee, S.-W., Oh, B.-K., Sanedrin, R.G., Salaita, K., Fujigaya, T., and Mirkin, C.A. (2006) *Adv. Mater.*, **18**, 1133–1136.

206 Li, B., Zhang, Y., Hu, J., and Li, M. (2005) *Ultramicroscopy*, **105**, 312–315.

207 Zhang, H., Li, Z., and Mirkin, C.A. (2002) *Adv. Mater.*, **14**, 1472–1474.

208 Li, S., Szegedi, S., Goluch, E., and Liu, C. (2008) *Anal. Chem.*, **80**, 5899–5904.

209 Hyun, J., Kim, J., Craig, S.L., and Chilkoti, A. (2004) *J. Am. Chem. Soc.*, **126**, 4770–4771.

210 Lim, J.-H. and Mirkin, C.A. (2002) *Adv. Mater.*, **14**, 1474–1476.

211 Maynor, B.W., Filocamo, S.F., Grinstaff, M.W., and Liu, J. (2002) *J. Am. Chem. Soc.*, **124**, 522–523.

212 Liu, X., Guo, S., and Mirkin, C.A. (2003) *Angew. Chem., Int. Ed.*, **42**, 4785–4789.

213 Maedler, C., Chada, S., Cui, X., Taylor, M., Yan, M., and LaRosa, A. (2008) *J. Appl. Phys.*, **104**, 14311/1–14311/4.

214 Su, M. and Dravid, V.P. (2002) *Appl. Phys. Lett.*, **80**, 4434–4436.

215 Noy, A., Miller, A.E., Klare, J.E., Weeks, B.L., Woods, B.W., and DeYoreo, J.J. (2002) *Nano Lett.*, **2**, 109–112.

216 Pena, D.J., Raphael, M.P., and Byers, J.M. (2003) *Langmuir*, **19**, 9028–9032.

217 Mulder, A., Onclin, S., Péter, M., Hoogenboom, J.P., Beijleveld, H.,

ter Maat, J., Garciá-Parajó, M.F., Ravoo, B.J., Huskens, J., van Hulst, N.F., and Reinhoudt, D.N. (2005) *Small*, **1**, 242–253.
218 Yu, M., Nyamjav, D., and Ivanisevic, A. (2005) *J. Mater. Chem.*, **15**, 649–652.
219 Lee, S.W., Sanedrin, R.G., Oh, B.-K., and Mirkin, C.A. (2005) *Adv. Mater.*, **17**, 2749–2753.
220 Zhang, H., Chung, S.-W., and Mirkin, C.A. (2003) *Nano Lett.*, **3**, 43–45.
221 Sheu, J.-T., Wu, C.-H., and Chao, T.-S. (2006) *Jpn. J. Appl. Phys.*, **45**, 3693–3698.
222 Maynor, B.W., Li, Y., and Liu, J. (2001) *Langmuir*, **17**, 2575–2578.
223 Li, Y., Maynor, B.W., and Liu, J. (2001) *J. Am. Chem. Soc.*, **123**, 2105–2106.
224 Liu, X., Fu, L., Hong, S., Dravid, V.P., and Mirkin, C.A. (2002) *Adv. Mater.*, **14**, 231–234.
225 Fu, L., Liu, X., Zhang, Y., Dravid, V.P., and Mirkin, C.A. (2003) *Nano Lett.*, **3**, 757–760.
226 Gundiah, G., John, N.S., Thomas, P.J., Kulkarni, G.U., Rao, C.N.R., and Heu, S. (2004) *Appl. Phys. Lett.*, **84**, 5341–5343.
227 Roy, D., Muny, M., Colombi, P., Bhattacharyya, S., Salvetat, J.-P., Cumpson, P.J., and Saboungi, M.-L. (2007) *Appl. Surf. Sci.*, **254**, 1394–1398.
228 Ding, L., Li, Y., Chu, H., Li, X., and Liu, J. (2005) *J. Phys. Chem. B*, **109**, 22337–22340.
229 Su, M., Liu, X., Li, S.-Y., Dravid, V.P., and Mirkin, C.A. (2002) *J. Am. Chem. Soc.*, **124**, 1560–1561.
230 Dimers, L.M. and Mirkin, C.A. (2001) *Angew. Chem., Int. Ed.*, **40**, 3069–3071.
231 Maynor, B.W., Li, J., Lu, C., and Mirkin, C.A. (2004) *J. Am. Chem. Soc.*, **126**, 6409–6413.
232 Basnar, B., Weizmann, Y., Cheglakov, Z., and Willner, I. (2006) *Adv. Mater.*, **18**, 713–718.
233 Zou, S., Maspoch, D., Wang, Y., Mirkin, C.A., and Schatz, G.C. (2007) *Nano Lett.*, **7**, 276–280.
234 Porter, L.A. Jr, Choi, H.C., Schmeltzer, J.M., Ribbe, A.E., Elliott, L.C.C., and Buriak, J.M. (2002) *Nano Lett.*, **2**, 1369–1372.
235 Degenhart, G.H., Dordi, B., Schönherr, H., and Vansco, G.J. (2004) *Langmuir*, **20**, 6216–6224.
236 Chi, Y.S. and Choi, I.S. (2006) *Adv. Funct. Mater.*, **16**, 1031–1036.
237 Long, D.A., Unal, K., Pratt, R.C., Malkoch, M., and Frommer, J. (2007) *Adv. Mater.*, **19**, 4471–4473.
238 Liu, G.-Y. and Salmeron, M.B. (1994) *Langmuir*, **10**, 367–370.
239 Xiao, X.D., Liu, G.-Y., Charych, D.H., and Salmeron, M.B. (1995) *Langmuir*, **11**, 1600–1604.
240 Liu, G.-Y. and Xu, S. (1999) *ACS Symp. Ser.*, **272**, 199–208.
241 Chwang, A.B., Granstrom, E.L., and Frisbie, C.D. (2000) *Adv. Mater.*, **12**, 285–288.
242 Garno, J.C., Yang, Y., Amro, N.A., Cruchon-Dupeyrat, S., Chen, S., and Liu, G.-Y. (2003) *Nano Lett.*, **3**, 389–395.
243 Zhou, D., Bruckbauer, A., Ying, L., Abell, C., and Klenerman, D. (2003) *Nano Lett.*, **3**, 1517–1520.
244 Kaholek, M., Lee, W.-K., LaMattina, B., Caster, K.C., and Zauscher, S. (2004) *Nano Lett.*, **4**, 373–376.
245 Kaholek, M., Lee, W.-K., Ahn, S.-J., Ma, H., Caster, K.C., Zauscher, B., and LaMattina, S. (2004) *Chem. Mater.*, **16**, 3688–3696.
246 Headerick, J.E., Armstrong, M., Cratty, J., Hammond, S., Sheriff, B.A., and Berrie, C.L. (2005) *Langmuir*, **21**, 4117–4122.
247 Seo, K. and Borguet, E. (2006) *Langmuir*, **22**, 1388–1391.
248 Shi, J. and Cremer, P.S. (2008) *J. Am. Chem. Soc.*, **130**, 2718–2719.
249 Liu, M., Amro, N.A., and Liu, G.-Y. (2008) *Annu. Rev. Phys. Chem.*, **59**, 367–386.
250 Xu, S. and Liu, G.Y. (1997) *Langmuir*, **13**, 127–129.
251 Xu, S., Laibinis, P.E., and Liu, G.Y. (1998) *Langmuir*, **14**, 9356–9361.
252 Xu, S., Miller, S., Laibinis, P.E., and Liu, G.Y. (1999) *Langmuir*, **15**, 7244–7251.
253 Wadu-Mesthridge, K., Xu, S., Amro, N.A., and Liu, G.-Y. (1999) *Langmuir*, **15**, 8580–8583.
254 Xu, S., Amro, N.A., and Liu, G.-Y. (2001) *Appl. Surf. Sci.*, **175–176**, 649–655.

255 Liu, J.-F., Cruchon-Dupeyrat, S., Garno, J.C., Frommer, J., and Liu, G.-Y. (2002) *Nano Lett.*, **2**, 937–940.

256 Liu, M., Amro, N.A., Chow, C.S., and Liu, G.-Y. (2002) *Nano Lett.*, **2**, 863–867.

257 Liu, G.Y. and Amro, N.A. (2002) *Proc. Natl Acad. Sci. USA*, **99**, 5165–5170.

258 Liu, G.-Y. (2005) *Langmuir*, **21**, 1972–1978.

259 Yu, J.J., Tan, Y.H., Li, X., Kuo, P.-K., and Liu, G.-Y. (2006) *J. Am. Chem. Soc.*, **128**, 11574–11581.

260 Case, M.A., McLendon, G.L., Hu, Y., Vanderlick, T.K., and Scoles, G. (2003) *Nano Lett.*, **3**, 425–429.

261 Hu, Y., Das, A., Hecht, M.H., and Scoles, G. (2005) *Langmuir*, **21**, 9103–9109.

262 Staii, C., Wood, D.W., and Scoles, G. (2008) *J. Am. Chem. Soc.*, **130**, 640–646.

263 Staii, C., Wood, D.W., and Scoles, G. (2008) *Nano Lett.*, **8**, 2503–2509.

264 Kenseth, J.R., Harnisch, J.A., Jones, V.W., and Porter, M.D. (2001) *Langmuir*, **17**, 4105–4112.

265 Jang, C.-H., Stevens, B.D., Carlier, P.R., Calter, M.A., and Ducker, W.A. (2002) *J. Am. Chem. Soc.*, **124**, 12114–12115.

266 Nuraje, N., Banerjee, I.A., MacCuspie, R.I., Yu, L., and Matsui, H. (2004) *J. Am. Chem. Soc.*, **126**, 8088–8089.

267 Wadu-Mesthrige, K., Amro, N.A., Garno, J.C., Xu, S., and Liu, G.-Y. (2001) *Biophys. J.*, **80**, 1891–1899.

268 Yu, J.-J., Nolting, B., Tan, Y.H., Li, X., Gervay-Hague, J., and Liu, G.-Y. (2006) *Nanobiotechnology*, **1**, 201–210.

269 Zhou, D., Wang, X.K., Birch, L., Rayment, T., and Abell, C. (2003) *Langmuir*, **19**, 10557–10562.

270 Chung, S.-W., Presley, A.D., Elhadj, S.K., Hok, S., Hah, S.S., Chernow, A.A., Francis, M.B., Eaton, B.E., Feldheim, D.L., and DeYoreo, J.J. (2008) *Scanning*, **30**, 159–171.

271 Wang, X., Zhou, D., Rayment, T., and Abell, C. (2003) *Chem. Commun.*, 474–475.

272 Amro, N.A., Xu, S., and Liu, G.-Y. (2000) *Langmuir*, **16**, 3006–3009.

273 Ngunjiri, J.N., Kelley, A.T., LeJeune, Z.M., Lewandowski, J.R.K., Li, B.R., Serem, W.K., Daniels, S.L., Lusker, K.L., and Garno, J.C. (2008) *Scanning*, **30**, 123–136.

274 Lee, M.V., Hoffman, M.T., Barnett, K., Geiss, J.M., and Smentkowski, V.S. (2006) *J. Nanosci. Nanotechnol.*, **6**, 1639–1643.

275 Lee, M.V., Nelson, K.A., Hutchins, L., Becerril, H.A., Cosby, S.T., Blood, J.C., Wheeler, D.R., Davis, R.C., Woolley, A.T., Harb, J.N., and Linford, M.R. (2007) *Chem. Mater.*, **19**, 5052–5054.

276 Sugimura, H., Okiguchi, K., and Nakagiri, N. (1996) *Jpn. J. Appl. Phys.*, **35**, 3749–3753.

277 Sugimura, H., Okiguchi, K., Nakagiri, N., and Mayashita, M. (1996) *J. Vac. Sci. Technol. B*, **14**, 4140–4143.

278 Kim, J., Oh, Y., Le, H., Shin, Y., and Park, S. (1998) *Jpn. J. Appl. Phys.*, **37**, 7148–7150.

279 Lee, H., Jan, Y.K., Bae, E.J., Lee, W., Kim, S.M., and Lee, S.H. (2002) *Curr. Appl. Phys.*, **2**, 85–90.

280 Sugimura, H., Takai, O., and Nakagiri, N. (1999) *J. Electroanal. Chem.*, **473**, 230–234.

281 Ara, M., Graaf, H., and Tada, H. (2002) *Appl. Phys. Lett.*, **80**, 2565–2567.

282 Graaf, H., Baumgärtel, T., Vieluf, M., and von Borczyskowski, C. (2008) *Superlattices Microstruct.*, **44**, 402–410.

283 Ara, M., Graaf, H., and Tada, H. (2002) *Jpn. J. Appl. Phys.*, **41**, 4894–4897.

284 Xie, X.N., Chung, H.J., Sow, C.H., and Wee, A.T.S. (2004) *Chem. Phys. Lett.*, **388**, 446–451.

285 Li, Q., Zheng, J., and Liu, Z. (2003) *Langmuir*, **19**, 166–171.

286 Zheng, J., Chen, Z., and Liu, Z. (2000) *Langmuir*, **16**, 9673–9676.

287 Tully, D.C., Wilder, K., Trimble, J.M., Fréchet, A.R., and Quate, C.F. (1999) *Adv. Mater.*, **11**, 314–318.

288 Rolandi, M., Suez, I., Dai, H., and Fréchet, J.M. (2004) *Nano Lett.*, **4**, 889–893.

289 Yoshinobu, T., Suzuki, J., Kurooka, H., Moon, W.C., and Iwasaki, H. (2003) *Electrochim. Acta*, **48**, 3131–3135.

290 He, M., Ling, X., Zhang, J., and Liu, Z. (2005) *J. Phys. Chem. B*, **109**, 10946–10951.

291 Martinez, R.V., Garcia, F., Garcia, R., Coronado, E., Forment-Aliaga, A., Romero, F.M., and Tatay, S. (2007) *Adv. Mater.*, **19**, 291–295.

292 Sugimura, H., and Nakagiri, N. (1997) *J. Am. Chem. Soc.*, **119**, 9226–9229.

293 Sugimura, H., Hanji, T., Hayashi, K., and Takai, O. (2002) *Adv. Mater.*, **14**, 524–526.

294 Graaf, H., Vieluf, M., and von Borczyskowski, C. (2007) *Nanotechnology*, **18**, 265306/1–265306/5.

295 Kim, Y., Kang, I., Choi, S.K., Choi, K., and Yi, J. (2005) *Microelectron. Eng.*, **81**, 341–348.

296 Shin, M., Kim, T., Kwon, C., Kim, S.K., Park, J.B., and Lee, H. (2006) *Jpn. J. Appl. Phys.*, **45**, 2076–2081.

297 Shin, M., Kwon, C., Kim, S.K., Kim, H.J., Roh, Y., Hong, B., Park, J.B., and Lee, H. (2006) *Nano Lett.*, **6**, 1334–1338.

298 Lee, W.-K., Caster, K.C., Kim, J., and Zauscher, S. (2006) *Small*, **2**, 848–853.

299 Kim, S.M., Ahn, S.J., Lee, H., Kim, E.R., and Lee, H. (2002) *Ultramicroscopy*, **91**, 165–169.

300 Kim, S.M. and Lee, H. (2003) *J. Vac. Sci. Technol. B*, **21**, 2398–2403.

301 Lee, H., Bae, E., and Lee, W. (2001) *Thin Solid Films*, **393**, 237–242.

302 Lee, W.B., Oh, Y., Kim, E.R., and Lee, H. (2001) *Synth. Met.*, **117**, 305–306.

303 Lee, W., Kim, E.R., and Lee, H. (2002) *Langmuir*, **18**, 8375–8380.

304 Lee, W., Lee, H., and Chun, M.S. (2005) *Langmuir*, **21**, 8839–8843.

305 Jang, J.-W., Sanedrin, R.G., Maspoch, D., Hwang, S., Fujigaya, T., Jeon, Y.-M., Vega, R.A., Chen, X., and Mirkin, C.A. (2008) *Nano Lett.*, **8**, 1451–1455.

306 Bourgoin, J.P., Sudiwala, R.V., and Palacin, S. (1996) *J. Vac. Sci. Technol. B*, **14**, 3381–3385.

307 Lee, H., Kim, S.A., Ahn, J., and Lee, H. (2002) *Appl. Phys. Lett.*, **81**, 138–140.

308 Ahn, S.J., Jang, Y.K., Kim, S.A., Lee, H., and Lee, H. (2002) *Ultramicroscopy*, **91**, 171–176.

309 Yam, C.M., Gu, J., Li, S., and Cai, C. (2005) *J. Colloid Interface. Sci.*, **285**, 711–718.

310 Choi, I., Kang, S.K., Lee, J., Kim, Y., and Yi, J. (2006) *Biomaterials*, **27**, 4655–4660.

311 Maoz, R., Cohen, S.R., and Sagiv, J. (1999) *Adv. Mater.*, **11**, 55–61.

312 Maoz, R., Frydman, E., Cohen, S.R., and Sagiv, J. (2000) *Adv. Mater.*, **12**, 725–731.

313 Pignataro, B., Panebianco, S., Consalvo, C., and Licciardello, A. (1999) *Surf. Interface Anal.*, **27**, 396–400.

314 Pignataro, B., Licciardello, A., Cataldo, S., and Marletta, G. (2003) *Mater. Sci. Eng. C*, **23**, 7–12.

315 Andruzzi, L., Nickel, B., Schwake, G., Rädler, J.O., Sohn, K.E., Mates, T.E., and Kramer, E.J. (2007) *Surf. Sci.*, **601**, 4984–4992.

316 Wouters, D., Willems, R., Hoeppener, S., Flipse, C.F.J., and Schubert, U.S. (2005) *Adv. Funct. Mater.*, **15**, 938–944.

317 Wouters, D., Hoeppener, S., and Schubert, U.S. (2009) *Angew. Chem., Int. Ed.*, **48**, 1732–1739.

318 Hoeppener, S., van Schaik, J.H.K., and Schubert, U.S. (2006) *Adv. Funct. Mater.*, **16**, 76–82.

319 Cai, Y. and Ocko, B.M. (2005) *J. Am. Chem. Soc.*, **127**, 16287–16291.

320 Hoeppener, S., Maoz, R., and Sagiv, J. (2003) *Nano Lett.*, **3**, 761–767.

321 Hoeppener, S., Maoz, R., and Sagiv, J. (2006) *Adv. Mater.*, **18**, 1286–1290.

322 Haensch, C., Hoeppener, S., and Schubert, U.S. (2009) *Nanotechnology*, **20**, 135302 (6pp).

323 Druzhinina, T., Weltjens, W., Hoeppener, S., and Schubert, U.S. (2009) *Adv. Funct. Mater.*, **19**, 1287–1292.

324 Hoeppener, S. and Schubert, U.S. (2005) *Small*, **1**, 628–632.

325 Wouters, D. and Schubert, U.S. (2003) *Langmuir*, **19**, 9033–9038.

326 Becer, C.R., Haensch, C., Hoeppener, S., and Schubert, U.S. (2007) *Small*, **3**, 220–225.

327 Liu, S., Maoz, R., Schmid, G., and Sagiv, J. (2002) *Nano Lett.*, **2**, 1055–1060.

328 Liu, S., Maoz, R., and Sagiv, J. (2004) *Nano Lett.*, **4**, 845–851.

329 Kolb, H.C., Finn, M.G., and Sharpless, K.B. (2001) *Angew. Chem., Int. Ed.*, **40**, 2004–2021.

330 Huisgen, R. (1984) in *1,3-Dipolar Cycloaddition Chemistry*, vol. 1 (ed. A. Padwa,), John Wiley & Sons, pp. 1–176.

331 Wang, X.-Y., Kimyonok, A., and Weck, M. (2006) *Chem. Commun.*, 3933–3935.

332 Flemming, D.A., Thode, C.F., and Williams, M.E. (2006) *Chem. Mater.*, **18**, 2327–2334.

333 Li, H., Cheng, F., Duft, A.M., and Adronov, A. (2005) *J. Am. Chem. Soc.*, **127**, 14518–14524.

334 Fernandez-Megia, E., Correa, J., and Riguera, R. (2006) *Biomacromolecules*, **7**, 3104–3111.

335 Moses, J.E. and Moorhouse, A.D. (2007) *Chem. Soc. Rev.*, **36**, 1249–1262.

336 Basner, B., Xu, J., Li, D., and Willner, I. (2007) *Langmuir*, **23**, 2293–2296.

337 Hoeppener, S., Susha, A.S., Rogach, A.L., Feldmann, J., and Schubert, U.S. (2006) *Curr. Nanosci.*, **2**, 135–141.

338 Wouters, D., and Schubert, U.S. (2005) *J. Mater. Chem.*, **15**, 2353–2355.

339 Hoeppener, S., Maoz, R., Cohen, S.R., Chi, L., Fuchs, H., and Sagiv, J. (2002) *Adv. Mater.*, **14**, 1036–1041.

340 Doron, A., Katz, E., and Willner, I. (1995) *Langmuir*, **11**, 1313–1317.

341 Zhu, T., Fu, X., Mu, T., Wang, J., and Liu, Z. (1999) *Langmuir*, **15**, 5197–5199.

342 Sato, T., Brown, D., and Johnon, B.F.G. (1997) *Chem. Commun.*, 1007–1008.

343 Yonezawa, T., Onoune, S.-Y., and Kunitake, T. (1998) *Adv. Mater.*, **10**, 414–416.

344 Liu, S., Zhu, T., Hu, R., and Liu, Z. (2002) *Phys. Chem. Chem. Phys.*, **4**, 6059–6062.

345 Maoz, R., Frydman, E., Cohen, S.R., and Sagiv, J. (2000) *Adv. Mater.*, **12**, 424–429.

346 Hoeppener, S., van Schaik, J.H.K., Wei, G., and Schubert, U.S. (2005) 13th International Conference on STM, 3–8 July, Sapporo, Japan, p. 234.

347 Chowdhury, D., Maoz, R., and Sagiv, J. (2007) *Nano Lett.*, **7**, 1770–1778.

348 Cheeco, A., Cai, Y., Gang, O., and Ocko, B.M. (2006) *Ultramicroscopy*, **106**, 703–708.

349 Cheeco, A., Gang, O., and Ocko, B.M. (2006) *Phys. Rev. Soc.*, **96**, 56104/1–56104/4.

350 Cai, Y. (2008) *Langmuir*, **24**, 337–343.

351 Woodsen, M. and Liu, J. (2007) *Phys. Chem. Chem. Phys.*, **9**, 207–225.

352 Blacklede, C., Engebretson, D.A., and McDonald, J.D. (2000) *Langmuir*, **16**, 8317–8323.

353 Müller, W.T., Klein, D.L., Lee, T., Clarke, J., McEuen, P.L., and Schultz, P.G. (1995) *Science*, **268**, 272–273.

354 Davis, J.J., Coleman, K.S., Bagshaw, K.L., and Busuttil, C.B. (2005) *J. Am. Chem. Soc.*, **127**, 13082–13083.

355 Davis, J.J., Bagshaw, C.B., Busuttil, K.L., Hanyu, Y., and Coleman, K.S. (2006) *J. Am. Chem. Soc.*, **128**, 14135–14141.

356 Blasdel, L.K., Banerjee, S., and Wang, S.S. (2008) *Langmuir*, **18**, 5055–5057.

357 Péter, M., Li, X.-M., Huskens, J., and Reinhoudt, D.N. (2004) *J. Am. Chem. Soc.*, **126**, 11684–11690.

358 Zorbas, V., Kanungo, M., Bains, S.A., Mao, Y., Hemraj-Benny, T., Misewich, J.A., and Wong, S.S. (2005) *Chem. Commun.*, 4598–4600.

359 Matsubara, S., Yamamoto, H., Oshima, K., Mouri, E., and Matsuoka, H. (2002) *Chem. Lett.*, 886–887.

360 Wang, J., Kenseth, J.R., Jones, V.W., Green, J.-B.D., McDermott, M.T., and Porter, M.D. (1997) *J. Am. Chem. Soc.*, **119**, 12796–12799.

361 Long, D.A., Unal, K., Pratt, R.C., Malkoch, M., and Frommer, J. (2007) *Adv. Mater.*, **19**, 4471–4473.

362 Szoszkiewicz, R., Okada, T., Jones, S.C., Li, T.-D., King, W.P., Marder, S.R., and Rieda, E. (2007) *Nano Lett.*, **7**, 1064–1069.

363 Fresco, Z.M., Suez, L., Backer, S.A., and Fréchet, J.M. (2004) *J. Am. Chem. Soc.*, **126**, 8374–8375.

364 Fresco, Z.M., and Fréchet, J.M. (2005) *J. Am. Chem. Soc.*, **127**, 8302–8303.

365 Pavlovic, E., Quist, A.P., Gelius, U., Nyholm, L., and Oscarsson, S. (2003) *Langmuir*, **19**, 4217–4221.

366 Pavlovic, E., Oscarsson, S., and Quist, A.P. (2003) *Nano Lett.*, **3**, 779–781.

367 Sugimura, H., Lee, S.-H., Saito, N., and Takai, O. (2004) *J. Vac. Sci. Technol. B*, **22**, L44–L46.

368 Saito, N., Lee, S.-H., Takahiro, I., Hieda, J., Sugimura, H., and Takai, O. (2005) *J. Phys. Chem. B*, **109**, 11602–11605.

369 Yam, C.M. and Kakkar, A.K. (1995) *J. Chem. Soc., Chem. Commun.*, 907–909.

370 Petrucci, M.G.L. and Kakkar, A.K. (1998) *Organometallics*, **17**, 1798–1811.

371 Linford, M.R., Fenter, P., Eisenberger, P.M., and Chidsey, C.E.D. (1995) *J. Am. Chem. Soc.*, **117**, 3145–3155.

372 Ge, S., Kojio, K., Takahara, A., and Kajiyama, T. (1998) *J. Biomater. Sci., Polym. Ed.*, **9**, 131–150.

373 Hoffmann, C. and Tovar, G.E.M. (2006) *J. Colloid Interface Sci.*, **295**, 427–435.

374 Yap, F.L. and Zhang, Y. (2007) *Biomaterials*, **28**, 2328–2338.

375 de la Rica, R., Baldi, A., Mendoza, E., Paulo, A.S., Llobera, A., and Fernandez-Sanchez, C. (2008) *Small*, **4**, 1076–1079.

376 Khatri, O.P., Han, J., Ichii, T., Murase, K., and Sugimura, H. (2008) *J. Phys. Chem. C*, **112**, 16182–16185.

377 Changa, L.-W., Yeha, Y.-C., and Lueb, J.-T. (2008) *Sensors & Transducers Journal*, **91**, 91–99.

378 Pang, I., Kim, S., and Lee, J. (2007) *Surf. Coat. Technol.*, **201**, 9426–9431.

379 Feng, Y., Zhou, Z., Ye, X., and Xiong, J. (2003) *Sens. Actuators A*, **108**, 138–143.

380 Netzer, L., Iscovici, R., and Sagiv, J. (1983) *Thin Solid Films*, **99**, 235–241.

381 Maoz, R., and Sagiv, J. (1987) *Langmuir*, **3**, 1045–1051.

382 Wasserman, S.R., Tao, Y.-T., and Whitesides, G.M. (1989) *Langmuir*, **5**, 1074–1087.

383 Natarajan Balachander, N. and Sukenik, C.N. (1990) *Langmuir*, **6**, 1621–1627.

384 Lee, Y.W., Reed-Mundell, J., Zull, J.E., and Sukenik, C.N. (1993) *Langmuir*, **9**, 3009–3014.

385 Cook, M.J., Hersans, R., McMurdo, J., and Russell, D.A. (1996) *J. Mater. Chem.*, **6**, 149–154.

386 Appelhans, D., Ferse, D., Adle, H.-J.P., Plieth, W., Fikus, A., Grundke, K., Schmitt, F.-J., Bayer, T., and Adolphi, B. (2000) *Colloids Surf. A*, **161**, 203–212.

387 Wang, Y., Cai, J., Rauscher, H., Behm, R.J., and Goedel, W.A. (2005) *Chem. Eur. J.*, **11**, 3968–3978.

388 Petrucci, M.G.L. and Kakkar, A.K. (1999) *Chem. Mater.*, **11**, 269–276.

389 Zhang, M., Desai, T. and Ferrari, M. (1998) *Biomaterials*, **19**, 953–960.

390 Sharma, S., Johnson, R.W., and Desai, T.A. (2003) *Appl. Surf. Sci.*, **206**, 218–229.

391 Chi, Y.S., Lee, J.K., Choi, S.-G., and Lee, I.S. (2004) *Langmuir*, **20**, 3024–3027.

392 Fryxell, G.E., Rieke, P.C., Wood, L.L., Engelhard, M.H., Williford, R.E., Graff, G.L., Campbell, A.A., Wiacek, R.J., Lee, L., and Halverson, A. (1996) *Langmuir*, **12**, 5064–5075.

393 Shyue, J.-J. and De Guire, M.R. (2004) *Langmuir*, **20**, 8693–8698.

394 Haensch, C., Ott, C., Hoeppener, S., and Schubert, U.S. (2008) *Langmuir*, **24**, 10222–10227.

395 Haensch, C., Chiper, M., Ulbricht, U., Winter, A., Hoeppener, S., and Schubert, U.S. (2008) *Langmuir*, **24**, 12981–12985.

396 Herzer, N., Hoeppener, S., Schubert, U.S., Fuchs, H., and Fischer, U.C. (2008) *Adv. Mater.*, **20**, 346–351.

397 Hoeppener, S. and Schubert, U.S. (2009) Electro-oxidative lithography and self-assembly concepts for bottom-up nanofabrication, in *Applied Scanning Probe Methods XIII* (eds B. Bhushan and H. Fuchs), ch. 20, Springer, pp. 45–67.

398 Lummerstorfer, T. and Hoffmann, H. (2004) *J. Phys. Chem. B*, **108**, 3963–3966.

399 Rohde, R.D., Agnew, H.D., Yeo, W.-S., Bailey, R.C., and Heath, J.R. (2006) *J. Am. Chem. Soc.*, **128**, 9518–9525.

400 Ciampi, S., Boecking, T., Kilian, K.A., James, M., Harper, J.B., and Gooding, J.J. (2007) *Langmuir*, **23**, 9320–9329.

401 Ostaci, R.-V., Damiron, D., Capponi, S., Vignaud, G., Leger, L., Grohens, Y., and Drockenmuller, E. (2008) *Langmuir*, **24**, 2732–3739.

402 Haensch, C., Hoeppener, S., and Schubert, U.S. (2008) *Nanotechnology*, **19**, 35703/1–35703/7.

403 Gallant, N.D., Lavery, K.A., Amis, E.J., and Becker, M.L. (2007) *Adv. Mater.*, **19**, 965–969.

404 Rozkiewicz, D.I., Janczewski, D., Verboom, W., Ravoo, B.J., and Reinhoudt, D.N. (2006) *Angew. Chem., Int. Ed.*, **45**, 5292–5296.

405 Michel, O. and Ravoo, B.J. (2008) *Langmuir*, **24**, 12116–12118.

406 Ku, S.-Y., Wong, K.-T., and Bard, A.J. (2008) *J. Am. Chem. Soc.*, **103**, 2392–2393.

407 Maoz, R., Cohen, H., and Sagiv, J. (1998) *Langmuir*, **14**, 5988–5993.

408 Flink, S., van Veggel, F.C.J.M., and Reinhoudt, D.N. (2001) *J. Phys. Org. Chem.*, **14**, 407–415.

409 Zhang, G.-J., Tanii, T., Zako, T., Hosaka, T., Miyake, T., Kanari, Y., Funatsu, T., and Ohdomari, I. (2005) *Small*, **1**, 833–837.

410 Fabre, B. and Hauquier, F. (2006) *J. Phys. Chem. B*, **110**, 6848–6855.

411 Chen, M.-S., Dulcey, C.S., Chrisey, L.A., and Dressick, W.J. (2006) *Adv. Funct. Mater.*, **16**, 774–783.

412 Miyake, T., Tanii, T., Kato, K., Zako, T., Funatsu, T., and Ohdomari, I. (2007) *Nanotechnology*, **18**, 305304/1–305304/6.

413 Crivillers, N., Mas-Torrent, M., Perruchas, S., Roques, N., Vidal-Gancedo, J., Veciana, J., Rovira, C., Basabe-Desmonts, L., Ravoo, B.J., Crego-Calama, M., and Reinhoudt, D.N. (2007) *Angew. Chem., Int. Ed.*, **46**, 2215–2219.

414 Duan, X., Sadhu, V.B., Perl, A., Peter, M., Reinhoudt, D.N., and Huskens, J. (2008) *Langmuir*, **24**, 3621–3627.

415 La, Y.-H., Jung, Y.J., Kim, H.J., Kang, T.-H., Ihm, K., Kim, K.-J., Kim, B., and Park, J.W. (2003) *Langmuir*, **19**, 4390–4395.

416 Rozkiewicz, D.I., Ravoo, B.J., and Reinhoudt, D.N. (2005) *Langmuir*, **21**, 6337–6343.

417 Rozkiewicz, D.I., Kraan, Y., Werten, M.W.T., de Wolf, F.A., Subramaniam, V., Ravoo, B.J., and Reinhoudt, D.N. (2006) *Chem. Eur. J.*, **12**, 6290–6297.

418 Rogero, C., Chaffey, B.T., Mateo-Marti, E., Sobrado, J.M., Horrocks, B.R., Houlton, A., Lakey, J.H., Briones, C., and Martin-Gago, J.A. (2008) *J. Phys. Chem. C*, **112**, 9308–9314.

419 Maury, P., Peter, M., Crespo-Biel, O., Ling, X.Y., Reinhoudt, D.N., and Huskens, J. (2007) *Nanotechnology*, **18**, 44007/1–44007/9.

420 Maury, P., Escalante, M., Peter, M., Reinhoudt, D.N., Subramaniam, V., and Huskens, J. (2007) *Small*, **3**, 1584–1592.

421 Netzer, L. and Sagiv, J. (1983) *J. Am. Chem. Soc.*, **105**, 674–676.

422 Maoz, R. and Sagiv, J. (1985) *Thin Solid Films*, **132**, 135–151.

423 Miyake, T., Tanii, T., Kato, K., Hosaka, T., Kanari, Y., Sonobe, H., and Ohdomari, I. (2006) *Chem. Phys. Lett.*, **426**, 361–364.

424 Frydman, E. (1999) Organic self-assembling monolayers as templates for deposition of inorganic materials, Ph.D. thesis, Weizmann Institute, Rehovot, Israel.

425 Sekkat, Z., Wood, J., Geerts, Y., and Knoll, W. (1996) *Langmuir*, **12**, 2976–2980.

426 Hozumi, A., Taoda, H., Saitoa, T., and Shirahatab, N. (2008) *Surf. Interface Anal.*, **40**, 408–411.

427 Brandow, S.L., Chen, M.-S., Aggarwal, R., Dulcey, C.S., Calvert, J.M., and Dressick, W.J. (1999) *Langmuir*, **15**, 5429–5432.

428 Hong, L., Sugimura, H., Furukawa, T., and Takai, O. (2003) *Langmuir*, **19**, 1966–1969.

429 Hadziioannou, G., Patel, S., Granick, S., and Tirrell, M. (1986) *J. Am. Chem. Soc.*, **108**, 2869–2876.

430 Dan, N. and Tirrell, M. (1993) *Macromolecules*, **26**, 4310–4315.

431 Belder, G.F., ten Brinke, G., and Hadziioannou, G. (1997) *Langmuir*, **13**, 4102–4105.

432 Chang, Y.-C. and Frank, C.W. (1996) *Langmuir*, **12**, 5824–5829.

433 Husseman, M., Malmstrom, E.E., McNamara, M., Mate, M.,

Mecerreyes, D., Benoit, D.G., Hedrick, J.L., Mansky, P., Huang, E., Russell, T.P., and Hawker, C.J. (1999) *Macromolecules*, **32**, 1424–1431.

434 Piech, M., George, M.C., Bell, N.S., and Braun, P.V. (2006) *Langmuir*, **22**, 1379–1382.

435 Dong, R., Krishnan, S., Baird, B.A., Lindau, M., and Ober, C.K. (2007) *Biomacromolecules*, **8**, 3082–3092.

436 Wu, T., Gong, P., Szleifer, I., Vlcek, P., Subr, V., and Genzer, J. (2007) *Macromolecules*, **40**, 8756–8764.

437 Jonas, A.M., Hu, Z., Glinel, K., and Huck, W.T.S. (2008) *Macromolecules*, **41**, 6859–6863.

438 Chen, J.-K., Hsieh, C.-Y., Huang, C.-F., Li, P.-M., Kuo, S.-W., and Chang, F.-C. (2008) *Macromolecules*, **41**, 8729–8736.

439 Brinks, M.K., Hirtz, M., Chi, L., Fuchs, H., and Studer, A. (2007) *Angew. Chem., Int. Ed.*, **46**, 5231–5233.

440 Santer, S., Kopyshev, A., Yang, H.-K., and Ruehe, J. (2006) *Macromolecules*, **39**, 3056–3064.

441 Inaoka, S. and Collard, D.M. (1999) *Langmuir*, **15**, 3752–3758.

442 Heise, A., Menzel, H., Yim, H., Foster, M.D., Wieringa, R.H., Schouten, A.J., Erb, V., and Stamm, M. (1997) *Langmuir*, **13**, 723–728.

443 Choi, I.S. and Langer, R. (2001) *Macromolecules*, **34**, 5361–5363.

444 Yoon, K.R., Chi, Y.S., Lee, K.-B., Lee, J.K., Kim, D.J., Koh, Y.-J., Joo, S.-W., Yund, W.S., and Choi, I.S. (2003) *J. Mater. Chem.*, **13**, 2910–2914.

445 Jeon, N.L., Choi, I.S., Whitesides, G.M., Kim, N.Y., Laibinis, P.E., Harada, Y.Y., Finnie, K.R., Girolami, G.S., and Nuzzo, R.G. (1999) *Appl. Phys. Lett.*, **75**, 4201–4203.

446 Harris, R.F., Ricci, M.J., Farrer, R.S., Praino, J., Miller, S.J., Saleh, B.E.A., Teich, M.C., and Fourkas, J.T. (2005) *Adv. Mater.*, **17**, 39–41.

447 Oosterling, M.L.C.M., Sei, A., and Schouten, A.J. (1992) *Polymer*, **33**, 4394–4400.

448 Kim, I.-J. and Faust, R. (2003) *J. Macromol. Sci. A*, **40**, 991–1008.

449 Ingall, M.D.K., Honeyman, C.H., Mercure, J.V., Bianconi, P.A., and Kunz, R.R. (1999) *J. Am. Chem. Soc.*, **121**, 3607–3613.

450 Park, J.-W. and Thomas, E.L. (2002) *J. Am. Chem. Soc.*, **124**, 514–515.

451 Denis, F.A., Pallandre, A., Nysten, B., Jonas, A.M., and Dupont-Gillain, C.C. (2005) *Small*, **1**, 984–991.

452 Brough, B., Christman, K.L., Wong, T.S., Kolodziej, C.M., Forbes, J.G., Wang, K., Maynard, H.D., and Ho, C.-M. (2007) *Soft Matter*, **3**, 541–546.

453 Gaubert, H.E. and Frey, W. (2007) *Nanotechnology*, **18**, 135101/1–135101/7.

454 Dekeyser, C.M., Buron, C.C., Mc Evoy, K., Dupont-Gillain, C.C., Marchand-Brynaert, J., Jonas, A.M., and Rouxhet, P.G. (2008) *J. Colloid Interface Sci.*, **324**, 118–126.

455 Cecchet, F., De Meersman, B., Demoustier-Champagne, S., Nysten, B., and Jonas, A.M. (2006) *Langmuir*, **22**, 1173–1181.

456 LeMieux, M.C., Julthongpiput, D., Bergman, K.N., Cuong, P.D., Ahn, H.-S., Lin, Y.-H., and Tsukruk, V.V. (2004) *Langmuir*, **20**, 10046–10054.

457 Qu, M., Zhang, Y., He, J., Cao, X., and Zhang, J. (2008) *Appl. Surf. Sci.*, **255**, 2608–2612.

458 Dutta, S., Perring, M., Barrett, S., Mitchell, M., Kenis, P.J.A., and Bowden, N.B. (2006) *Langmuir*, **22**, 2146–2155.

459 Ryan, D., Parviz, B.A., Linder, V., Semetey, V., Sia, S.K., Su, J., Mrksich, M., and Whitesides, G.M. (2004) *Langmuir*, **20**, 9080–9088.

460 del Campo, A., Boos, D., Spiess, H.W., and Jonas, U. (2005) *Angew. Chem., Int. Ed.*, **44**, 4707–4712.

461 Renault, J.P., Bernard, A., Juncker, D., Michel, B., Bosshard, H.R., and Delamarche, E. (2002) *Angew. Chem., Int. Ed.*, **41**, 2320–2323.

462 Geissler, M., McLellan, J.M., Chen, J., and Xia, Y. (2005) *Angew. Chem., Int. Ed.*, **44**, 3596–3600.

463 Zammatteo, N., Jeanmart, L., Hamels, S., Courtois, S., Louette, P., Hevesi, L., and Remacle, J. (2000) *Anal. Biochem.*, **280**, 143–150.

464 Beyer, M., Felgenhauer, T., Bischoff, F.R., Breitling, F., and Stadler, V. (2006) *Biomaterials*, **27**, 3505–3514.

3
Physical, Chemical, and Biological Surface Patterning by Microcontact Printing

Jan Mehlich and Bart Jan Ravoo

3.1
Introduction

The technique of printing, which was invented by humankind many thousands of years ago, has through the years undergone steady improvements, notably due to the progress of technology [1–3]. Printing usually involves three components: an ink; an appropriate surface; and a stamp or a press.

Contact printing is an efficient method for pattern transfer, in which a conformal contact between the stamp and the surface of the substrate is the key to success. Printing has the advantage of simplicity and convenience: once a stamp has been made available, multiple copies of the pattern can be produced by repeated inking and printing. Printing is an additive process and, in comparison to lithography, the wastage of material is minimized. Printing can be used to pattern large areas. Furthermore, although contact printing is most suitable for two-dimensional (2-D) patterning, it can also be used to generate three-dimensional (3-D) structures through its combination with other processes.

Microcontact printing (μCP) was developed during the early 1990s by Whitesides and coworkers [4]. Just like conventional printing, μCP also involves an ink, a substrate and a stamp; however, in contrast to the dyes that are normally used for printing, the inks for μCP are printed in monomolecular layers and, instead of paper, clothing, stone or wood, the surfaces for μCP are usually ultraflat metal, silicon, or glass substrates. But, perhaps the most remarkable difference is that, instead of macroscopic patterns, the stamps for μCP carry features in the micrometer range [5], or even at the nanoscale [6, 7]. As a result, within less than two decades μCP has emerged as a straightforward and cheap bench-top method for the preparation of both microstructured and nanostructured surfaces.

In this chapter, the potential of μCP is reviewed with regards to the chemical, physical and biological patterning of surfaces by μCP. First, a general introduction to the method of μCP will be provided, including a short discussion of the main advantages and limitations of the process in the preparation of microstructured and nanostructured surfaces. The broad range of inks that can be printed in monolayers

Nanotechnology, Volume 8: Nanostructured Surfaces. Edited by Lifeng Chi

ISBN: 978-3-527-31739-4

by μCP, subject to a suitable modification of stamp and substrate, will then be described. In addition, an outline will be provided of the way in which μCP can provide physical surface structures by "soft lithography," and how μCP can be used to prepare biological microarrays. Exactly how μCP can be used to induce and direct chemical reactions on a surface will also be discussed. Finally, the major innovations that have been proposed to improve the resolution of μCP will be detailed, followed by a brief outlook of the situation.

At this point, it should be emphasized how quickly μCP has found widespread application throughout the scientific community. In fact, according to the ISI Web of Knowledge, almost 1000 articles involving μCP – including two recent reviews [8, 9] – have been produced to date. However, rather than simply provide an exhaustive review, the decision was taken to highlight the most important developments and applications of μCP for the preparation of microstructured and nanostructured surfaces.

3.2 What is Microcontact Printing?

In its most simple version, μCP is a nonphotolithographic method that readily provides patterned self-assembled monolayers (SAMs) with submicron lateral resolution. It offers remarkable experimental simplicity, and can be performed in almost any laboratory, without a need for "clean room" conditions. Moreover, as μCP is a cheap and straightforward process, it has rapidly found widespread applications in different areas of research since its invention during the early 1990s [4].

3.2.1 The "Master"

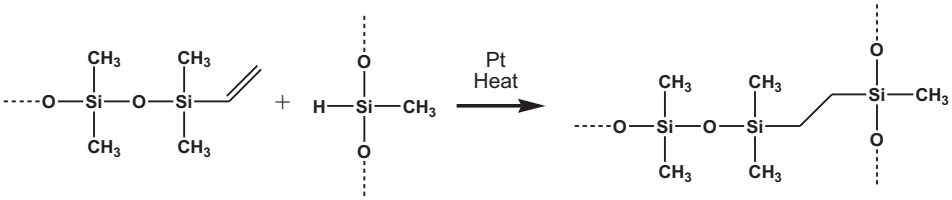
The initial step in each μCP experiment is the design and production of a "master" that can be designed with the help of simple computer software. The desired pattern is first transferred from the master to a surface of choice. When a master pattern has been established by common photolithographic methods on a flat substrate, such as a silicon wafer, it can easily be replicated by making elastomeric stamps with the negative image of the master. To achieve this, a liquid polymer precursor is poured onto the master, allowed to polymerize, and then released from the master such that the pattern is transferred as a microrelief structure at the surface of the hardened polymer. This stamp is then "inked" with the molecules that are to be printed, either by wetting the surface of the stamp with a solution of the ink molecules, by immersing the stamp in a solution of the ink, or simply by placing the stamp on an ink pad. In this situation, small, low-molecular-weight inks will be absorbed into the polymer network of the stamp, whereas large, high-molecular-weight inks, as well as nanoparticles (NPs) and colloids, will be coated onto the surface of the stamp. When the inked stamp is placed on a substrate, in those protruding areas where the stamp is in conformal contact with the substrate, the ink will be transferred. However,

in the receding areas of the stamp there will be no contact with the substrate, and no ink will be transferred. The substrate may be a metal or metal oxide, a silicon wafer or glass, a polymer film, or a SAM, while the ink should possess functional groups that allow its chemisorption onto the surface. In the seminal studies conducted by Whitesides and coworkers, the stamp was prepared from polydimethylsiloxane (PDMS), the ink was *n*-octadecylthiol (ODT), and the substrate was a silicon wafer coated with a thin film of gold [4]. The entire procedure is illustrated schematically in Figure 3.1.

3.2.2 The Stamp

The stamp, which is key to the success of μCP as it is used to generate the pattern, is usually prepared from silicone polymers, among which PDMS (available commercially as Sylgard 184) is the most commonly used. Sylgard 184 is not only transparent but also has a low viscosity before being cured; both features are highly favorable when producing stamps for use in soft lithography. PDMS is also easy to handle and inexpensive (the typical cost of a stamp is much less than €1). Notably, PDMS has a very low resistance to most nonpolar solvents, and although it will not dissolve in such solvents it will undergo substantial deformation as a result of its swelling. Fortunately, however, the stamp will regain its original shape when the solvent has evaporated. Consequently, ink solutions should preferably be prepared in polar solvents such as ethanol, methanol, or water.

Sylgard 184 is a two-component heat-curing system; that is, it consists of a base and a curing agent mixed in a ratio of 10 : 1. The elastomer base is a PDMS with terminal ethylene groups, while the curing agent consists of much shorter PDMS chains, with many of the methyl groups substituted by hydrogen atoms. In the presence of Pt (in Sylgard 184, Pt is added to the base component) the polymerization follows the reaction shown in Scheme 3.1.



Scheme 3.1 Pt-catalyzed cross-linking of poly(dimethylsiloxane) (PDMS) with curing agent.

In the curing reaction of PDMS, the details of which have been elucidated [10], Pt(II) is initially coordinated by two terminal ethylene groups of the precursor polymer. In an oxidative addition – that is, when Pt(II) is oxidized to Pt(IV) – a hydrosiloxane unit of a curing agent molecule becomes coordinated to Pt. Then, after a migratory insertion of the hydrogen atom to one of the ethylene groups, the connection between the curing agent moiety and the PDMS polymer is made in a

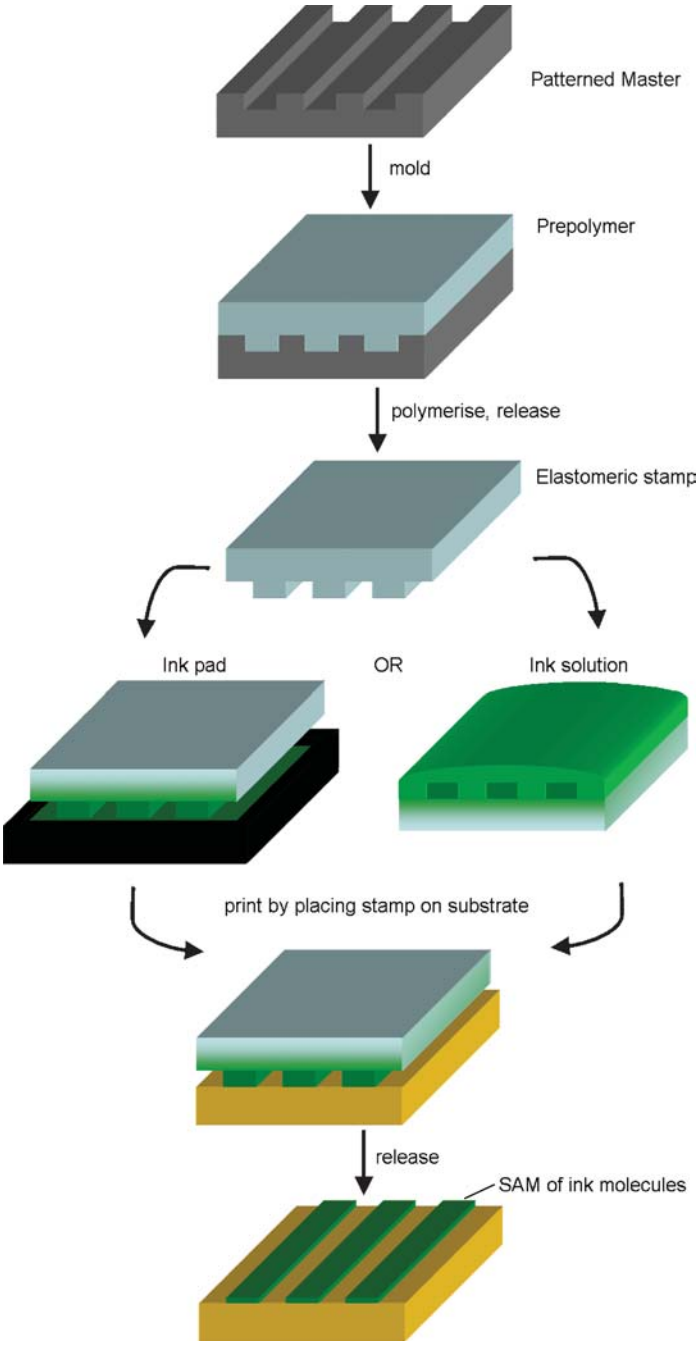


Figure 3.1 The principle of microcontact printing (μCP). The stages include: molding of a stamp; inking of a stamp; printing on a suitable substrate; and release of the stamp from the substrate. In the seminal studies of Whitesides and coworkers the stamp was produced from poly(dimethylsiloxane) (PDMS), the ink was *n*-octadecylthiol (ODT), and the substrate a silicon wafer coated with a thin film of gold [4].

reductive elimination – that is, Pt(IV) is reduced to Pt(II). Subsequently, another ethylene-terminated polymer can coordinate to Pt(II), and a further crosslink can be made such that the result is a more or less dense network of crosslinked polymer chains. Together, the curing time and temperature determine the extent of crosslinking, and hence the elasticity and the stiffness of the stamp. A significant shrinkage of the stamp must be taken into account when curing at high temperatures ($>100\,^{\circ}C$).

PDMS stamps are mainly characterized by the two opposing properties of stiffness and elasticity, which are expressed in the Young's modulus (which, for a PDMS stamp is typically about 1.5 MPa). On the one hand, a stamp should be mechanically stable or the pattern will be blurred upon contact with the substrate; this means that the stamp must be sufficiently stiff. On the other hand, as a conformal contact between the stamp and substrate is required, the elasticity of the stamp must be substantial. The stiffer the stamp, the greater will be the reduction in sagging and collapse of the stamp upon demolding. A stiffer stamp will also generally improve the accuracy of replication. Whilst the main disadvantage of a stiffer stamp is an increased brittleness, a greater elasticity will compensate for an uneven surface and ensure conformal contact also with uneven surfaces.

The principle of conformal contact is illustrated in Figure 3.2. Conformal contact comprises the macroscopic adaptation to the overall shape of the substrate, as well as the microscopic adaptation of a soft polymer layer to a rough surface, leading to an intimate contact without voids. The elastic adaption is caused by adhesion forces such that, even without the application of any external pressure, the stamp can spontaneously compensate for some degree of substrate roughness, depending on the material's properties [11]. The elastomer (the light gray layer in Figure 3.2) compensates for local surface roughness amplitudes of up to 1 μm, whereas long-range warp (wavelengths $>100\,\mu m$) is compensated by the flexibility of the backplane (the dark gray layer in Figure 3.2, which may be a metal, glass, or polymer). Conformal contact benefits from a low Young's modulus and a moderate (yet sufficient) work of adhesion.

The quality of the stamp also depends on the dimensions and depth of the pattern. The pattern dimensions can be characterized by the *aspect ratio* and the *fill ratio*. As illustrated in Figure 3.3, a microrelief stamp can be defined according to the width w of the protruding features, the height h of a protruding features, and the distance d between two protruding features (Figure 3.3a). The aspect ratio is the height h of features divided by their width, w, while the fill ratio is given by the width w of the features divided by their distance, d. Features of high aspect ratio ($h/w > 2$) exhibit

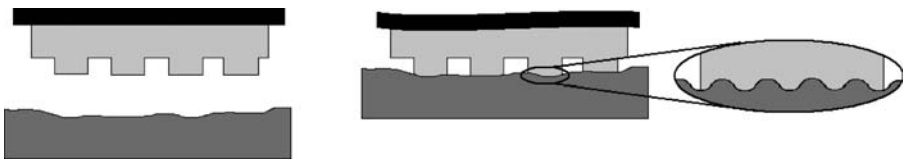


Figure 3.2 Conformal contact of an elastomer stamp (light gray) and a solid surface (dark gray).

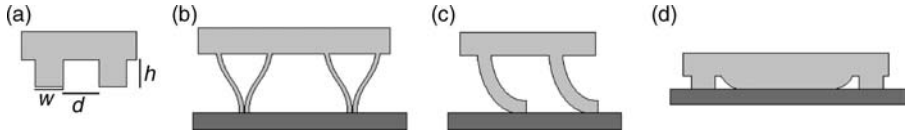


Figure 3.3 Effect of feature dimensions on the stability of an elastomer microrelief stamp. For details, see the text.

lateral instabilities (Figure 3.3b and c), where the structures collapse while peeling off the template or during the inking process due to capillary action (Figure 3.3b), or they collapse against the substrate such that a side of the feature comes into contact with the substrate (Figure 3.3c). On the other hand, voids in a stamp with a low fill ratio ($h/d < 0.2$) are susceptible to sagging (Figure 3.3d) [11–13].

Typically, μCP is used for printing at the microscale – that is, with smallest features of about 0.5 μm. A number of factors determine the smallest features that can be printed, with the fundamental limits to printing being determined by three main constraints: (i) the minimum size of features in the stamp; (ii) the lateral dimensions and resolution of the ink; and (iii) the adhesion and spreading of the ink at the substrate surface. The smallest feature in the stamp depends on the size of features within the master, the fidelity of the molding process, and the ability of the elastomer mold to retain nanoscale features. Distortion of the stamp while in contact with the printed surface also limits the minimum size of the transferred feature. In the best case, composite stamps of PDMS can retain 100 nm features without collapse [14, 15]. Some strategies proposed to extend the resolution of μCP into the nanoscale are outlined in Section 3.7.

Inspired by the pioneering studies of Whitesides and colleagues, many research groups have since shown that the nature of the ink, the stamp, and the substrate can be widely modified, not only to improve the printing quality but also to exploit μCP for a broad range of applications. These achievements are described in detail in the following sections.

3.3
Inks and Stamps for Microcontact Printing

The role of μCP was first demonstrated for the preparation of patterned SAMs of *n*-alkylthiols on gold substrates [4, 16], and this remains today the most widely studied and best established use of the technique. Several points have been identified as being responsible for the success of this particular combination, namely:

- *n*-Alkylthiols are rapidly chemisorbed onto gold surfaces by the formation of a coordinative bond between gold and sulfur atoms.
- A dense and highly ordered monolayer of molecules is formed spontaneously, due to strong van der Waals interactions between the long alkyl chains.
- The self-limiting nature of the forming monolayer favors its confinement to the area of contact.

The basic process for forming patterned SAMs of *n*-alkylthiols on gold is conceptually simple. The stamp is first impregnated with a 1 mM solution of *n*-alkylthiol in ethanol, and then placed in contact with a clean gold surface; this causes the thiols to diffuse from the stamp onto the surface, where they assemble into an ordered monolayer. However, investigations into the details of this self-assembly phenomenon have suggested the process to be complex and to depend on a number of parameters, including the choice of the SAM-forming molecules, the concentration of molecules in the ink solution, the contact time, and the pressure applied to the stamp [17–19].

The mechanisms for the mass transport of thiols during μCP include (at least) the following:

- Diffusion from the bulk of the stamp to the interface between the stamp and the surface of the gold contacted by the stamp.
- Diffusion away from the edges of the stamp and across the surface of the gold substrate.
- Vapor transport through the gas phase (see Figure 3.1).

The first of these mechanisms is important for the formation of SAMs in regions where the stamp should be in contact with the surface; however, very little information is available regarding relevant parameters such as the rates of diffusion of *n*-alkylthiols (or other nonpolar molecules) in PDMS. The second and third mechanisms are important for understanding and controlling the lateral diffusion of SAMs into regions that are not contacted by the stamp. These are unwanted processes that lead to distortions of the lateral dimensions of the printed features and gradients of surface coverage at the edges of the printed structures. Whilst the relative contributions of each of these mechanisms in the formation of the SAMs in the area contacted by the stamp and in the noncontact area are not completely understood [19], much is known regarding the mechanisms of SAM formation, and the structure of *n*-alkylthiol SAMs on gold in particular. In general, these principles should be the same for μCP-mediated SAM formation in the contact areas between the stamp and the gold substrate. As with SAMs formed from solution, in μCP-controlled SAMs the monolayer not only contains perfect domains of slightly tilted aligned molecules but also invariably includes areas with less-ordered molecules, or even "collapsed" orientations with the molecules not standing upright but laying flat on the surface (Figure 3.4).

Patterned SAMs formed by μCP can be easily visualized using a range of techniques that include scanning electron microscopy (SEM), scanning probe microscopy (SPM), secondary ion mass spectrometry (SIMS), condensation figures observed in optical microscopy, and surface-enhanced Raman microscopy. In Figure 3.5 are shown the lateral force microscopy (LFM) images of patterned SAMs of *n*-hexadecanethiol (HDT) on gold [20]. In this case, the surface was patterned with HDT, after which the remaining regions were covered with 16-mercaptohexadecanoic acid (MUA) by immersing the patterned sample in an ethanolic solution of MUA. Relatively high frictional forces between the probe and the surface were detected in regions covered with a COOH-terminated SAM (light areas), while

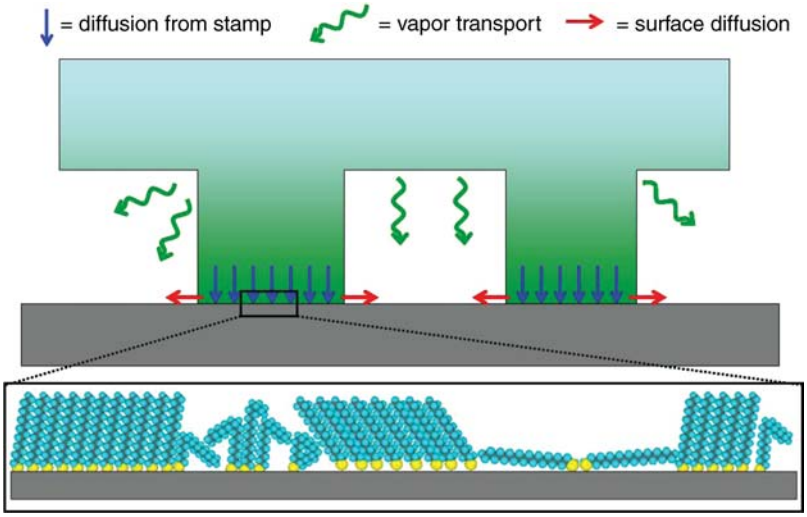


Figure 3.4 Ink transport in microcontact printing of *n*-alkylthiols and structure of *n*-alkylthiol SAMs on gold substrates.

relatively low frictional forces were measured in regions covered with a CH_3-terminated SAM (dark areas).

μCP is not limited to printing thiols on gold, however. Indeed, it has been shown that, subject to a suitable modification of stamp and substrate, other small molecules may also be printed. The μCP-mediated formation of patterned monolayers of *n*-alkylsilanes such as *n*-octadecyltrichlorosilane (OTS) adsorbed onto oxide surfaces such as glass, Al_2O_3, SiO_2, and ITO has been investigated in detail [21]. It has been found that the OTS chains can pack with densities approaching those found in bulk hydrocarbon crystals, but that even the highest-quality printed monolayers of *n*-alkylsilanes lack the long-range ordering found for *n*-alkylthiol SAMs on Au and Ag. It is generally believed that the adsorption of *n*-alkyltrichlorosilanes and other *n*-alkylsilanes with hydrolyzable bonds proceeds on hydrated surfaces via the formation of silanols as intermediates, which then react in turn laterally or with surface OH groups to form a network polymer which is covalently bound (to some degree) to the surface [22]. The resultant films have significant mechanical, thermal, and chemical stability, with infrared (IR) spectroscopy, ellipsometry, and contact angle measurements indicating a high degree of structural organization in such films. The largely all-*trans* alkyl chains are usually found to be tilted at ~10° from the surface normal direction [23].

A PDMS stamp is quite hydrophobic and suitable for apolar ink molecules such as long-chain *n*-alkylthiols and *n*-alkylsilanes. However, polar and hydrophilic inks cannot be printed efficiently with PDMS stamps, and special surface treatments that enable the printing of such inks that otherwise would not adhere to PDMS due to its hydrophobic surface are required. When the PDMS surface is exposed to oxygen plasma or ozone, the surface becomes hydrophilic due to the formation of a thin and

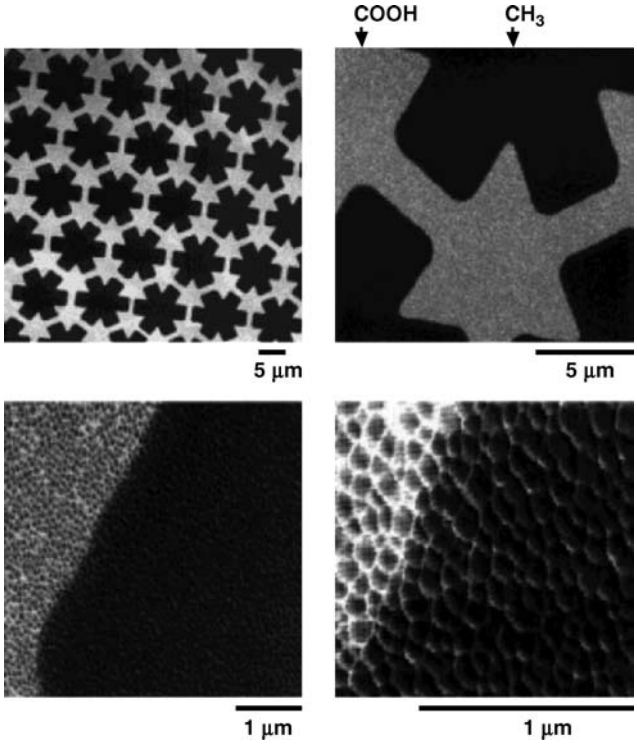


Figure 3.5 Lateral force microscopy at four different magnifications of a gold surface patterned with SAMs terminated in different head groups. Reproduced with permission from Ref. [20]; © 1995, American Chemical Society.

brittle silica-like layer that causes changes in the mechanical properties of PDMS. Owen and Smith [24] studied the formation of cracks in this silica-like layer, and showed that the cracks may allow the migration of low-molecular-weight PDMS fragments to the surface, leading to a recovery of the hydrophobic character of the PDMS surface. Hydrophobic recovery always occurs with time after exposure to oxygen plasma or ozone [25]. The chemical attachment of hydrophilic chlorosilanes and/or grafting of hydrophilic polymers on the oxidized PDMS surface to tune the surface energy of the PDMS stamp have been reported [26–28].

In view of the range of possible applications, there is today great interest in the patterning of bio(macro)molecules on surfaces. In particular, the mild conditions of µCP make it an attractive method for the patterning of biomolecules. For example, µCP can be used to transfer proteins onto a variety of substrate materials with-hydrophilic or hydrophobic surfaces, including bare and silanized glass, gold, silicon and silicon oxide, poly(styrene), poly(methylmethacrylate) (PMMA), and various monolayers on gold [29, 30]. One important advantage of µCP is that most proteins retain their biological activity after printing. When printing proteins on substrates, three types of immobilization can be distinguished, namely physisorption,

chemisorptions, and attachment to protein-resistant surfaces. Similar to printing proteins, the μCP of DNA calls for carefully tailored surface properties of the PDMS stamps, since DNA is a highly negatively charged polyelectrolyte due to its phosphate backbone and electrostatic interactions may determine its adsorption and transfer properties [31]. Since μCP has today become a valuable tool in the preparation of biological microarrays, Section 3.5 of this chapter is focused on the μCP of biomolecules.

Supramolecular μCP

Supramolecular μCP is a variation of μCP where the interaction between ink and substrate is tuned by supramolecular (noncovalent) interactions. In a series of publications, Reinhoudt and Huskens and colleagues have shown that gold, glass and silicon surfaces functionalized with β-cyclodextrin host molecules form a suitable platform ("molecular printboard") to print guest molecules [32, 33]. The interaction between ink and substrate in supramolecular μCP is highly specific, and can be tuned via the multivalency of the ink – that is, the number of interactions with the substrate. Among others, these research groups have printed redox active dendrimers [34] and proteins [35] on such molecular printboards.

The terms "bottom-up" and "top-down" refer to two strategies in nanotechnology that are also evident in μCP with macromolecules. The term "printing polymers," for example, can be appreciated in two ways. In the top-down approach, a pattern of polymers is created by printing a polymer ink, such that a pattern of small features – usually just a monolayer on a substrate – is formed from a bulk polymer. In the bottom-up approach, either a monomer is patterned followed by polymerization on the substrate, or a "seed" for polymer growth is transferred in a patterned manner onto the substrate by μCP. Surface-initiated polymerization results in covalently bound, dense polymer brushes. The same principle applies to NPs and nanotubes (NTs), which can either be printed as such (e.g., with NP and NT inks) or synthesized on the substrate by printing a template. Since this is an important soft lithographic method underlining the importance of μCP for nanotechnology, this topic is described in more detail in Section 3.4; however, it should be noted that printing catalysts for the post-printing formation of much larger structures is possible. This illustrates the fact that μCP is not restricted to organic ink molecules (such as all inks discussed so far), but may also be applied to inorganic compounds.

Macromolecules are particularly useful as inks for μCP, as they tend to adhere strongly to the contact area and diffuse only slowly into noncontact areas. Simple polymers such as PMMA can be patterned using μCP, subject to some modification of the printing process [36]. As a stamp which is directly inked with a PMMA solution in chloroform will become deformed due to swelling, the stamp must be inked using the Langmuir–Schaefer film transfer technique. In this case, the inked stamp was brought into contact with a layer of PMMA on water, which resulted in a thin PMMA layer being deposited on the PDMS stamp. Printing the inked stamp on a silicon oxide substrate then led to the creation of patterned PMMA layers on the surface.

Silicon wafers were patterned with dendrimers of poly(amidoamine) (PAMAM) [6], resulting in 140 nm-wide lines of a single dendrimer layer. Patterns of amine-terminated PAMAM were used as stabilizers for the growth of photoluminescent CdS NPs, simultaneously functioning as adhesive layer between the particles and the silicon surface [37]. Amine-terminated PAMAM was also used to pattern reactive dendrimers on activated SAMs on gold [38]. The deposition of dendrimer multilayers on several substrates by μCP, and the effect of ink concentration, contact time and inking method have also been recently studied [39].

Unlike small apolar ink molecules such as *n*-alkylthiols, which are absorbed into the stamp and then transferred upon contact between the stamp and substrate, high-molecular-weight inks are not absorbed by the stamps but are merely adsorbed to the surface of the stamp. Hence, most of the ink is transferred in a single printing step, and re-inking is necessary after every print; however, this issue can be resolved by changing the composition and structure of the stamp. In this case, agarose has been exploited as a stamp material as it offers certain advantages, especially for printing large molecules such as biomolecules or even cells [40]. The high permeability of agarose for water makes it suitable for printing water-soluble biomacromolecules. In addition, the agarose stamp functions as an ink reservoir that releases the ink molecules slowly, which in turn enables multistep printing without re-inking.

More recently, an alternative method for microstructuring various polymer-based materials was developed, termed phase-separation micromolding (PSμM) [41, 42]. This versatile microfabrication technique can be used to structure a broad range of polymers, including block copolymers, and biodegradable and conductive polymers, without the need for clean room facilities. The method relies on the phase separation of a polymer solution while in contact with a structured mold. For this, a mixture of polymers is cast onto a patterned mold and then placed in a nonsolvent (e.g., water) where polymer chains of the soluble component leave the bulk-producing pores, which immediately become filled by the nonsolvent (Figure 3.6). By using such porous materials as stamps, high-molecular-weight polar inks such as poly(propylene imine) (PPI) dendrimer, HIgG-Fc protein, and functionalized silica NPs were successfully transferred from the stamps to the substrates [43]. The pores not only enable the attachment of such large molecules, but also serve as an ink reservoir for

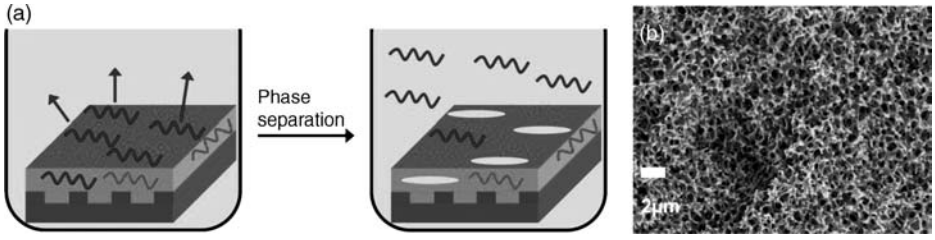


Figure 3.6 (a) The principle of phase separation micromolding: a polymer mixture on a mold is exposed to a nonsolvent; (b) Scanning electron microscopy image of a porous stamp. Reproduced with permission from Ref. [43]; © 2009, American Chemical Society.

repeated printing steps, without a need for re-inking and with no loss of printing quality. With these inks, the PDMS stamps can be used for only one printing step before showing a significant reduction in ink transfer.

Metals in the form of salts [44] and metal colloids [45] can also be patterned on appropriate substrates by using μCP. The μCP of colloids provides access to submicron metal structures, and is a flexible technique that allows patterning on a variety of substrates, including glass, (Si/SiO_2), and polymers. Moreover, both flat and curved surface substrates can be used without any loss of resolution. The μCP of colloids can also be used to produce free-standing metal structures and metal films with different thicknesses. For example, Hidber and coworkers used μCP to selectively seed substrates with palladium particles [45] by first coating a stamp with tetra-alkylammonium- or tetraoctadecylammonium-stabilized palladium NPs. The stamp was then contacted with a silanized substrate and the NPs were transferred, followed by electroless metallization with copper (Figure 3.7). The silane layer ensured bonding between the substrate and the NPs [45, 46]. Later, Kind and coworkers coated a stamp with a catalytic precursor ink [47], whereby the stamp was first brought into contact with a titanium-coated substrate, after which the Pd(II) in the ink and the titanium on the surface reacted chemically to form a Pd(0) catalytic pattern. Some additional applications of electroless deposition (ELD) are described in Section 3.4; these include the μCP of colloids for the preparation of surface-enhanced Raman scattering (SERS)-active substrates by attaching silver NPs to gold NP structures that have been patterned using μCP [48].

One final method should be mentioned that does not follow the typical principle of printing an ink onto a substrate but, without doubt, must be regarded as a variation of microcontact printing. Chen and coworkers recently reported the concept of "microcontact deprinting," which involved a microstructured poly(styrene) stamp placed on a monolayer of poly(styrene)-*block*-poly(2-vinylpyridine) micelles on a silicon wafer [49]. Lifting off the stamp caused the micelles in the contact area to be removed (Figure 3.8); these micelles served as initiators for the growth of NPs, such that a patterned bottom-up fabrication of NPs could be achieved.

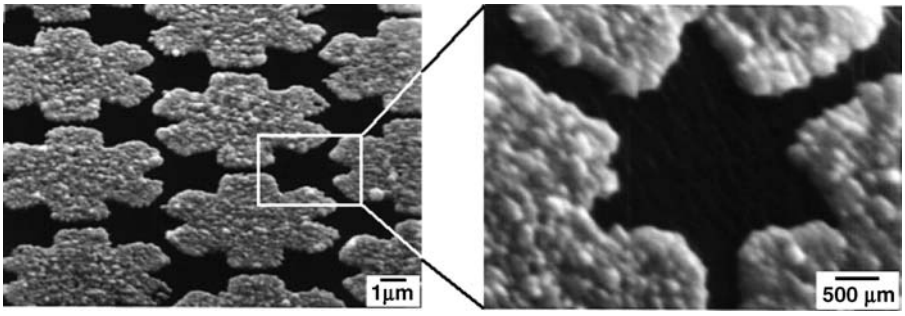


Figure 3.7 Scanning electron microscopy images of copper microstructures grown on poly (styrene) substrates patterned with palladium nanoparticles (NPs). Reproduced with permission from Ref. [46]; © 1996, American Chemical Society.

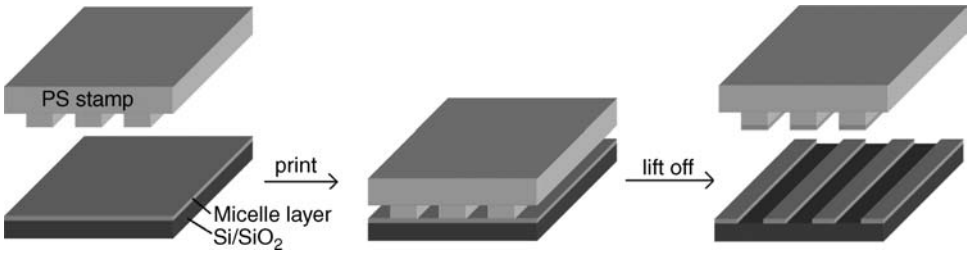


Figure 3.8 The principle of microcontact deprinting. A microstructured poly(styrene) stamp is placed on a monolayer of block copolymer micelles on silicon, and selectively removes the micelles in the contact area at lift-off.

In summary, although μCP was originally developed for the patterning of *n*-alkylthiol SAMs on gold substrates, it represents a straightforward method for the patterning of a wide variety of inks on all types of substrate, subject to an appropriate combination of ink, stamp, and substrate.

3.4 Microcontact Printing and Soft Lithography

Patterned SAMs on solid substrates are important for nanoscience and nanotechnology in two ways:

- They represent a nanostructured material that is easy to prepare and useful for the study of interfacial phenomena that are influenced by nanometer-scale topographies and composition.
- They serve as suitable templates for fabricating microstructures and nanostructures.

Some examples of interfacial phenomena studied with SAMs on thin films include wetting [50–52], corrosion [53, 54], adhesion [52, 55], tribology [56, 57], charge transfer through molecules [58, 59], nucleation and growth of crystals on surfaces [60], and model surfaces for biochemistry and cell biology [61, 62]. These studies depend primarily on the synthesis of SAMs with specific compositions, both in the plane of the surface and out of plane. However, some – such as electron-transfer processes – are extremely sensitive to the nanometer-scale thickness of the SAM. Other applications (such as resistance to etching and protein adsorption, modified electrodes for electrochemistry) rely on the ability of SAMs to prevent the diffusion of other molecules to the surface of the underlying substrate. The application of μCP, and its variations within the area of soft lithography, are detailed in the following subsections.

Hydrophobic SAMs of long-chain *n*-alkylthiols (16 carbons or more) can be used to protect metal films from aqueous wet etching [63]. Moreover, a combination of this ability with μCP makes it possible to fabricate microstructures and nanostructures composed of gold, silver, copper, palladium, platinum, and gold–palladium alloys.

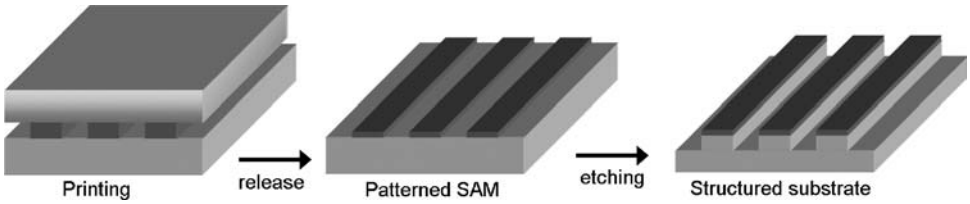


Figure 3.9 Lithography by μCP. A patterned SAM is used as an etch mask for the underlying metal substrate.

Indeed, this was the first application of μCP, and in fact was most likely the intention of its invention [64]. The principle is simple (see Figure 3.9): a substrate patterned by μCP with a SAM of an etch-resistant molecule is treated with an etching solution ("wet etching") or with an etching beam ("dry etching"). The SAM protects the underlying substrate areas from etching, so that the substrate material is removed only in the noncontact areas. This results in the creation of protruding features in the substrate that are similar to those of the stamp, and negative to those of the master that served as the mold for the stamp.

The parameters that determine the minimum dimensions and quality (as measured by the density of pinhole defects on etching and on the edge roughness) of the structures include the composition of the SAM, the density of defects in the SAM, the selectivity of the wet chemical etch, and the morphology of the thin film. A number of etching agents can selectively dissolve regions that are not derivatized with a SAM, and the compositions of these have been developed empirically. The addition of amphiphiles (e.g., *n*-octanol) or use of polymeric complexing agents (e.g., poly(ethyleneimine)) decreases the number of pits and pinholes produced in the surfaces of etched structures, controls the vertical profile of the edges of etched features, and also enables the use of SAMs as resists to pattern thick (>1 μm) electrodeposited films [65]. In the past, the density of pinholes in the SAM and the roughness of the edges of etched features have limited the use of μCP and selective wet etching for fabricating structures with lateral dimensions below 500 nm in gold [66, 67]. However, alternative substrates such as palladium or gold–palladium alloys ($Au_{60}Pd_{40}$) make it possible to generate etched structures that have smaller edge roughness and fewer pinholes than comparable structures in gold when SAMs are used as etch resists. An interphase of PdS that is formed between the bulk metal and the hydrophobic SAM enhances the contrast between the patterned and unpatterned regions [68]. An additional advantage of palladium and gold–palladium alloys as substrates is that they have small grain sizes (~15–30 nm); such morphology is better suited than that of gold (grain sizes ~35–75 nm) for fabricating metal lines with widths as small as 50 nm [69–71]. Unlike gold, palladium is compatible with complementary metal oxide semiconductor (CMOS) manufacturing processes [72]. μCP has also been used to etch Au/Ti layers on GaAs-based materials, and to this end layers of titanium and gold were first evaporated on top of GaAs/AlGaAs quantum well structures and then selectively etched away, using μCP-printed SAMs to protect particular areas of

the surface [73]. Finally, the exposed GaAs could be etched away, transferring the pattern.

In conventional μCP followed by etching, the resultant topology is a positive replica of the stamp, and hence the negative of the master. However, in a newer variation of μCP the resultant topology is the negative of the stamp, and hence the positive of the master. This variation of μCP, which is termed "positive μCP" ((+)μCP) [74, 75], involves pattern replication by printing with a poorly etch-resistant ink, followed by immersion of the sample in a second, etch-resistant adsorbate solution that fills the available areas and acts as a resist in the etching step. An additional advantage of (+)μCP is that stamps with a high filling ratio can be used to replicate master features with a low filling ratio. Originally, pentaerythritol tetrakis(3-mercaptopropionate) (PTMP) was proposed as a positive ink, because it forms a stable SAM on gold and copper, is not replaced by etch-resistant thiols such as ODT, and does not provide significant etch resistance [74, 76, 77]. Whilst (+)μCP complements "standard" μCP, both techniques share similar attributes in terms of optimal contrast and resolution for patterning a metal substrate layer by printing and etching it selectively (Figure 3.10).

μCP can also be exploited to generate patterned metal structures or polymer brushes or nano-objects such as carbon nanotubes (CNTs) and NPs. As mentioned in Section 3.4, the patterned formation or growth of these structures by printing an initiator or a catalyst, followed by polymerization or nucleation reactions or selective deposition to build up the structures, represents an elegant "bottom-up" strategy to obtain nanostructured surfaces.

The ELD of metals onto patterned supporting metal features previously attached to substrates by μCP serves as a straightforward means of producing patterned metal structures on surfaces. The ability of printing to transfer chemical reagents from an elastomeric stamp to a substrate can be used to direct the ELD of copper [45–47]. Electroless deposition is a wet chemical metallization process that involves the reduction of a salt from solution onto a surface, using a reducing agent as the electron source. The presence of a catalyst on this surface is necessary to initiate ELD before the deposition can proceed in an autocatalytic manner (see Figure 3.11). The combination of printing with the ELD of a metal is of both scientific and technological interest.

A bottom-up approach is also followed when producing polymer brushes on substrates. These polymer brushes can, for example, increase the etch resistance of a monolayer on a gold substrate, with the initiator either being printed directly or being backfilled after printing another ink (Figure 3.12). Atom transfer radical polymerization (ATRP) can be used to grow polymer chains from an initiator template [78, 79], while $(BrC(CH_3)_2COO(CH_2)_{10}S)_2$ is often used as a polymerization initiator for the formation of polymer brushes of PMMA and various other poly(methacrylates) [78].

Following a similar patterning strategy, single-walled CNTs have been grown – using a chemical vapor deposition (CVD) technique – from methane and hydrogen on iron nitrate catalyst patterns, or on initiator polymer patterns prepared by μCP on suitable substrates (Figure 3.13) [80, 81]. It has also been shown that improved stamp production techniques can improve the quality of printing the catalyst necessary for NT growth [82]. Another approach enables the patterning of a substrate with isolated

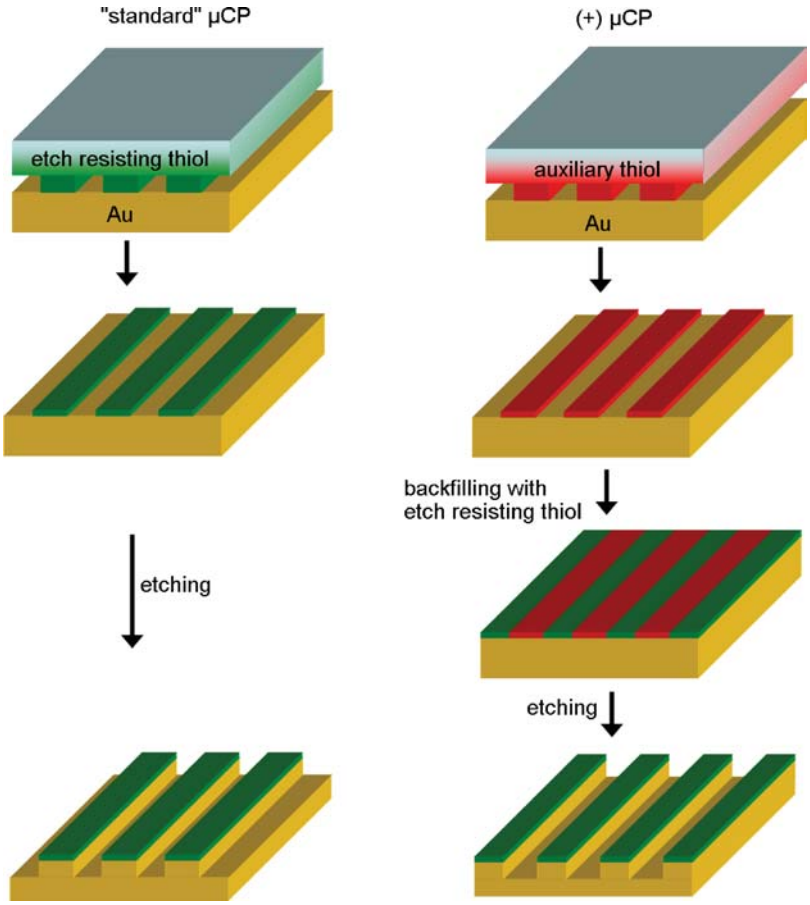


Figure 3.10 Lithography by (+)μCP compared to "standard" μCP. In standard μCP, a SAM is patterned by μCP, while the patterned SAM serves as an etch mask. Etching provides the positive of the stamp (and hence the negative of the master). In (+)μCP, a poorly etch-resistant SAM is patterned by μCP; the patterned SAM is then back-filled with a strongly etch-resistant SAM. Etching provides the negative of the stamp (and hence the positive of the master).

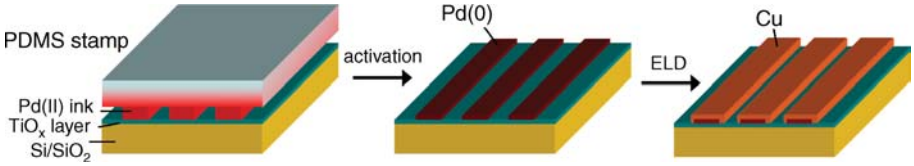


Figure 3.11 Metal structures by μCP. Left: Pd(II) salt is printed onto a TiO_x layer on Si/SiO_2. Center: Pd(II) is reduced to Pd(0). Right: Pd(0) catalyzes the electroless deposition of copper.

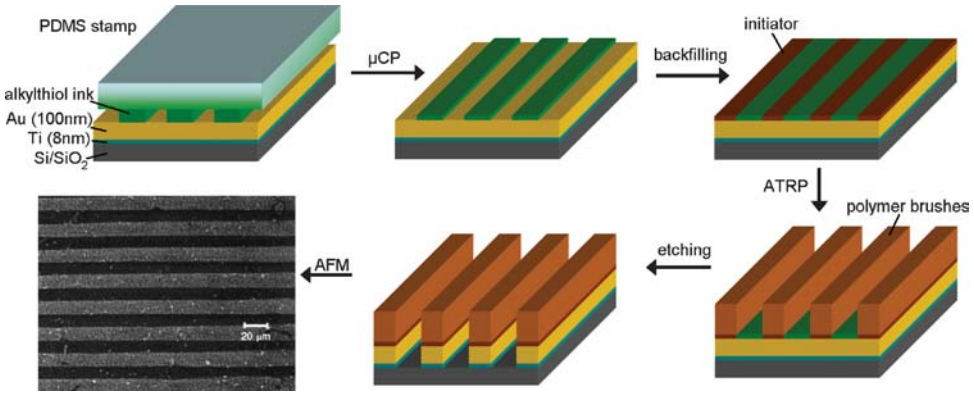


Figure 3.12 Formation of patterned polymer brushes by μCP and atomic force microscopy (AFM) image of patterned gold (dark areas, protected by PMMA) on glass (light areas). AFM image reproduced from Ref. [78]; © 2000, American Chemical Society.

CNTs by using a composite stamp, in this case, the growth of straight CNTs between the patterns was observed, and a method to promote the controlled growth of such isolated nano-objects considered conceivable [83].

Nanotransfer Printing

It is also possible to print a patterned thin film of metal by μCP. The process of transferring a solid nanofilm from a stamp with nanoscale patterned features to a substrate is referred to as "nanotransfer printing" [84–87], in which the stamp can be either a soft or a hard material such as PDMS or silicon. A typical procedure for this μCP technique, using a PDMS stamp, is illustrated in Figure 3.14. In this case, the stamp was coated with a continuous layer of gold (~20 nm thick), without an adhesion layer between the gold and the PDMS. The stamp was then brought into contact with a substrate coated with a dithiol (e.g., 1,8-octanedithiol) [85, 88], after

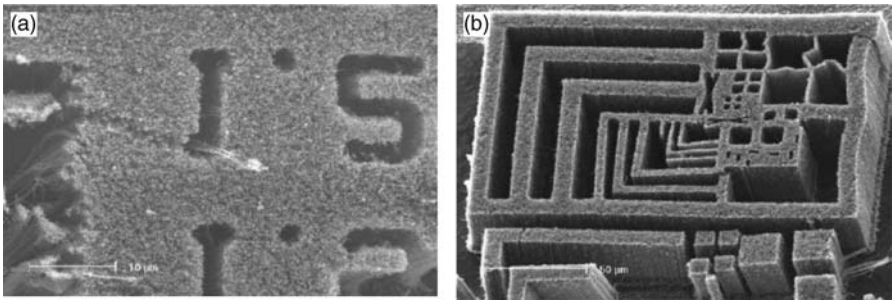


Figure 3.13 Growth of carbon nanotubes after patterning a substrate with iron catalyst with μCP. Image reproduced from Ref. [80]; © 2002, Elsevier.

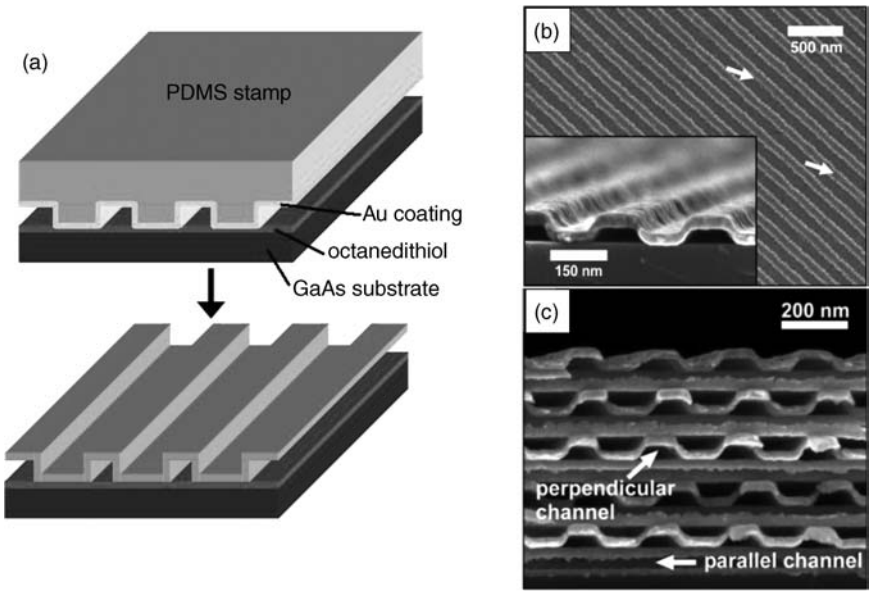


Figure 3.14 (a) The principle of nanotransfer printing (nTP); (b) SEM image of layer of gold on substrate; (c) SEM image of multiple layers after repeating the printing steps. Image reproduced from Ref. [91]; © 2003, American Chemical Society.

which the dithiol formed a SAM on the substrate (GaAs in this case) and the exposed thiol group was bound covalently to the gold layer in the regions of contact. Subsequent removal of the elastomeric stamp from the substrate left the gold layer bound to both the SAM and the underlying substrate. As an alternative, "cold welding" [89, 90] between two metal surfaces could be used to transfer the structured metal film, such that 3-D structures could be fabricated by repeating this procedure [91]. Notably, nanotransfer printing avoids any harsh processing conditions, allows the transfer of nanostructures to be achieved in one combined step [85, 92], and can also be used to pattern features with a lateral resolution of at least 70 nm and an edge roughness down to 10 nm [85, 93]. Another method for releasing the structured film relies on condensation reactions between surface-bound silanols (Si–OH) and/or titanols (Ti–OH) [85, 94]. Techniques that rely on noncovalent interactions between the metal film and the substrate have also been explored, with the minimum dimensions of transferred features currently in excess of 100 nm [95, 96].

As a contact printing technique, nanotransfer printing is well suited to transferring electrodes to fragile surfaces; in particular, it can be used to pattern parallel lines and circular dots as electrical contacts on SAMs [93], with such discontinuous structures adhering to the substrate under "Scotch tape" adhesive tests [92]. The components of devices fabricated directly on plastic substrates include complementary inverter circuits, organic thin-film transistors, capacitors, and electrostatic lenses. This nanotransfer patterning technique can also be used to transfer arrays of sacrificial

etch masks and ferromagnetic stacks of cobalt. The morphology and continuity of the transferred metal structure is important in functional devices, with the uniformity of the metal film depending on the wetting and grain size of the metal on the stamp. Typically, a thin adhesion layer (<2 nm thick) will improve the uniformity of a gold layer on the PDMS stamp and in the transferred layer, although a metal film on an elastomeric stamp may crack due to thermal expansion during metal deposition. Such cracking can be prevented by the rapid deposition of metal and by cooling the stamp [88, 97]. The stress in the metal film from thermal expansion may also be avoided by depositing the metal onto a stamp with a higher thermal conductivity than PDMS (e.g., silicon or gallium arsenide), although the surfaces of these stamps must first be modified with a release layer. Mechanical stress during printing can also introduce cracks into the metal structure.

An alternative approach to printing structured materials is that of "decal transfer printing" [98, 99], whereby a structure (e.g., a PDMS membrane or isolated PDMS features) is transferred from one planar surface to another. In this case, the PDMS decals are made to adhere reversibly to the first substrate (a PDMS slab) [100, 101] while forming covalent bonds with the second substrate. The PDMS slab serves as a handle for patterning continuous or discontinuous features that are otherwise difficult to manipulate. Decal transfer printing can be used to transfer submicrometer features; however, extending the technique to nanoscale features will require further investigations of the interfacial adhesion between the PDMS (or other) substrate and the decal.

3.5
Microcontact Printing and Biological Arrays

During recent years, biological microarrays have rapidly developed into a essential tool for high-throughput genomic and proteomic analysis. Today, "DNA chips" are useful for large-scale parallel analyses of genome sequences and gene expression, for the detection of viruses and other pathogens, for monitoring mRNA expression, and for the classification and evaluation of tumors [102]. Similarly, "protein chips" are valuable for high-throughput diagnostics and drug discovery [103]. By comparison, "carbohydrate chips" have received much less attention to date [104].

Ideally, a biological microarray would have the following properties:

- A high and homogeneous probe density for optimal signal read-out.
- A submicron spot size and nanoscale spot resolution for high data density.
- Many thousands of different probes spotted identically and rapidly for large probe arrays.
- Simple, parallel manufacturing and analysis.

It is evident that state-of-the-art microarray technology falls short of this ideal, with inhomogeneous spots resulting when printing from pins or pipettes due to the evaporation of solvent. In particular, higher probe concentrations may remain at the edges (causing the "coffee-stain effect"), or the probe molecules may aggregate at only

a few points within a spot. Hence, spot sizes are typically in the 50 μm range and are separated by at least 50 μm; moreover, significantly smaller spots can only be produced accurately by using time-consuming SPM-based serial processes. Although, admittedly, soft lithography methods such as μCP are not suitable for patterning multiple probes simultaneously, μCP has become a useful tool for the preparation of biological microarrays.

Biological microarrays can be provided in an "indirect" manner via conventional μCP, using molecular inks that provide a 2-D template for the selective adhesion of biomolecules, as well as of living cells. The earliest examples of this approach originate from the Whitesides group, who printed *n*-alkylthiol patterns on a gold-coated substrate and filled the noncontact area with an oligo(ethyleneglycol)-terminated thiol [105]. Subsequently, extracellular matrix (ECM) proteins (such as fibronectin, collagen and laminin) will be adsorbed onto the hydrophobic area, which in turn causes living cells of various types to adhere preferentially to the ECM-modified areas of the surface. In this way, μCP can be used to create microarrays of cells; moreover, as the cells are generally much larger than the resolution limit of μCP, the size and shape of the cell can be directed by its adhesion to a substrate patterned by μCP [106].

μCP can also be used to provide biological microarrays in a "direct" manner, since many biomolecules such as proteins, lipids or oligonucleotides may serve as suitable inks for μCP. Notably, the rather high molecular weight of biomolecules will enhance the formation of well-defined, high-contrast patterns as their diffusion is limited. The transfer of biomolecules from the stamp to the substrate by μCP depends on the surfaces properties of the stamp and substrate. Whilst the simplest μCP approach for patterning of biomolecules involves the direct transfer of ink molecules adsorbed onto the stamp to a target substrate by conformal contact, several important factors must be considered in this respect. Notably, the affinity of the biomolecule towards the stamp and substrate must be tailored such that it is higher for the latter than for the former. In addition, as μCP should not cause denaturation it is preferable that the use of hydrophobic stamps and substrates is avoided in the μCP of proteins. Finally, the biomolecule should, ideally, be printed in such a way that all of the active sites are exposed to the target molecules.

The first reports of the μCP of proteins were made in 1998 [29], when the process was deemed to be very straightforward, and involved: (i) the inking of a PDMS stamp with an aqueous protein solution; (ii) a period of incubation; (iii) air-drying of the stamp; and (iv) bringing the stamp into conformal contact with the substrate (Figure 3.15) [29, 107]. Tan and coworkers have demonstrated that both stamp and substrate wettability is crucial for biomolecule transfer [108]; indeed, a minimum wettability of the substrate was seen to be required for the successful μCP of proteins, but this would be lessened if the wettability of the stamp were to be reduced. Tan and coworkers also found the mechanism for the μCP of protein to differ from protein adsorption because: (i) those surfaces that are resistant to protein adsorption in an aqueous environment are susceptible to μCP under ambient conditions; and (ii) the amount of immobilized proteins and the wettability of the substrate varied gradually for adsorption, but displayed a threshold wettability for μCP.

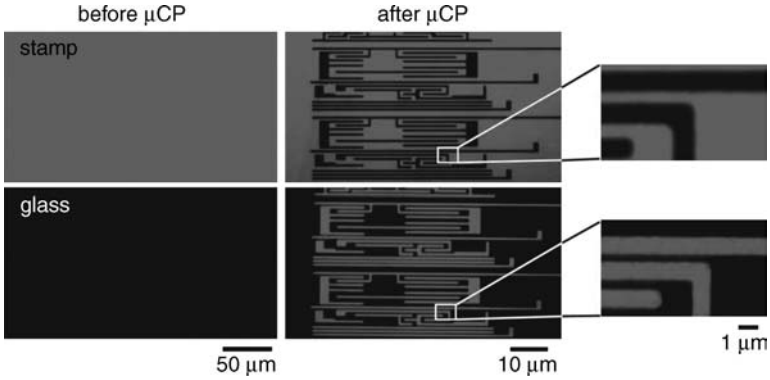


Figure 3.15 μCP of proteins. An IgG protein solution is incubated on the top of an elastomeric stamp. After drying, the stamp is brought into conformal contact with the glass substrate and transfer of proteins occurs only in the area of contact between the stamp and the substrate. Image reproduced from Ref. [29]; © 1998, American Chemical Society.

The patterning of proteins by μCP was demonstrated on different types of substrate, including glass, metal oxides, metals, and polymers. The concept of a direct μCP of protein onto a glass substrate was further extended to the fabrication of *single protein arrays* such as antibodies (e.g., immunoglobulin G; IgG) and green fluorescent proteins (GFPs) on glass [109]. ECM proteins such as laminins have also been patterned by μCP on silicon wafers to guide the growth of neurons for bioelectronic purposes [110]. It should be noted that, whilst many proteins are chemisorbed to gold substrates, the chemisorption often involves a reduction of the disulfides in the protein, leading in turn to denaturation. Gold-binding polypeptide (GBP) represents an interesting example of a protein that can be applied to direct μCP on gold surfaces [111]. Notably, GBP does not contain any cysteine residues that are known generally to form a covalent bond with gold; hence, the binding of GBP is independent of thiols, and offers a new means of interaction between the biomolecule and the surface. Likewise, a GBP–GFP–His_6 fusion protein could be printed directly onto a gold surface in a mixture of bovine serum albumin (BSA) and surfactant, such that the protein pattern could be applied as a template for the high-throughput assays of both protein–protein and DNA–DNA interactions [111]. Kwak and coworkers patterned cytochrome C onto gold surfaces using a nonmodified PDMS stamp [112]; in this case, the cytochrome C was used as an ink, while the protein arrays were transferred directly from the stamp to a SAM of mercaptohexanoic acid (MHA) on gold. Active enzymes were also successfully patterned using SAMs on gold surfaces [113]; for example, the metalloprotein azurin was printed on a glass substrate that had been modified with mercaptosilane and which allowed site-specific binding of the protein. The pattern obtained was investigated, using immunofluorescence, with anti-azurin serum [114].

Poly(lysine) was microcontact-printed onto a clean, nonmodified glass surface via an oxidized PDMS stamp by using electrostatic interactions between the positively charged polypeptide and the negatively charged glass surface [115]. Delamarche and

coworkers proposed the use of hydrophilic PDMS stamps modified with poly (ethylene oxide) silanes [27, 116]. In this case, the modification was conducted by oxidation of the PDMS stamp and reaction with 3-aminopropyltriethoxysilane (APTES), followed by a reaction to conjugate the surface amino groups with poly (ethylene glycol) (PEG) chains. When PEG is grafted onto oxidized PDMS stamps it acts as a protein repellent layer, and this property was utilized to design a flat stamp with regions that could attract proteins (nonmodified PDMS) and regions modified with PEG that have protein-repellent properties [117]. The local modification of PDMS was conducted by oxidation in O_2-plasma with the application of metal mask (those areas covered by the mask were neither oxidized nor modified). When proteins are applied to such a stamp, they are directed towards its hydrophobic areas; hence, protein IgG was successfully transferred to the glass substrates and immobilized in a well-defined pattern with high accuracy and contrast. In a different approach, when the PEG-modified stamp (according to the procedure described above) was contacted with another flat, dry, nonmodified PDMS stamp (the "ink pad") that had been incubated in IgG buffer solution, a homogeneous layer of proteins was transferred to the PEG regions of the other stamp. The latter could then be contacted with a glass substrate and used to pattern IgG proteins.

Proteins may also be patterned on a flat stamp by using a microfluidic network [30], such that the patterned flat stamp can be contacted with glass or another substrate so as to transfer the pattern. Recently, this approach was extended to the high-resolution μCP of proteins, whereby a flat PDMS stamp was patterned with a nanoscale PEG pattern by using dip-pen nanolithography (DPN). The nanopattern could then be replicated as a protein nanopattern on a glass substrate [118].

Another means of overcoming the problems of PDMS stamps with regards to wettability and compatibility with aqueous solutions, would be to select alternative materials for fabrication of the stamp. One versatile approach here would be the use of agarose hydrogel stamps as a mold for transferring water-soluble biomolecules [40]. Since an agarose hydrogel stamp is highly permeable to water, it would also function as an ink reservoir; consequently, multiple stamping would be possible, without any need for intermediate re-inking of the stamp.

μCP of DNA

In the μCP of DNA, it is necessary to modify the stamp surface to ensure DNA–stamp attraction, and such modifications can be carried out by the addition of APTES, which confers a positive charge to the surface [31]. Perhaps the greatest benefit for the μCP of DNA is its ability to print multiple arrays from a single loaded stamp; indeed, this could ultimately result in both cost-saving and time-saving processes, especially for gene expression studies when it is the *ratio* of bound to labeled molecules that is important, and *not* the total amount of material present [31]. Subsequently, a much more efficient method of transferring the micropatterns of DNA and RNA to a surface was achieved by modifying the PDMS stamp with positively charged PPI dendrimers (creating the "dendri stamp") in a "layer-by-layer" arrangement (Figure 3.16) [119, 120]. The electrostatic interactions between dendrimers and

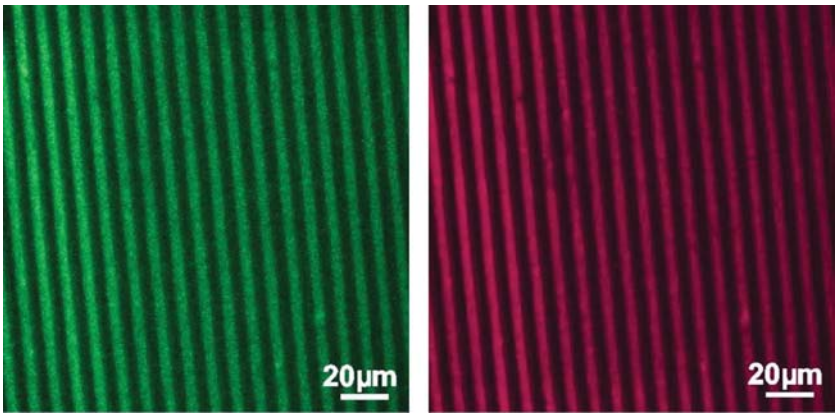


Figure 3.16 Transfer printing of DNA and RNA using "dendri-stamps." A PDMS stamp is oxidized and coated with a cationic PPI dendrimer. DNA and RNA bind to the dendrimer-coated stamp in a layer-by-layer arrangement. If the DNA (or RNA) is functionalized with an amine, it can be printed on an aldehyde-terminated self-assembled monolayer. The image shows a simultaneous fluorescence micrograph of DNA patterns after hybridization between a fluorescein-labeled probe that was obtained by μCP (left image) and its complementary Cy5-labeled target (right image). Image reproduced from Ref. [119]; © 2007, American Chemical Society.

oligonucleotides ensured a successful transfer of DNA or RNA to the target surface, while imine chemistry [119] or "click" chemistry [120] could be applied to bind the covalently modified DNA and RNA molecules to a chemically functionalized substrate (see Section 3.6). An alternative approach to patterning DNA molecules on the surface was proposed by Xu and coworkers [121], who prepared an "amphiphilic" DNA by attaching a hydrophobic alkyl chain to the 3′ or 5′ end, such that the hydrophobic tail enhanced the adsorption of DNA to the hydrophobic PDMS stamp. This in turn allowed for the efficient transfer and delivery of DNA to the surface.

It is also of interest to pattern phospholipids by μCP (Figure 3.17) [122]. Supported lipid bilayers are very fragile assemblies that are formed by lipids organized into two opposing leaflets on hydrophilic surfaces, such as glass or mica substrates, and can be patterned onto solid substrates. For phospholipid patterning, however, the μCP technique used differs slightly from that used with proteins or DNA. First, the bilayer must be formed on the oxidized PDMS stamp from the buffer solution by fusion of liposomes to the stamp surface. Second, the printing should be carried out in water, otherwise the bilayer will lose its structure. This method allows an efficient and reliable transfer of membrane patches to glass surfaces, which are of particular interest in investigations of biological membranes in general, and in the behavior of membrane proteins in particular. Alternatively, lipid bilayers can be patterned indirectly by μCP: in this case, either a template of proteins can be printed to which the lipid bilayer vesicles are fused, or a supported lipid bilayer is selectively removed in the contact areas by blotting through μCP with a bare stamp [123].

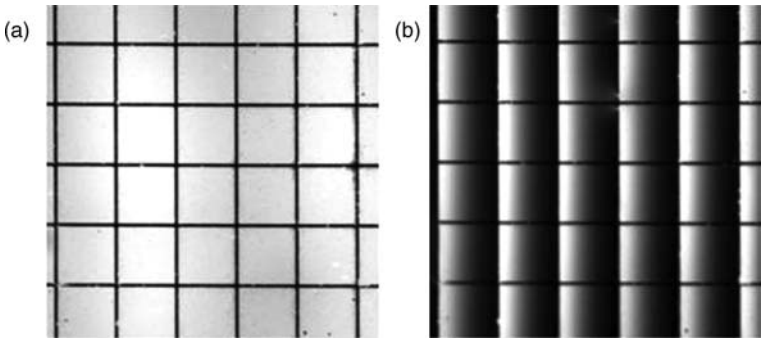


Figure 3.17 Patterning of phospholipids bilayers by μCP. (a) Fluorescence image of a patterned supported lipid bilayer that was printed onto a glass surface (egg phosphatidylcholine with 1 mol% Texas Red). The bright regions are fluorescence from the labeled lipids, and the dark grid pattern is the bare glass surface; (b) Fluorescence image taken after an electric field was briefly applied parallel to the bilayer plane, creating a steady-state gradient of the negatively charged labeled lipids and demonstrating long-range mobility. The dimensions of both images were 560 μm × 560 μm. Image reproduced from Ref. [122]; © 2001, American Chemical Society.

Affinity Contact Printing

Recently, an interesting concept for the simultaneous μCP of multiple probes has been proposed. "Affinity contact printing" (αCP) [124] relies on inking the surface of a PDMS stamp with antigens as "capture molecules" which allows the subsequent binding of selected antibodies from a solution containing mixtures of proteins (Figure 3.18). Affinity stamps were prepared by modification of the PDMS stamp with APTES and a crosslinker to produce an activated, hydrophilic surface. This activated stamp was then used to couple antigens to selected areas using: (i) microwells; (ii) microfluidic networks; and (iii) μCP. By repeating this procedure with a different type of antigen, the stamp could be functionalized with a pattern of various antigens, which would be valuable for microarray applications. When several types of antigen are immobilized on an activated stamp they can be exposed to a solution of different antibodies, so as to extract and immobilize a "matching partner." The captured antibodies can then be printed onto a glass substrate to form microarrays of antibodies.

An alternative form of affinity contact printing was demonstrated by Jang and coworkers [125], which relied on the modification of a PDMS stamp with aminosilanes and succinic anhydride to introduce carboxylic acid groups on the surface; this was followed by the immobilization of a monoclonal antibody (mAb) to the epidermal growth factor receptor (EGFR). The EGFR-antibody-modified stamp was then incubated with a solution of membrane proteins from cell membrane extracts and crude cell lysates. The stamp was contacted with a gold substrate that had been modified with an amino-terminated monolayer and, after μCP, the substrate was covered with a nematic liquid crystal (LC) film. The orientation of the LC film was

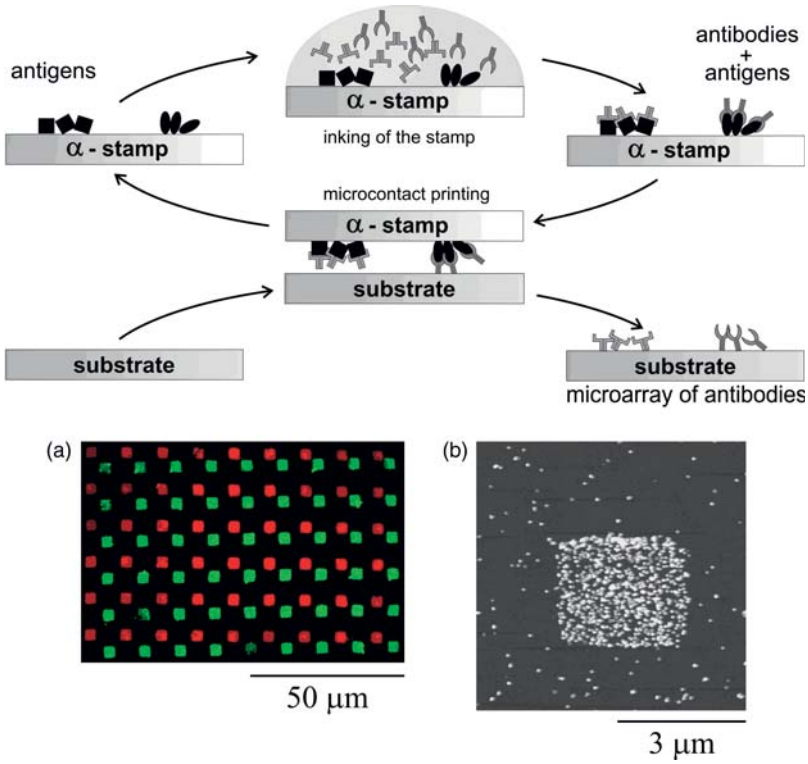


Figure 3.18 Upper panel: Affinity contact printing (αCP) relies on inking the surface of a PDMS stamp with antigens as "capture molecules"; this allows the subsequent binding of selected antibodies from a solution containing mixtures of proteins. Lower panel: (a) Fluorescence microscopy image showing the placement of the TRITC-anti-chicken and FITC-anti-goat antibodies from a stamp onto a glass substrate; (b) AFM image obtained on a spot of the array in which the printed anti-goat antibodies bound to Au-labeled goat antigens presented in solution. Detection of this binding was monitored by staining the Au labels with electroless-deposited silver particles of average diameter 80 nm. Images (a) and (b) reproduced from Ref. [124]; © 2002, Wiley-VCH.

found to be different on the amino-terminated surface and the regions of the surface presenting EGFR, thus providing a simple, label-free method for optically detecting the presence of EGFRs on the surface. In a similar manner, the group of Abbott used αCP to immobilize proteins that subsequently can be imaged with LCs [126]. This method relies on the covalent modification of PDMS stamp with a biotinylated BSA. In this case, the BSA-functionalized stamp was inked with anti-biotin IgG and brought into conformal contact with an amino-modified gold surface. After printing, the protein pattern was imaged by spreading an LC film onto the surface.

An interesting concept of pattern transfer of DNA using μCP was introduced by the groups of Crooks [127] and Stellaci [128]. These strategies relied on the fabrication of DNA arrays onto flat solid supports by immobilization via the 5′ end functionalized with, for example, amine linkers. The DNA array was further hybridized with

complementary strands that possessed a capturing group at the 5′ end. These groups could be reacted with functionalized surfaces simply by bringing the surface into conformal contact with the DNA array. After reaction, the two surfaces could be separated and the pattern of single-stranded DNA transferred to the surface that was in contact with the patterned hybridized array. As a result, the strands would be mechanically separated and the new arrays of patterned DNA could be used for the next transfer of microarrays (Figure 3.19).

In conclusion, μCP has emerged as a versatile tool for the preparation of biological microarrays. Interesting methods for the simultaneous μCP of multiple inks have recently been reported. The highly selective molecular recognition of biomolecules may also be used in the replication of arrays by μCP.

3.6
Microcontact Printing and Surface Chemistry

One striking feature of μCP is the short contact time required to form a dense monolayer of ink on the substrate. Although, typically, the contact times are approximately 1 min, μCP has also been performed with millisecond contacts of the stamp and substrate [129]. In contrast, it takes several hours to prepare an *n*-alkylthiol or *n*-alkylsilane SAM from solution [130]. These observations indicate that μCP is a particularly effective method for preparing SAMs, even if this

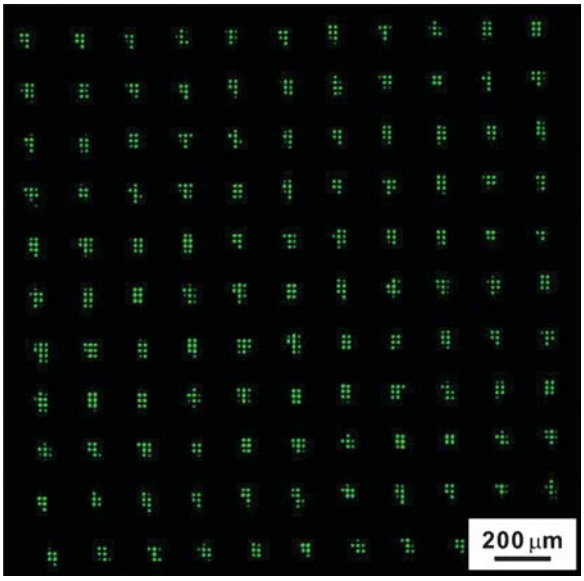


Figure 3.19 Fluorescence micrograph of an RNA microarray on a PDMS surface, which was fabricated using a master DNA array of 2500 spots (~70 μm in diameter). Image reproduced from Ref. [127]; © 2007, American Chemical Society.

involves a slow condensation reaction (as in the case of *n*-alkylsilane SAMs). This conclusion raises the interesting point of whether μCP could be used to accelerate surface reactions, for example by printing a molecular ink on top of a SAM [131]. It is known that reactions on SAMs are typically several orders of magnitude slower than reactions in solution [132]. It can be argued that the steric hindrance and conformational restraints encountered at the surface of a SAM (or any other surface) reduce the frequency of effective intermolecular collisions, and hence enhance the activation energy of reaction. Yet, it is likely that this kinetic barrier is more than compensated when a stamp saturated or densely covered with ink is brought into conformal contact with a SAM in which most of the reactive groups are exposed at the surface. In this case, a bimolecular reaction should benefit from the nanoscale confinement of highly concentrated reagents in the contact area between a stamp and a substrate.

Indeed, it has been demonstrated several times by Whitesides and colleagues that amides are formed when amines – small molecules as well as polymers – are printed on an anhydride-terminated or active ester-terminated SAM on gold [133]. However, it must be noted that this result is not surprising given the reactivity of amines towards anhydrides and active esters.

In 2004, Huck and coworkers described the formation of peptides by printing N-protected amino acids onto an amine-terminated SAM on gold (Figure 3.20) [133]. Of course, peptide bonds do not spontaneously form from carboxylic acids and amines under ambient conditions, and it was proposed by Huck that "...the nanoscale confinement of the ink at the interface between the stamp and the SAM, in combination with the pre-organization of the reactants in the SAM, facilitates the formation of covalent bonds" [134]. In a remarkable experiment, it was shown that the consecutive μCP of as many as 20 peptide nucleobases resulted in the formation of an oligopeptide nucleic acid that could selectively bind a complementary strand of DNA. These findings point to the fascinating potential of surface chemistry by μCP, that complex biomacromolecules could be synthesized simply by printing the monomers in the appropriate sequence!

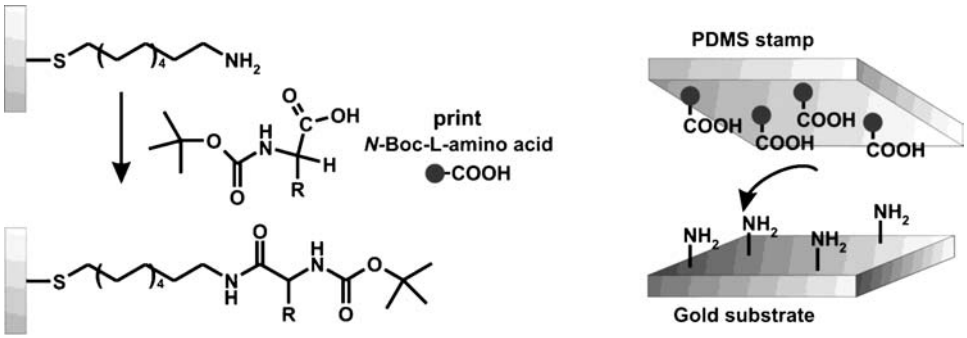


Figure 3.20 Peptide synthesis by μCP. An oxidized PDMS stamp is inked with an *N*-Boc-*L*-amino acid and pressed into a contact against an amino-functionalized gold substrate to yield a covalent peptide bond. Boc = *t*-butoxycarbonyl.

Subsequently, it was also shown that imines can be formed in a few minutes under ambient conditions by printing amines onto aldehyde-terminated SAMs [135]. The reaction of aldehydes and amines is an equilibrium reaction that is generally unfavorable unless water is removed from the reaction mixture. In addition, this reaction was applied to the preparation of biological microarrays: the μCP of RGD-containing proteins on aldehyde-terminated SAMs was used to direct the adhesion of cells in microarrays [136], while the μCP of amine-modified DNA on aldehyde-terminated SAMs provided DNA microarrays [119].

These reports point to a second advantage of surface chemistry by μCP, namely that biological arrays are generally prepared on transparent substrates (preferably glass) so that they can be read out with fluorescence; however, as biomolecules are not compatible with alkoxysilanes, an indirect immobilization and patterning method would be required for glass substrates. Microcontact chemistry on an intermediate SAM fulfills this requirement.

The Huisgen 1,3-dipolar cycloaddition of alkynes and azides can also be induced by μCP (Figure 3.21) [137]. The Cu(I)-catalyzed cycloaddition of alkynes and azides [138] is a prime example of "click chemistry" (i.e., a chemical reaction with near-quantitative yield, mild reaction conditions, and short reaction time) [139] that has found widespread use for the bio-orthogonal ligation of biomolecules to surfaces; moreover, its combination with μCP constitutes an attractive method for the preparation of microarrays. Triazoles are formed within minutes when an alkyne is printed on an azide-terminated SAM on a silicon wafer or glass, even in the absence of a Cu(I) catalyst that is normally used to accelerate this type of "click chemistry" [137]. It should be emphasized that the solution reaction in the absence of Cu(I) is slow unless electron-poor alkynes are used. The Huisgen cycloaddition induced by μCP was investigated in detail by printing a set of fluorescent alkynes on azide-terminated SAMs on glass substrates (J. Mehlich and B.J. Ravoo, unpublished results). When fluorescence microscopy was then used to monitor the extent

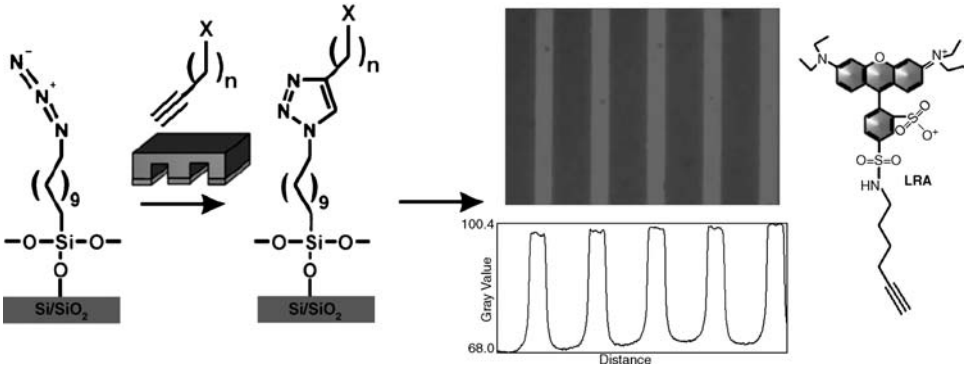


Figure 3.21 1,3-Dipolar cycloaddition reaction by μCP. Triazoles are formed within minutes when an alkyne is printed on an azide-terminated SAM on a silicon wafer or glass, even in the absence of a Cu(I) catalyst that is normally used to accelerate this type of "click chemistry."

of reaction on the glass surface, it was shown that the rate of cycloaddition depended on the reactivity of the alkyne and on the presence of Cu(I). Although the cycloaddition would be accelerated by Cu(I), it also proceeded readily in the absence of Cu(I).

"Click chemistry" by μCP was applied to print microarrays of alkyne-modified DNA [120] and alkyne-modified carbohydrates [140] on azide-terminated SAMs on Si wafers and glass. It was observed that, although DNA could be printed without a Cu(I) catalyst, the surface density of carbohydrates was low in the absence of a Cu(I) catalyst. In particular, for the preparation of biological microarrays it is advantageous to exclude the toxic Cu(I) catalyst.

μCP can also be used for the heterogeneous catalysis of chemical reactions in the contact area between stamp and substrate. In its most simple form, an oxidized PDMS stamp can be used as a heterogeneous acid catalyst to accelerate a hydrolysis reaction on a SAM. It has been shown that silylether-terminated SAMs are hydrolyzed to produce hydroxyl-terminated SAMs upon contact with an oxidized PDMS stamp for 5–10 min [141]. It was also shown that FMOC protecting groups can be removed from an amine-terminated SAM by contact with a piperidine-modified poly (urethane) stamp [142]. In this way, catalytic μCP can be used to replicate the microstructure of the stamp in the form of a chemical contrast on the substrate, without any ink transfer ("printing without ink"). In a more sophisticated approach to catalytic μCP, Toone and coworkers have shown that enzymes immobilized in a poly (acrylamide) stamp can induce the cleavage of surface-immobilized DNA in the area of contact between stamp and substrate [143].

Very recently, the heterogeneous catalysis of the Huisgen cycloaddition of alkynes and azides by μCP was reported. In this case, a microstructured PDMS stamp covered with a thin film of Cu (which had been air-oxidized to Cu_2O) was used to induce the cycloaddition of alkynes on an azide-terminated SAM on gold [144]. It was shown that the cycloaddition by μCP would proceed to completion (i.e., until all reactive sites on the surface were occupied) within a few hours if a Cu-coated stamp was used.

Finally, it must be emphasized that spatially controlled surface chemistry induced by μCP is not limited to reactions on SAMs on inorganic substrates. In particular, transparent polymer films represent attractive substrates for reactions induced by μCP. An early example of microcontact chemistry on polymer films was that described by Chilkoti and coworkers, who oxidized poly(olefin) and poly(ester) substrates, activated them with pentafluorophenyl esters, and subsequently patterned the polymer surface with amine-terminated biotin, using μCP [145]. In similar fashion, amine-modified proteins and DNA can be printed on *N*-hydroxysuccinimide-activated poly(methacrylate) films [146] (of course, these results should be expected, given the inherent reactivity of amines and active esters). Recently, the Cu (I)-catalyzed cycloaddition of alkynes and azides was also used to pattern the surface of a poly(alkyne) film by the μCP of azide-terminated biotin [147]. To this end, the Cu (I) catalyst was printed on a poly(alkyne) film covered with a thin layer of azide-terminated biotin. It is to be expected that less-reactive polymer films could also be functionalized with spatially patterned molecular monolayers.

Table 3.1 Surface chemical reactions induced by microcontact printing (μCP).

Substrate	Ink	Product	Catalyst	Reference(s)
Anhydride	Amine	Amide	—	[133]
Active ester	Amine	Amide	—	[145, 146]
Amine	Carboxylic acid	Amide	—	[134]
Aldehyde	Amine	Imine	—	[119, 135, 136]
Azide	Alkyne	Triazole	—	[120, 137]
Azide	Alkyne	Triazole	Cu(I)	[140]
Alkyne	Azide	Triazole	Cu(I)	[147]
Azide	Alkyne	Triazole	Cu stamp	[144]
Si-protected alcohol	—	Alcohol	Ox stamp	[141]
FMOC-protected amine	—	Amine	Pip stamp	[142]

In many respects, reactions induced by μCP follow the principles of click chemistry: near-quantitative yield (i.e., complete surface coverage); mild reaction conditions; and short reaction times [139]. Most interestingly, however, the scope of the reactions – including condensation, cycloaddition, nucleophilic substitution, and deprotection – continues to expand, and an overview of those induced by μCP to date is provided in Table 3.1.

Notably, reactions carried out using μCP can be combined with heterogeneous catalysis and applied to functionalize polymer films with molecular monolayers. The confinement of heterogeneous catalysts on a microstructured stamp also opens up the possibility to react highly volatile reagents with the substrate, which is not possible in conventional μCP. The limited resolution of μCP could possibly be overcome by using flat stamps with nanostructured heterogeneous catalysts (see Section 3.7). In summary, it can be foreseen that surface reactions induced by μCP will provide a straightforward and versatile method for surface chemistry in general, and for the fabrication of (bio)molecular microarrays and nanoarrays in particular.

3.7 From Micro to Nano: Increasing the Resolution of Microcontact Printing

As the name indicates, μCP is typically used for the structuring of surfaces at the microscale. A number of factors limit the resolution of conventional μCP to about 0.5 μm. The first major limitation is inherent to the flexible nature of the elastomer PDMS stamp. PDMS has a Young's modulus of about 1.5 MPa, which is soft enough to ensure a conformal contact with substrates so as to facilitate ink transfer. Unfortunately, however, a PDMS stamp can be easily deformed, and this imposes a limit on the aspect ratio (i.e., the height of a microstructure divided by its width) as well as the fill ratio (i.e., the width of a structure divided by the distance) of the stamp (see Figure 3.3). If the aspect ratio is too high (i.e., tall structures close together), then the microstructures will buckle and stick together. For conventional μCP with

n-alkylthiol inks on gold, an aspect ratio of about 1 is considered optimal. However, if the fill ratio is too low (i.e., small structures far apart) then the stamp will sag and touch the substrate also in the noncontact area. A second major limitation to the resolution of μCP resides in the ink transfer from stamp to substrate: although, ideally, the ink should be transferred exclusively in the contact areas, the ink in fact tends to diffuse and spread to the noncontact areas during printing. In particular, when printing small features with low-molecular-weight inks, diffusion and spreading of the ink outside the contact area will adversely affect the edge resolution of μCP. The strategies proposed to extend the resolution of μCP into the nanoscale will be outlined in the following subsections.

One obvious approach would be to use "stiffer" stamps for μCP, so that deformations would occur less readily. For example, PDMS can be crosslinked more extensively, so that its Young's modulus would increase to about 10 MPa [11, 15]. Although such "hard" PDMS stamps would still be soft enough for conformal contact, their relief structure would less readily deform. Structures as small as 80 nm can be accurately replicated using hard PDMS stamps. Alternatively, stamps can be prepared from acryloxy perfluoropoly(ethers), which have a Young's modulus of about 10 MPa [148], or from poly(urethane acrylates), which have a modulus of about 20 MPa [149], or from poly(olefins), which can have a value of more than 40 MPa [18]. It has been shown that such "rigid" stamps can be used for the μCP of proteins in lines of 100 nm width with 3 μm periodicity (Figure 3.22) [150]. Deformation of the stamp can also be reduced by using a PDMS stamp on a rigid support [115, 151] that would prevent the stamp from sagging in the noncontact area, such that a substantially lower fill ratio would be possible. Another useful improvement is that of "submerged" μCP, where the μCP of *n*-alkylthiols is performed in water instead of

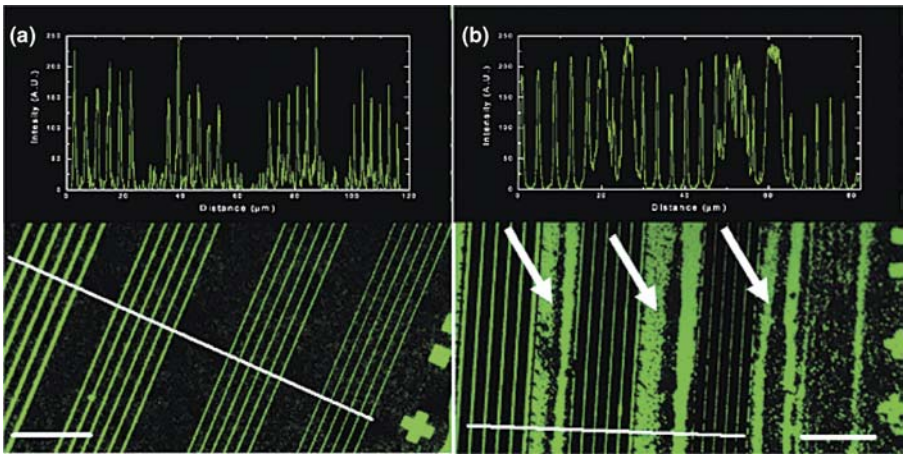


Figure 3.22 High-resolution μCP of proteins in lines of 100 nm width with 3 μm periodicity. μCP with poly(olefin) stamps (a) is clearly superior to μCP with conventional PDMS stamps (b). Image reproduced from Ref. [150]; © 2003, American Chemical Society.

air [152]. Here, the role of the water is to support the relief structure of the stamp, so that stamps with aspect ratios of 15 : 1 and higher can be used.

The diffusion and spreading of ink molecules into the noncontact areas can, of course, be limited by reducing the contact time of stamp and substrate. To this end, Wolf *et al*. have proposed "high-speed μCP" [129], which allows *n*-alkylthiol inks to be printed on gold substrates within 10 ms, while the resultant SAMs in the contact areas serve as effective etch resists, allowing the noncontact metal areas to be selectively etched. In fact, it has been shown that ultrafast μCP can be readily used for μCP at the submicron scale. As an alternative, the diffusion and spreading of ink into the noncontact area can be limited by inking the stamp through an ink pad, such that the stamp is inked only in the contact area [153].

Another obvious improvement would be to use inks that had a low diffusion coefficient and a low tendency to spread across the substrate. The diffusion rate of an ink correlates with its molecular weight; thus, a higher molecular weight will limit the diffusion of an ink into noncontact areas. To this end, the use of "heavy inks" has been proposed as an alternative to the simple *n*-alkylthiols, in particular for applications in high-resolution μCP for lithography. An early example of this so-called "nanocontact printing" was described by Huck *et al*., who used a dendrimer ink and a submicron-structured PDMS stamp to print 140 nm-wide lines with 70 nm periodicity [8]. Along the same lines, others have designed "multivalent inks" that have multiple functional groups capable of binding to the substrate [154]; the enhanced surface adsorption of multivalent inks reduces the spreading of the ink into the noncontact areas of the substrate. Dendrimers have also been particularly useful in this area since, by using dendrimer inks with multiple thioether end groups, it is possible to perform (+)μCP followed by a wet etch with a resolution better than 100 nm (Figure 3.23) [155]. As most biomolecules have a much higher molecular weight than *n*-alkylthiols, their diffusion will be negligible, such that they will be particularly suited for μCP at the submicron scale.

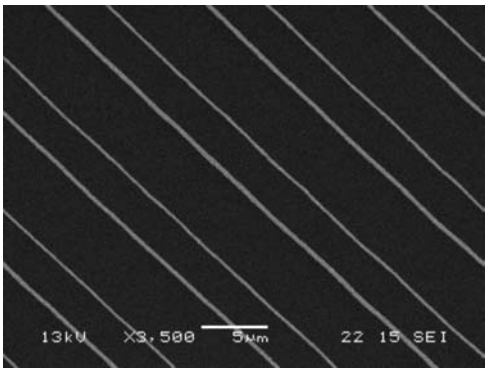


Figure 3.23 High resolution (+)μCP with dendrimer inks; 100 nm gold lines were prepared by (+)μCP of thioether dendrimers for 2 min, dipping in ODT for 6 s, and etching in Fe(III)/thiourea for 2.5 min.

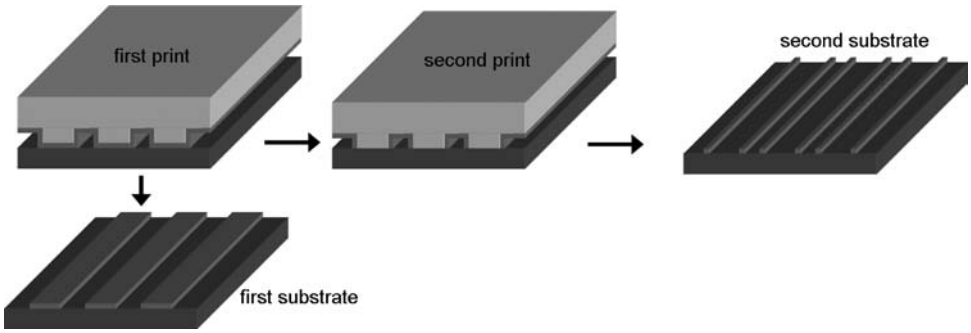


Figure 3.24 The principle of edge transfer printing.

An alternative approach to improve the resolution of μCP is that of *edge transfer lithography* (Figure 3.24) [156–158], in which the edges of micrometer-sized stamps are used to reproduce submicrometer structures. After removing the top layer of the ink that is required to be transferred from the protruding features of a stamp, the stamp is brought in conformal contact with another substrate, whereupon any residual ink at the edges of the features is transferred to the substrate.

In the case of a conventional PDMS stamp with a microrelief surface structure, the selective transfer of ink occurs essentially due to a rapid transfer in the contact area and a slow (ideally, negligible) transfer in the noncontact area. It could be argued that the air-filled voids between the microstructures at the stamp surface pose a diffusion barrier for the ink, since volatile inks can easily cross this barrier but nonvolatile inks cannot. A major advance in the resolution of μCP involves a radically altered design of stamps; instead of exploiting the voids in the microrelief pattern as a diffusion barrier, it is possible to impose a diffusion barrier on a *flat* PDMS stamp [159]. For example, by oxidation of the PDMS surface, a thin silicon oxide film is created, which is essentially impermeable to apolar inks. If the oxidation is directed by a mask, then a flat stamp with a surface pattern of silicon oxide on PDMS will result that can be used for the μCP of alkylthiol, which are transferred exclusively in the nonoxidized area. The properties of the diffusion barrier and stability of the stamp may be improved by coating the silicon oxide film with a fluorinated silane SAM, and even volatile, low-molecular-weight inks can be printed with such chemically patterned flat stamps. Moreover, because the stamp is flat, all problems due to deformation of the microrelief surface structures are circumvented. Hence, the resolution of the stamp is now limited only by the resolution of the oxidation mask, which can be made by conventional photolithography and/or electron beam lithography (EBL). Recently, the resolution of μCP with flat stamps was further improved by using DPN to "write" a nanostructured oxidation mask on the surface of a flat PDMS stamp [118]. Following oxidation, the stamp can be functionalized with hydrophilic and fluorinated silanes. It was demonstrated that the chemical nanopattern on the stamp can be replicated on a gold substrate in the form of a nanopatterned alkylthiol SAM, which can serve as a nanoscale etch mask. Among others, the patterns included a nanoscale map of the USA! Regular arrays of gold dots with a diameter of 80 nm spaced by

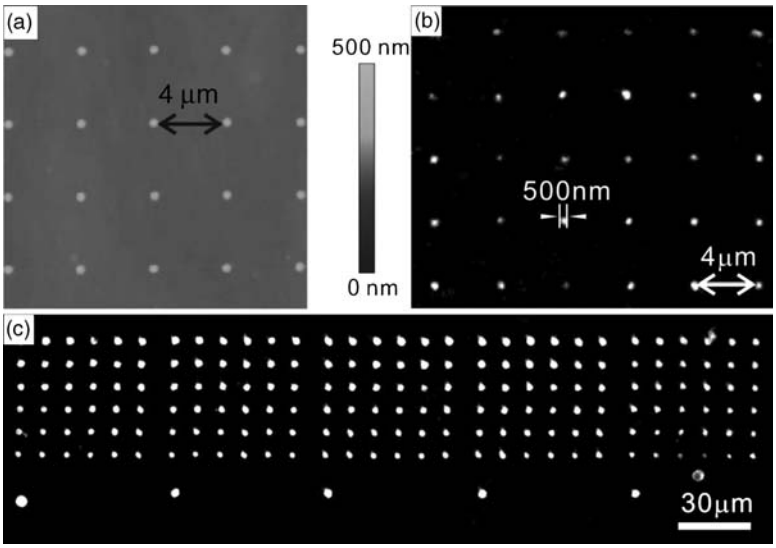


Figure 3.25 High-resolution μCP with a flat PDMS stamp that was nanostructured by DPN. (a) An AFM topography of the PEG pattern used for fabricating a flat stamp for μCP of proteins; (b) A fluorescence image of the printed TRITC-conjugated IgG; (c) A fluorescent image of the printed TRITC pattern on glass. Image reproduced from Ref. [118]; © 2008, Wiley-VCH.

10 μm can be readily obtained using this method. Alternatively, the nanopatterned stamp can be used to print proteins in a nanoscale array on glass (Figure 3.25).

In summary, the resolution of μCP may be substantially better than 100 nm. In particular, the design of chemically patterned, flat elastomer stamps has radically improved the resolution of μCP. Although the nanostructuring of surfaces by μCP is less straightforward than conventional microstructuring by μCP, "nanocontact printing" represents an attractive bench-top method for the preparation of nanoscale patterns of a range of active structures in a parallel manner.

3.8 Conclusions and Outlook

μCP was first developed for the preparation of patterned SAMs of *n*-alkylthiols on gold substrates. Inspired by the pioneering studies of Whitesides and coworkers, many research groups have shown that the nature of the ink, the stamp, and the substrate can be widely modified in order to improve the printing quality and to exploit the possibilities of μCP for a range of applications in materials and life sciences.

μCP has proven to be a valuable method for the preparation of microstructured and nanostructured surfaces, and has emerged as a versatile tool for the preparation of biological microarrays. Interesting methods for the simultaneous μCP of multiple biological inks have recently been reported. The highly selective molecular

recognition of biomolecules may also be used in the replication of arrays by μCP. Furthermore, it can be foreseen that surface reactions induced by μCP will provide a straightforward and versatile method for surface chemistry in general, and for the fabrication of (bio)molecular microarrays and nanoarrays in particular.

References

1 Carter, T.F. (1955) *The Invention of Printing in China and its Spread Westward*, Ronald Press Co., New York.
2 Kipphan, H. (2000) *Handbuch der Printmedien: Technologien und Produktionsverfahren*, Springer, Berlin.
3 Adams, J.M., Faux, D.D., and Rieber, J.J. (1996) *Printing Technology*, 4th edn, Delmare Publishers, Albany, NY.
4 Kumar, A. and Whitesides, G.M. (1993) *Appl. Phys. Lett.*, **63**, 2002.
5 Xia, Y., Venkateswaran, N., Qin, D., Tien, J., and Whitesides, G.M. (1998) *Langmuir*, **14**, 363.
6 Li, H., Kang, D.-J., Blamire, M.G., and Huck, W.T.S. (2002) *Nano Lett.*, **2**, 347.
7 Li, H., Muir, B.V.O., Flichet, G., and Huck, W.T.S. (2003) *Langmuir*, **19**, 1963.
8 Ruiz, S.A. and Chen, C.S. (2007) *Soft Matter*, **3**, 168.
9 Perl, A., Reinhoudt, D.N., and Huskens, J. (2009) *Adv. Mater.*, **21**, 2257.
10 Stein, J., Lewis, N.L., Gao, Y., and Scott, R.A. (1999) *J. Am. Chem. Soc.*, **121**, 3693.
11 Bietsch, A. and Michel, B. (2000) *J. Appl. Phys.*, **88**, 4310.
12 Delamarche, E., Schmid, H., Michel, B., and Biebuyck, H. (1997) *Adv. Mater.*, **9**, 741.
13 Hui, C.Y., Jagota, A., Lin, Y.Y., and Kramer, E.J. (2002) *Langmuir*, **18**, 1394.
14 Odom, T.W., Love, J.C., Wolfe, D.B., Paul, K.E., and Whitesides, G.M. (2002) *Langmuir*, **18**, 5314.
15 Schmid, H. and Michel, B. (2000) *Macromolecules*, **33**, 3042.
16 Wilbur, J.L., Kumar, A., Kim, E., and Whitesides, G.M. (1994) *Adv. Mater.*, **6**, 600.
17 Guo, Q., Teng, X., and Yang, H. (2004) *Nano Lett.*, **4**, 1657.
18 Trimbach, D., Feldman, K., Spencer, N.D., Broer, D.J., and Bastiaansen, C.W.M. (2003) *Langmuir*, **19**, 10957.
19 Delamarche, E., Schmid, H., Bietsch, A., Larsen, N.B., Rothuizen, H., Michel, B., and Biebuyck, H. (1998) *J. Phys. Chem. B*, **102**, 3324.
20 Wilbur, J.L., Biebuyck, H.A., MacDonald, J.C., and Whitesides, G.M. (1995) *Langmuir*, **11**, 825.
21 Jeon, N., Finnie, K., Branshaw, K., and Nuzzo, R. (1997) *Langmuir*, **13**, 3382.
22 Brzoska, J.B., Azouz, I.B., and Rondelez, F. (1994) *Langmuir*, **10**, 4367.
23 Allara, D.L., Parikh, A.N., and Rondelez, F. (1995) *Langmuir*, **11**, 2357.
24 Owen, M.J. and Smith, P.J. (1994) *J. Adhes. Sci. Technol.*, **8**, 1063.
25 Hillborg, H., Tomczak, N., Olah, A., Schonherr, H., and Vancso, G.J. (2004) *Langmuir*, **20**, 785.
26 Hu, S., Ren, X., Bachman, M., Sims, C.E., Li, G.P., and Allbritton, N. (2002) *Anal. Chem.*, **74**, 4117.
27 Delamarche, E., Donzel, C., Kamounah, S.S., Wolf, H., Geissler, M., Stutz, R., Schmid-Winkel, P., Michel, B., Mathieu, H.J., and Schaumburg, K. (2003) *Langmuir*, **19**, 8749.
28 Sadhu, V.B., Perl, A., Peter, M., Rozkiewicz, D.I., Engbers, G., Ravoo, B.J., Reinhoudt, D.N., and Huskens F J. (2007) *Langmuir*, **23**, 6850.
29 Bernard, A., Delamarche, E., Schmid, H., Michel, B., Bosshard, H.R., and Biebuyck, H. (1998) *Langmuir*, **14**, 2225.
30 Bernard, A., Renault, J.-P., Michel, B., Bosshard, H.R., and Delamarche, E. (2000) *Adv. Mater.*, **12**, 1067.
31 Lange, S.A., Benes, V., Kern, D.P., Hoerber, J.K.H., and Bernard, A. (2004) *Anal. Chem.*, **76**, 1641.
32 Auletta, T., Dordi, B., Mulder, A., Sartori, A., Onclin, S., Bruinink, C.M., Peter, M., Nijhuis, C.A., Beijleveld, H., Schonherr, H., Vancso, G.J., Casnati, A., Ungaro, R., Ravoo, B.J., Huskens, J., and Reinhoudt,

D.N. (2004) *Angew. Chem., Int. Ed.*, **43**, 369.
33 Bruinink, C.M., Nijhuis, C.A., Peter, M., Dordi, B., Crespo Biel, O., Auletta, T., Mulder, A., Schönherr, H., Vancso, G.J., Huskens, J., and Reinhoudt, D.N. (2005) *Chem. Eur. J.*, **11**, 3988.
34 Nijhuis, C.A., Sinha, J.K., Wittstock, G., Huskens, J., Ravoo, B.J., and Reinhoudt, D.N. (2006) *Langmuir*, **22**, 9770.
35 Ludden, M.L.W., Mulder, A., Schulze, K., Subramaniam, V., Tampe, R., and Huskens, J. (2008) *Chem. Eur. J.*, **14**, 2044.
36 Kim, Y., Kim, D., Park, J., Shin, G., Kim, G., and Ha, J. (2008) *Langmuir*, **24**, 14289.
37 Wu, X.C., Bittner, A.M., and Kern, K. (2004) *Adv. Mater.*, **16**, 413.
38 Degenhart, G.H., Dordi, B., Schonherr, H., and Vancso, G.J. (2004) *Langmuir*, **20**, 6216.
39 Kohli, N., Dvornic, P.R., Kaganove, S.N., Worden, R.M., and Lee, I. (2004) *Macromol. Rapid Commun.*, **25**, 935.
40 Campbell, C.J., Smoukov, S.K., Bishop, K.J.M., and Grzybowski, B.A. (2005) *Langmuir*, **21**, 2637.
41 Vogelaar, L., Barsema, J.N., van Rijn, C.J.M., Nijdam, W., and Wessling, M. (2003) *Adv. Mater.*, **15**, 1385.
42 Vogelaar, L., Lammertink, R.G.H., Barsema, J.N., Nijdam, W., Bolhuis-Versteeg, L.A.M., van Rijn, C.J.M., and Wessling, M. (2005) *Small*, **1**, 645.
43 Xu, H., Ling, X., van Bennekom, J., Duan, X., Ludden, M.J.W., Reinhoudt, D.N., Wessling, M., Lammertink, R.G.H., and Huskens, J. (2009) *J. Am. Chem. Soc.*, **2**, 797.
44 Allen, C.G., Dorr, J.C., Khandekar, A.A., Beach, J.D., Schick, I.C., Schick, E.J., Collins, R.T., and Kuech, T.F. (2007) *Thin Solid Films*, **515**, 6812.
45 Hidber, P.O., Helbig, W., Kim, E., and Whitesides, G.M. (1996) *Langmuir*, **12**, 1375.
46 Hidber, P.O., Nealey, P.F., Helbig, W., and Whitesides, G.M. (1996) *Langmuir*, **12**, 5209.
47 Kind, H., Geissler, M., Schmid, H., Michel, B., Kern, K., and Delamarche, E. (2000) *Langmuir*, **16**, 6367.
48 Xue, M., Zhang, Z., Zhu, N., Wang, F., Zhao, X., and Cao, T. (2009) *Langmuir*, **25**, 4347.
49 Chen, J., Mela, P., Möller, M., and Lensen, M.C. (2009) *ACS Nano*, **3**, 1451.
50 Whitesides, G.M., and Laibinis, P.E. (1990) *Langmuir*, **6**, 87.
51 Wenzl, I., Yam, C.M., Barriet, D., and Lee, T.R. (2003) *Langmuir*, **19**, 10217.
52 Colorado, R. Jr and Lee, T.R. (2003) *Langmuir*, **19**, 3288.
53 Burleigh, T.D., Gu, Y., Donahey, G., Vida, M., and Waldeck, D.H. (2001) *Corrosion*, **57**, 1066.
54 Jennings, G.K., Yong, T.-H., Munro, J.C., and Laibinis, P.E. (2003) *J. Am. Chem. Soc.*, **125**, 2950.
55 Petrenko, V.F. and Peng, S. (2003) *Can. J. Phys.*, **81**, 387.
56 Leggett, G.J. (2003) *Anal. Chim. Acta*, **479**, 17.
57 Ahn, H.-S., Cuong, P.D., Park, S., Kim, Y.-W., and Lim, J.-C. (2003) *Wear*, **255**, 819.
58 Adams, D.M., Brus, L., Chidsey, C.E.D., Creager, S., Creutz, C., Kagan, C.R., Kamat, P.V., Lieberman, M., Lindsay, S., Marcus, R.A., Metzger, R.M., Michel-Beyerle, M.E., Miller, J.R., Newton, M.D., Rolison, D.R., Sankey, O., Schanze, K.S., Yardley, J., and Zhu, X. (2003) *J. Phys. Chem. B*, **107**, 6668.
59 Salomon, A., Cahen, D., Lindsay, S., Tomfohr, J., Engelkes, V.B., and Frisbie, C.D. (2003) *Adv. Mater.*, **15**, 1881.
60 Aizenberg, J. (2000) *J. Chem. Soc., Dalton Trans.*, 3963.
61 Ostuni, E., Yan, L., and Whitesides, G.M. (1999) *Colloids Surf. B*, **15**, 3.
62 Mrksich, M. (2002) *Curr. Opin. Chem. Biol.*, **6**, 794.
63 Xia, Y., Zhao, X., and Whitesides, G.M. (1996) *Microelectron. Eng.*, **32**, 255.
64 Kumar, A., Biebuyck, H.A., Abbott, N.L., and Whitesides, G.M. (1992) *J. Am. Chem. Soc.*, **114**, 9188.
65 Geissler, M., Schmid, H., Bietsch, A., Michel, B., and Delamarche, E. (2002) *Langmuir*, **18**, 2374.
66 Wolfe, D.B., Love, J.C., Paul, K.E., Chabinyc, M.L., and Whitesides, G.M. (2002) *Appl. Phys. Lett.*, **80**, 2222.

67 Zhao, X.M., Wilbur, J.L., and Whitesides, G.M. (1996) *Langmuir*, **12**, 3257.

68 Love, J.C., Wolfe, D.B., Haasch, R., Chabinyc, M.L., Paul, K.E., Whitesides, G.M., and Nuzzo, R.G. (2003) *J. Am. Chem. Soc.*, **125**, 2597.

69 Carvalho, A., Geissler, M., Schmid, H., Michel, B., and Delamarche, E. (2002) *Langmuir*, **18**, 2406.

70 Michel, B., Bernard, A., Bietsch, A., Delamarche, E., Geissler, M., Juncker, D., Kind, H., Renault, J.P., Rothuizen, H., Schmid, H., Schmidt-Winkel, P., Stutz, R., and Wolf, H. (2001) *IBM J. Res. Dev.*, **45**, 697.

71 Love, J.C., Wolfe, D.B., Chabinyc, M.L., Paul, K.E., and Whitesides, G.M. (2002) *J. Am. Chem. Soc.*, **124**, 1576.

72 Wolf, S. (1990) *Silicon Processing for the VLSI Era*, Lattice Press, Sunset Beach.

73 Kim, E., Whitesides, G.M., Freiler, M.B., Levy, M., Lin, J.L., and Osgood, R.M. (1996) *Nanotechnology*, **7**, 266.

74 Delamarche, E., Geissler, M., Wolf, H., and Michel, B. (2002) *J. Am. Chem. Soc.*, **124**, 3834.

75 Saalmink, M., van der Marel, C., Stapert, H.R., and Burdinski, D. (2006) *Langmuir*, **22**, 1016.

76 Trimbach, D.C., Al-Hussein, M., de Jeu, W.H., Decré, M., Broer, D.J., and Bastiaansen, C.W.M. (2004) *Langmuir*, **20**, 4738.

77 Lee, M.S., Hong, S.-C., and Kim, D. (2004) *Jpn. J. Appl. Phys.*, **43**, 8347.

78 Shah, R.R., Merreceyes, D., Husemann, M., Rees, I., Abbott, N.L., Hawker, C.J., and Hedrick, J.L. (2000) *Macromolecules*, **33**, 597.

79 Tu, H., Heitzman, C.E., and Braun, P.V. (2004) *Langmuir*, **20**, 8313.

80 Huang, S., Dai, L., and Mau, A. (2002) *Physica B*, **322**, 333.

81 Gu, G., Philipp, G., Wu, X., Burghard, M., Bittner, A.M., and Roth, S. (2001) *Adv. Funct. Mater.*, **11**, 295.

82 Argyrakis, P., Teo, L., Stevenson, T., and Cheung, R. (2005) *Microelectron. Eng.*, **78–79** 647.

83 Casimirius, S., Flahaut, E., Laberty-Robert, C., Malaquin, L., Carcenac, F., Laurent, C., and Vieu, C. (2004) *Microelectron. Eng.*, **73–74**, 564.

84 Melosh, N.A., Boukai, A., Diana, F., Gerardot, B., Badolato, A., Petroff, P.M., and Heath, J.R. (2003) *Science*, **300**, 112.

85 Loo, Y.L., Hsu, J.W.P., Willett, R.L., Baldwin, K.W., West, K.W., and Rogers, J.A. (2002) *J. Vac. Sci. Technol. B*, **20**, 2853.

86 Jeon, S., Menard, E., Park, J.-U., Maria, J., Meitl, M., Zaumseil, J., and Rogers, J.A. (2004) *Adv. Mater.*, **16**, 1369.

87 Menard, E., Bilhaut, L., Zaumseil, J., and Rogers, J.A. (2004) *Langmuir*, **20**, 6871.

88 Schmid, H., Wolf, H., Allenspach, R., Riel, H., Karg, S., Michel, B., and Delamarche, E. (2003) *Adv. Funct. Mater.*, **13**, 145.

89 Lasky, J.B. (1986) *Appl. Phys. Lett.*, **48**, 78.

90 Tong, Q.Y. (2001) *Mater. Sci. Eng., B*, **B87**, 323.

91 Zaumseil, J., Meitl, M.A., Hsu, J.W.P., Acharya, B.R., Baldwin, K.W., Loo, Y.-L., and Rogers, J.A. (2003) *Nano Lett.*, **3**, 1223.

92 Loo, Y.L., Willett, R.L., Baldwin, K.W., and Rogers, J.A. (2002) *Appl. Phys. Lett.*, **81**, 562.

93 Loo, Y.L., Lang, D.V., Rogers, J.A., and Hsu, J.W.P. (2003) *Nano Lett.*, **3**, 913.

94 Loo, Y.L., Willett, R.L., Baldwin, K.W., and Rogers, J.A. (2002) *J. Am. Chem. Soc.*, **124**, 7654.

95 Wang, Z., Yuan, J., Zhang, J., Xing, R., Yan, D., and Han, Y. (2003) *Adv. Mater.*, **15**, 1009.

96 Helt, J.M., Drain, C.M., and Batteas, J.D. (2004) *J. Am. Chem. Soc.*, **126**, 628.

97 Wolfe, D.B., Love, J.C., Gates, B.D., Whitesides, G.M., Conroy, R.S., and Prentiss, M. (2004) *Appl. Phys. Lett.*, **84**, 1623.

98 Childs, W.R. and Nuzzo, R.G. (2002) *J. Am. Chem. Soc.*, **124**, 13583.

99 Childs, W.R. and Nuzzo, R.G. (2004) *Adv. Mater.*, **16**, 1323.

100 Jackman, R.J., Duffy, D.C., Cherniavskaya, O., and Whitesides, G.M. (1999) *Langmuir*, **15**, 2973.

101 Jackman, R.J., Duffy, D.C., Ostuni, E., Willmore, N.D., and Whitesides, G.M. (1998) *Anal. Chem.*, **70**, 2280.

102 Sassolas, A., Leca-Bouvier, B.D., and Blum, L.J. (2008) *Chem. Rev.*, **108**, 109.

103 Zhu, H. and Snyder, M. (2003) *Curr. Opin. Chem. Biol.*, **7**, 55.

104 Feizi, T., Fazio, F., Chai, W., and Wong, C.H. (2003) *Curr. Opin. Struct. Biol.*, **13**, 637.

105 Kane, R.S., Takayama, S., Ostuni, E., Ingber, D.E., and Whitesides, G.M. (1999) *Biomaterials*, **20**, 2363.

106 Chen, C.S., Mrksich, M., Huang, S., Whitesides, G.M., and Ingber, D.E. (1997) *Science*, **276**, 1425.

107 (a) Graber, D.J., Zieziulewicz, T.J., Lawrence, D.A., Shain, W., and Turner, J.N. (2003) *Langmuir*, **19**, 5431;(b) LaGraff, J.R. and Chu-LaGraff, Q. (2006) *Langmuir*, **22**, 4685.

108 Tan, J.L., Tien, J., and Chen, C.S. (2002) *Langmuir*, **18**, 519.

109 Renault, J.P., Bernard, A., Bietsch, A., Michel, B., Bosshard, H.R., Delamarche, E., Kreiter, E., Hecht, B., and Wild, U.P. (2003) *J. Phys. Chem. B*, **107**, 703.

110 Yeung, C.K., Lauer, L., Offenhausser, A., and Knoll, W. (2001) *Neurosci. Lett.*, **301**, 147.

111 Park, T.J., Lee, S.Y., Lee, S.J., Park, J.P., Yang, K.S., Lee, K.B., Ko, S., Park, J.B., Kim, T., Kim, S.K., Shin, Y.B., Chung, B.H., Ku, S.J., Kim, D.H., and Choi, I.S. (2006) *Anal. Chem.*, **78**, 7197.

112 Kwak, S.K., Lee, G.S., Ahn, D.J., and Choi, J.W. (2004) *Mater. Sci. Eng. C - Bio.*, **24**, 151.

113 Wilhelm, T. and Wittstock, G. (2002) *Langmuir*, **18**, 9485.

114 Biasco, A., Pisignano, D., Krebs, B., Cingolani, R., and Rinaldi, R. (2005) *Synthetic Met.*, **153**, 21.

115 James, C.D., Davis, R.C., Kam, L., Craighead, H.G., Isaacson, M., Turner, J.N., and Shain, W. (1998) *Langmuir*, **14**, 741.

116 Delamarche, E., Geissler, M., Bernard, A., Wolf, H., Michel, B., Hilborn, J., and Donzel, C. (2001) *Adv. Mater.*, **13**, 1164.

117 Geissler, M., Bernard, A., Bietsch, A., Schmid, H., Michel, B., and Delamarche, E. (2000) *J. Am. Chem. Soc.*, **122**, 6303.

118 Zheng, Z., Jang, J.W., Zheng, G., and Mirkin, C.A. (2008) *Angew. Chem., Int. Ed.*, **47**, 9951.

119 Rozkiewicz, D.I., Brugman, W., Kerkhoven, R.M., Ravoo, B.J., and Reinhoudt, D.N. (2007) *J. Am. Chem. Soc.*, **129**, 11593.

120 Rożkiewicz, D.I., Gierlich, J., Burley, G.A., Gutsmiedl, K., Carell, T., Ravoo, B.J., and Reinhoudt, D.N. (2007) *ChemBioChem*, **8**, 1997.

121 Xu, C., Taylor, P., Ersoz, M., Fletcher, P.D.J., and Paunov, V. (2003) *J. Mater. Chem.*, **13**, 3044.

122 Hovis, J.S. and Boxer, S.G. (2001) *Langmuir*, **17**, 3400.

123 Kung, L.A., Kam, L., Hovis, J.S., and Boxer, S.G. (2000) *Langmuir*, **16**, 6773.

124 Renault, J.P., Bernard, A., Juncker, D., Michel, B., Bosshard, H.R., and Delamarche, E. (2002) *Angew. Chem., Int. Ed.*, **41**, 2320.

125 Jang, C.-H., Tingey, M.L., Korpi, N.L., Wiepz, G.J., Schiller, J.H., Bertics, P.J., and Abbott, N.L. (2005) *J. Am. Chem. Soc.*, **127**, 8912.

126 Tingey, M.L., Wilyana, S., Snodgrass, E.J., and Abbott, N.L. (2004) *Langmuir*, **20**, 6818.

127 Kim, J. and Crooks, R.M. (2007) *Anal. Chem.*, **79**, 7267.

128 Yu, A.A., Savas, T.A., Taylor, G.S., Elie, A.G., Smith, H.I., and Stellacci, F. (2005) *Nano Lett.*, **5**, 1061.

129 Helmuth, J.A., Schmid, H., Stutz, R., Stemmer, A., and Wolf, H. (2006) *J. Am. Chem. Soc.*, **128**, 9296.

130 (a) Love, J.C., Estroff, L.A., Kriebel, J.K., Nuzzo, R.G., and Whitesides, G.M. (2005) *Chem. Rev.*, **105**, 1103; (b) Onclin, S., Ravoo, B.J., and Reinhoudt, D.N. (2005) *Angew. Chem., Int. Ed.*, **44**, 6282.

131 Ravoo, B.J. (2009) *J. Mater. Chem.*, **19**, 8902.

132 Sullivan, T.P. and Huck, W.T.S. (2003) *Eur. J. Org. Chem.*, 17.

133 (a) Yan, L., Zhao, X.M., and Whitesides, G.M. (1998) *J. Am. Chem. Soc.*, **120**, 6179; (b) Yan, L., Huck, W.T.S., Zhao, X.M., and Whitesides, G.M. (1999) *Langmuir*, **15**, 1208;(c) Lahiri, J., Ostuni, E., and Whitesides, G.M. (1999) *Langmuir*, **15**, 2055.

134 Sullivan, T.P., van Poll, M.L., Dankers, P.Y.W., and Huck, W.T.S. (2004) *Angew. Chem., Int. Ed.*, **43**, 4190.

135 Rozkiewicz, D.I., Ravoo, B.J., and Reinhoudt, D.N. (2005) *Langmuir*, **21**, 6337.

136 Rozkiewicz, D.I., Kraan, Y., Werten, M.W.T., de Wolf, F.A., Subramaniam, V.,

Ravoo, B.J., and Reinhoudt, D.N. (2006) *Chem. Eur. J.*, **12**, 6290.

137 Rozkiewicz, D.I., Jancewski, D., Verboom, W., Ravoo, B.J., and Reinhoudt, D.N. (2006) *Angew. Chem., Int. Ed.*, **45**, 5292.

138 (a) Rostovtsev, V.V., Green, L.G., Fokin, V.V., and Sharpless, K.B. (2002) *Angew. Chem., Int. Ed.*, **42**, 2596;(b) Tornøe, C.W., Christensen, C., and Meldal, M. (2002) *J. Org. Chem.*, **67**, 3057.

139 Kolb, H.C., Finn, M.G., and Sharpless, K.B. (2001) *Angew. Chem., Int. Ed.*, **40**, 2004.

140 Michel, O. and Ravoo, B.J. (2008) *Langmuir*, **24**, 12116.

141 Li, X.M., Peter, M., Huskens, J., and Reinhoudt, D.N. (2003) *Nano Lett.*, **3**, 1449.

142 Shestopalov, A.A., Clark, R.L., and Toone, E.J. (2007) *J. Am. Chem. Soc.*, **129**, 13818.

143 Snyder, P.W., Johannes, M.S., Vogen, B.N., Clark, R.L., and Toone, E.J. (2007) *J. Org. Chem.*, **72**, 7459.

144 Spruell, J.M., Sheriff, B.A., Rozkiewicz, D.I., Dichtel, W.R., Rohde, R.D., Reinhoudt, D.N., Stoddart, J.F., and Heath, J.R. (2008) *Angew. Chem., Int. Ed.*, **47**, 9927.

145 Hyun, J., Zhu, Y., Liebmann-Vinson, A., Beebe, T.P., and Chilkoti, A. (2001) *Langmuir*, **17**, 6358.

146 Feng, C.L., Vancso, G.J., and Schönherr, H. (2006) *Adv. Funct. Mater.*, **16**, 1306.

147 Nandivada, H., Chen, H.Y., Bondarenko, L., and Lahann F J. (2006) *Angew. Chem., Int. Ed.*, **45**, 3360.

148 Truong, T.T., Lin, R., Jeon, S., Lee, H., Maria, J., Gaur, A., Hua, F., Meinel, I., and Rogers, J.A. (2007) *Langmuir*, **23**, 2898.

149 Yoo, P.J., Choi, S.J., Kim, J.H., Suh, D., Baek, S.J., Kim, T.W., and Lee, H.H. (2004) *Chem. Mater.*, **16**, 5000.

150 Csucs, G., Künzler, T., Feldman, K., Robin, F., and Spencer, N.D. (2003) *Langmuir*, **19**, 6104.

151 Tormen, M., Borzenko, T., Steffen, B., Schmidt, G., and Molenkamp, L.W. (2002) *Microelectron. Eng.*, **61**, 469.

152 Bessueille, F., Pla-Roca, M., Mills, C.A., Martinez, E., Samitier, J., and Errachid, A. (2005) *Langmuir*, **21**, 12060.

153 Libouille, L., Bietsch, A., Schmid, H., Michel, B., and Delamarche, E. (1999) *Langmuir*, **15**, 300.

154 Liebau, M., Huskens, J., and Reinhoudt, D.N. (2001) *Adv. Funct. Mater.*, **11**, 147.

155 Perl, A., Peter, M., Ravoo, B.J., Reinhoudt, D.N., and Huskens, J. (2006) *Langmuir*, **22**, 7568.

156 Wu, X., Lenhert, S., Chi, L., and Fuchs, H. (2006) *Langmuir*, **22**, 7807.

157 Sharpe, R.B.A., Titulaer, B.J.F., Peeters, E., Burdinski, D., Huskens, J., Zandvliet, H.J.W., Reinhoudt, D.N., and Poelsema, B. (2006) *Nano Lett.*, **6**, 1235.

158 Xue, M., Yang, Y., and Cao, T. (2008) *Adv. Mater.*, **20**, 596.

159 Sharpe, R.B.A., Burdinski, D., Huskens, J., Zandvliet, H.J.W., Reinhoudt, D.N., and Poelsema, B. (2005) *J. Am. Chem. Soc.*, **127**, 10344.

4
Advances in Nanoimprint Lithography: 2-D and 3-D Nanopatterning of Surfaces by Nanoimoprint Lithography, Morphological Characterization, and Photonic Applications

Vincent Reboud, Timothy Kehoe, Nikolaos Kehagias, and Clivia M. Sotomayor Torres

4.1
Introduction

The advances in nanofabrication by emerging patterning methods made during the past fourteen years have been dramatic, ranging from laboratory-scale experiments reviewed in 2003 [1] to reports on some of these methods appearing in the road map of the most demanding industry, namely microelectronics [2]. Among these methods, which include self-assembly, microcontact printing, scanning probes and nanoimprint lithography (NIL), attention in this chapter is focused mainly on NIL, since it is perhaps the most mature of the then-emerging nanofabrication methods.

NIL as a nanopatterning process has intrinsically many advantages, since it combines simplicity with a wealth of functional materials based on polymers. At present, one trend is to develop NIL processes to fabricate three-dimensional (3-D) structures, tiered or wood-pile or combinations thereof, among others. There is an enormous demand to replicate polymers with combined micro- and nanometer features that require 3-D nanopatterning. In this chapter, details are provided of the many studies reported to date in 3-D NIL, in addition to those of the present authors' own contributions to the field. It will be seen that whilst 3-D NIL and resolution below 20 nm are still to be met, solid progress is being made nonetheless.

Today, the transition from a laboratory-scale method to a full-scale technology is still in progress, albeit well advanced, with commercial printing equipment available in the market, a preliminary set of NIL processes for specific applications [3], and continued rapid developments in tools, designs, stamps, process simulations, materials and applications. During this transition, the importance of metrology cannot be underestimated since, without nanometrology for the critical dimensions, it would be very difficult to transfer such applications to the production stage. The status of metrology in NIL is also described in the chapter, with an incursion into the physical properties of polymer films thinner than 10 nm. Clearly, there is still a long way to go in this area, and innovative methods are much in demand.

Nanotechnology, Volume 8: Nanostructured Surfaces. Edited by Lifeng Chi

ISBN: 978-3-527-31739-4

One application of NIL which has experienced major progress due to the functionality offered by polymers when mixed with nanoparticles (whereby they are converted into a nanocomposite), has been in the area of *photonics*. Here, the feature sizes are slightly larger and the tolerances slightly more relaxed than in, for example, electronic applications. Hence, the potential uses of NIL can be illustrated in the field of photonics by two device-like structures, and a perspective offered in this area of applications.

4.2
Three-Dimensional, Nanoimprint-Based Lithography

4.2.1
Background

In this section, an overview is provided of the nanoimprint-based techniques used to fabricate 3-D structures. During the past few years, several next-generation lithography (NGL) techniques have been developed and progressed beyond the 22 nm barrier node. To meet this demand, the semiconductor industry has for example used adapted photolithographic processes with ever-decreasing wavelengths – from optical, to ultraviolet (UV), to deep UV, to extreme UV – in an effort to beat the diffraction limit, but this led to dramatic increases in tool costs. In contrast, imprint-based lithography techniques require neither expensive projection optics, advanced illumination sources, nor specialized resist materials.

There are two basic approaches to imprint lithography (Figure 4.1) based on the use of temperature (thermal NIL) and/or ultraviolet light (UV-NIL) to transfer the

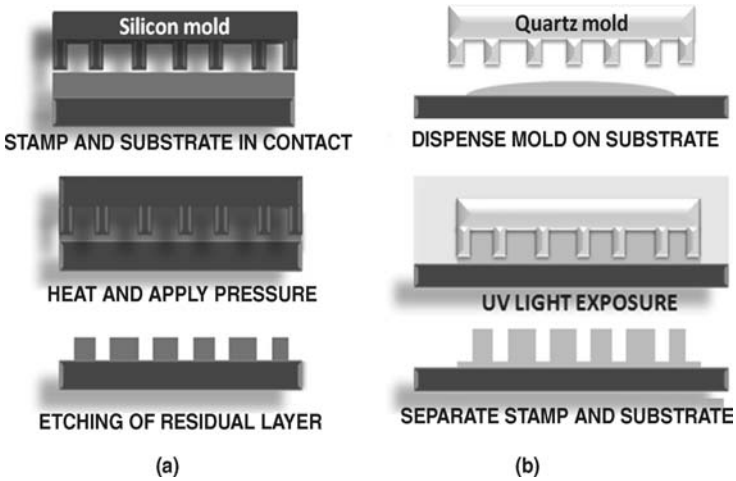


Figure 4.1 Schematic representations of (a) the thermal NIL process and (b) the UV-NIL process.

Table 4.1 Surface energies of common materials used in nanoimprint lithography. Values taken from Ref. [9].

Material	Surface energy ($mN\,m^{-1}$)
PMMA	41.1
PS	40.7
PTFE	15.6
−CF3 and −CF2	15–17
Silicon surface	20–26

mold to the imprinted resist. In thermal NIL (Figure 4.1a), as developed by S. Chou *et al.* [4, 5], a hard template (mold/stamp) such as silicon is used to imprint a thermoplastic polymer that is then heated to a temperature above the polymer glass transition temperature (T_g), while applying a relatively high pressure. After a specific time, which depends on the architecture and topography of the template, the polymer is cooled to a temperature below its T_g, at which point the stamp and substrate can be separated. In order to minimize the adhesion between the imprinted resist and the mold, a fluorine-based material is deposited onto the mold surface [6] to serve as an anti-adhesive agent.

The values of surface energy of the most frequently used materials are listed in Table 4.1. Features as small as 5 nm [7] have been reported, while in a separate report 200 mm wafer surfaces have been patterned successfully [8]. These features position NIL as a high-throughput lithographic technique capable of functioning within small- and medium-sized manufacturing companies.

In 1998, a modified nanoimprint process was developed at the University of Texas (UT-Austin) by the group of C.G. Wilson [10]. In this process, which was referred to as step and flash nanoimprint lithography (SFIL), a photo-curable liquid resist was molded in a step-and-repeat manner by applying UV light when the transparent mold was in contact with the resist (see Figure 4.1b). Unlike thermal NIL, the SFIL process was carried out at room temperature and used relatively low pressures, which in turn led to a significant reduction in the imprint time. The latest generation of SFIL tools has a specification for 300 mm wafers [11], with sub-50 nm resolution [12], and today several instruments are available commercially from several suppliers, including the EV Group, Molecular Imprints, Nanonex, Obducat, and Smart Equipment Technology.

The main achievements of thermal and UV-based NIL are summarized and compared in Table 4.2.

In comparison to the photolithography process, NIL-based methods are referred to as "1×1 processes"; that is, the resolution of both imprint-based techniques is determined by the resolution of the template. Unfortunately, this advantage of ultimate resolution could in time become a disadvantage, since any unintentional template artifact might be transferred with high fidelity to the molded material. It is for this reason that high-quality (noncontaminated) templates must be fabricated,

Table 4.2 Comparison of the two basic imprint techniques.

Technique	Smallest/ largest features in same print	Minimum pitch (nm)	Largest wafer printed (mm)	Overlay accuracy (nm)	T align, T print, T release, T cycle	No. of times stamp used
NIL	5 nm [13]/N/A	14	200 [14]	500	Minutes, 10 s, Min, 10–15 min	>50
SFIL	25 nm/μm	50	300 [15, 16] stamp size: $\sim 26 \times 26\,mm^2$	50 [17]	20 wafers per hour [18]	800

and nondestructive inspected methods employed to maintain the quality control of the system. An example of a nondestructive, nano-metrological technique is described in the following subsection.

Currently, imprint-based process are investigated for the manufacture of, for example, photonic crystal devices [19], micro- and nano-optical components [20], and media storage devices [21]. Today, one of the most interesting applications and capabilities of NIL techniques is the direct imprinting of multilevel structures; this is of particular interest in the semiconductor industry, as the simultaneous imprinting of multiple device levels can greatly reduce the number of steps (by about 50%) associated with back-end-of-line (BEOL) processing [22].

In 2005, a research group at the University of Texas proposed the use of a dual damascene imprint template, rather than a traditional lithographic method, to build the metal interconnect stack [23]. For such a process a hard template with tiered-like structures was required and, depending on the type of imprint method, either a transparent (UVNIL) or nontransparent (thermal NIL) mold was used. The imprinted material used would then need to be either a UV-curable resist or a thermoplastic film, respectively. Although, in most thermal NIL cases, the template used was of similar size to the imprinted substrate, the production of the mold was always very expensive. Moreover, in UVNIL (and particularly in SFIL mode) the template was much smaller (ca. $2 \times 2\,cm^2$) and the imprint proceeded in a step-and-repeat fashion. This led to an increase in throughput in SFIL, since lower temperatures and pressures were required compared to thermal NIL. However, these lower values also helped to achieve sub-100 nm alignment between the two layers.

A schematic representation of the multilayer imprint process is shown in Figure 4.2. In this procedure, a rigid transparent stamp is first brought into close proximity to, and in alignment with, the substrate. A low-viscosity photocurable resist is then dispensed onto the pre-patterned substrate, most likely using areas with metallic stripes. The stamp and substrate are then brought into contact, followed by UV light exposure when the cavities in the stamps has been filled by the resist. In most cases, the resist used is a low-molecular-weight monomer incorporating

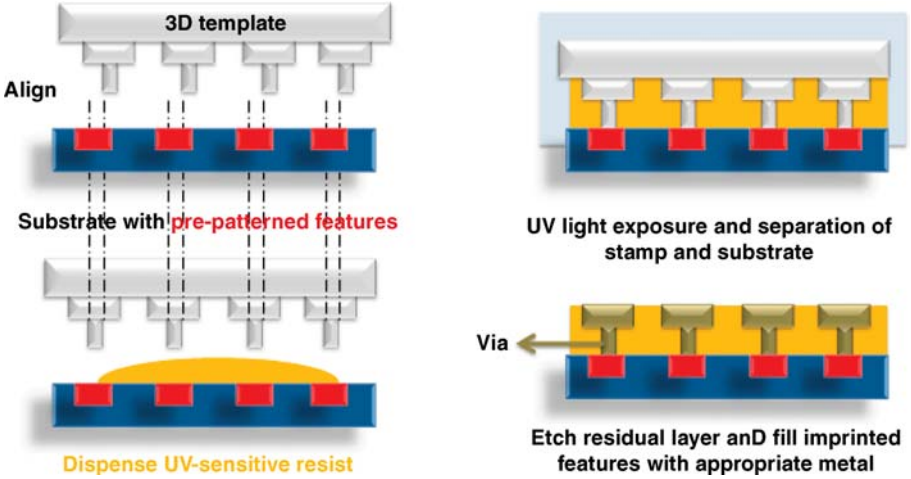


Figure 4.2 Fabrication process of a multilayer interconnection with UVNIL technology.

photoinitiator molecules. When the resist has polymerized the stamp can be removed, leaving behind an inverse patterned replica. As in thermal NIL, a thin residual layer always remains that can be removed by using a classical reactive ion etching (RIE) method. Infiltration of the opened (imprinted) areas is then carried out in order to connect the two levels. Compared to photolithographic methods, the imprint-based strategy for the fabrication of interconnected structures will reduce by half the steps required to wire the adjacent layers.

As described above, alignment of the via and upper wiring level is carried out at the template fabrication stage, such that only one alignment step per metal layer is required during the fabrication sequence [24]. Thus, it becomes clear that one of the most challenging issues of direct multilayer nanopatterning is the fabrication of 3-D stamps (Figure 4.3), and several techniques have been proposed for the creation of such templates. The "pros" and "cons" of each technique, together with representative scanning electron microscopy (SEM) images, are listed in Table 4.3. The data in Table 4.3 indicate that, although direct 3-D patterning can be achieved by NIL with a high throughput, it is still necessary to fabricate the master with a single lithographic technique (though a combination of techniques may sometimes be needed). There appear to be several limiting factors that hinder the use of these techniques for the mass production of 3-D devices, including high cost, low throughput, low resolution, and complexity of procedure.

In order to create more complex structures, several other methods involving a combination of different techniques have been investigated; details of these "nonconventional" 3-D patterning methods are listed in Table 4.4. For applications in photonics, and in particular for the creation of 3-D devices such as photonic crystals, it is important to minimize the number of steps in the stacking process. In this way, the excessive use of lift-off, sacrificial layer removal and etching that is normally performed in planar technologies, can be avoided. Moreover, alignment to

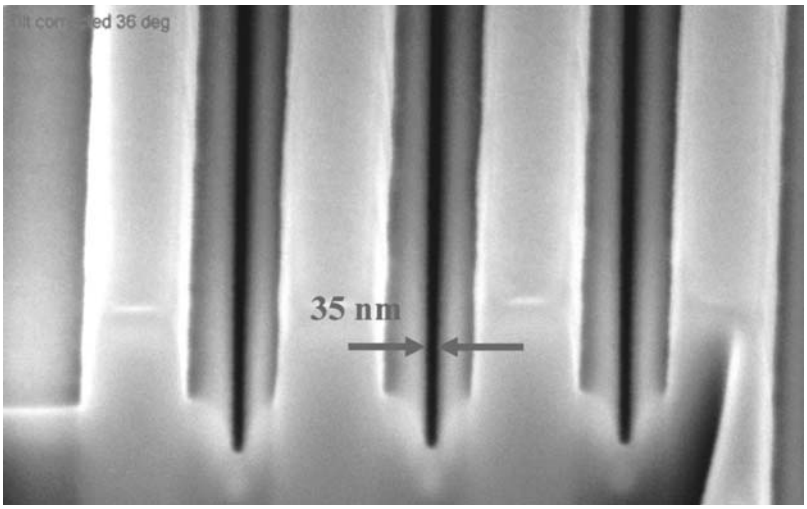


Figure 4.3 Tilted scanning electron microscopy (SEM) image of a tiered-like Si template fabricated by focused ion beam (FIB) lithography.

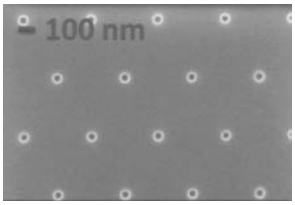
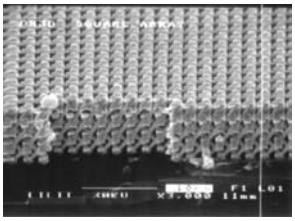
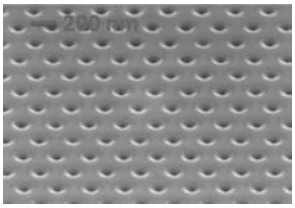
within $\sim\lambda/20$ is also necessary. One of the most difficult milestones for imprint-based technologies has been is to prove their ability to fabricate 3-D devices at the optical scales. Despite integration in the third dimension requiring compact, fast and cost-efficient device fabrication methods, several modified imprint-based techniques have been described as possible solutions to these problems.

Within the context of 3-D nanofabrication, two alternative fabrication processes have been developed, the details of which are presented below. Whilst both of these (imprint-based) techniques are capable of meeting the important requirements of high resolution and low cost, issues of alignment and reproducibility in large-scale areas have yet to be resolved. Hence, in the following subsections, after a brief discussion of the reversal nanoimprint technique and its potential, the limiting factors responsible for its poor uptake as a method for patterning 3-D device-like structures are outlined. In addition, as an alternative to the reversal imprint technique, a novel method of fabrication is introduced, namely reverse-contact UVNIL.

4.2.2
Reverse NIL

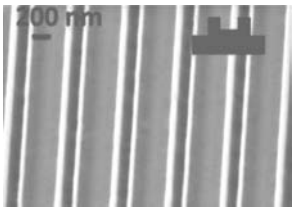
As discussed above, by using NIL it is possible to transfer a wide variety of stamp profiles into a polymer, although the 3-D patterning possibilities are limited. However, recent progress in nanoimprinting techniques has led to the introduction of a new method that is similar to NIL and might provide a solution to this problem. In the so-called "reverse nanoimprint" technique [30] or "bonding" process [31], the

Table 4.3 Conventional direct 3-D patterning techniques, indicating whether the technique is sequential or parallel.

Method	Advantages/Drawbacks	Features fabricated
Electron beam Resist is exposed to a focused electron beam. Two axes of rotation and enhanced control of focus for a true 3-D fabrication are needed (Sequential)	*Advantages:* High-resolution (sub-10 nm) *Drawbacks:* Low throughput Limited exposure depth Limited area Electron scattering in resist and substrate	 Sub-100 nm 2-D plasmonic crystal structure made by means of EBL
X-ray lithography Illumination of a X-ray mask. 3-D obtained by multiple exposures of the sample at tilted angles. (Parallel)	*Advantages:* Resolution sub-50 nm Aspect ratio ~20 High throughput *Drawbacks:* Complex process High cost	 Photonic crystal [25]
Focused ion beam Direct milling or growth of material by accelerated and focused ions. 3-D achieved by control of ion energy and tilted angles. (Sequential)	*Advantages:* High resolution (sub-20 nm) High aspect ratio ~40 *Drawbacks:* Low throughput Gaussian beam Limited area	 Photonic crystal structure with holes of 160 nm diameter
Two-photon lithography Photopolymer is exposed to a focused femtosecond laser beam. (Sequential)	*Advantages:* High resolution Arbitrary shapes In-depth writing *Drawbacks:* Low throughput Complex process	 Microbull [26]

(*Continued*)

Table 4.3 (*Continued*)

Method	Advantages/Drawbacks	Features fabricated
Direct 3-D NIL Deformation of a thermoplastic polymer by a rigid mold, followed by dry etching (Parallel)	*Advantages:* High resolution (sub-10 nm) High throughput *Drawbacks:* Low aspect ratio ~3 Nonarbitrary shapes	 Top view SEM image of direct 3-D structure. Insert shows a schematic of the profile

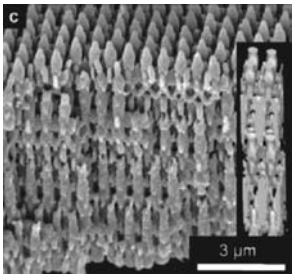
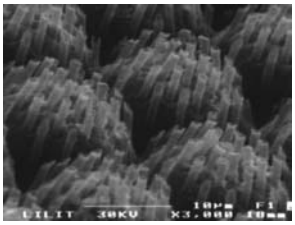
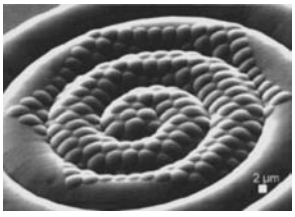
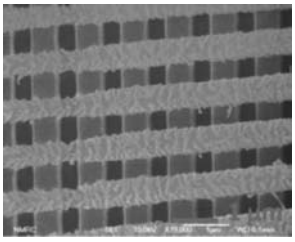
EBL = electron beam lithography.

polymer film is first spin-coated onto the stamp and then transferred to the substrate; by simply repeating the process on the same substrate, it is then possible to build 3-D polymer structures in a layer-by-layer fashion. Most importantly, the reverse NIL process can be carried out at a much lower temperature ($T_{imp} \sim T_g$) than conventional NIL, which allows the imprint cycles to be shortened. Unfortunately, one problem associated with this technique when using thermoplastic polymers is that each additional layer requires polymers with a progressively lower glass transition temperature (T_g) [32], and this in turn limits the number of layers formed. The reverse nanoimprint process where, depending on the size and density of the stamp features, either a "whole-layer transfer" mode or an "inking" mode can be observed, is shown schematically in Figure 4.4.

In recently conducted reverse imprinting experiments [33], two different transfer modes were observed, depending on the stamp topography. For features above 1 μm, and if the protrusions on the stamp were not too dense – that is, if the stamp presented a protrusion area of about 25% or less – then the "inking" transfer mode was observed. In this case, a positive copy of the stamp remained on the substrate because only the polymer on top of the stamp protrusions was transferred, and as a result no residual layer remained after reverse imprinting. This was due to the fact that spin-coating on such large and separated features created a non-flat film. An example of reverse imprint fabrication using the inking mode is shown in Figure 4.5, where a grating structure was imprinted on a 1 μm-period grating. In this case, even at a pressure of 60 bar, transfer of the top layer did not damage the underlying cured polymer grating.

Recently, Nakajima *et al*. [34] showed that, by controlling the temperatures of the mold and substrate at temperatures above and below T_g, respectively, the reverse imprint technique could be used to fabricate 3-D structures (nanochannels), using the same polymer [poly(methylmethacrylate); PMMA].

Table 4.4 Nonconventional direct 3-D patterning techniques.

Method	Advantage/Drawbacks	Features fabricated
Two-photon lithography and phase mask Photopolymer is exposed to an unfocused laser beam through conformable phase mask. (Parallel)	*Advantages:* Parallel In-depth writing *Drawback:* Complex process	 Photonic crystal [27]
Combination of NIL and X-ray lithography X-ray lithography on a substrate pre-patterned by NIL	*Advantages:* Complex shapes High throughput *Drawback:* High cost	 Pillars on hemispheres [26]
Combination of lithographic steps and wet etching	*Advantages:* High resolution Complex shapes at a relatively high throughput *Drawbacks:* Complex process Nonarbitrary shapes	 Complex 3-D stamp for NIL [28]
Reverse NIL	*Advantages:* Sub-200 nm resolution Suspended structure *Drawbacks:* Alignment issues Reproducibility issues	 Wood pile-like structures [29]

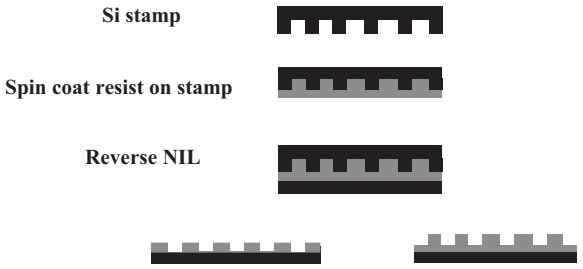


Figure 4.4 Schematics of the reverse nanoimprint process. Top to bottom: A polymer is first spin-coated onto the stamp, and then transferred to the flat or pre-patterned substrate.

4.2.3 Reverse-Contact UVNIL

A combination of the reversed NIL and UVNIL techniques has been shown to generate a promising nanofabrication technique termed reverse contact UVnanoimprint lithography (RUVNIL) [35–37]. This new technique has three main advantages: (i) the stamp does not need to be treated with an anti-adhesive layer; (ii) no residual layer remains after imprinting; and (iii) 3-D device-like structures can be obtained using the same polymer for each layer by repeating the procedure. Potentially, this method could be used to build up structures with several layers suitable, for example, in the fabrication of 3-D periodic structures. These might include photonic crystals with predicted defects, diffractive optical elements, and embedded channels for nano/micro fluidic devices for bioapplications.

This lithography process is illustrated schematically in Figure 4.6. A thin film of resist is first spin-coated onto the stamp (Figure 4.6b); this sacrificial polymer layer is used as an adherence promoter, as a planarization layer, and also to protect the stamp from being contaminated by the photocuring resist. A film of a UV crosslinkable polymer is then spin-coated onto the first layer (Figure 4.6c), after which the polymer bilayer is reverse-imprinted onto a flat or pre-patterned surface (Figure 4.6d). The stamp and substrate are then heated to a temperature above the T_g of mr-NIL 6000 (Figure 4.6e) and exposed to UV light. The stamp and substrate are separated just after a post-baking step (Figure 4.6f), thus ensuring a good adhesion between the polymer and the underlying substrate. Finally, both the unexposed polymer areas and the sacrificial layer are removed, leaving behind the negative features of the original stamp (Figure 4.6g). In this way, the oxygen plasma-etching step which is usually necessary in standard NIL is avoided.

Figure 4.7 shows the SEM images of 3-D structures fabricated by repeating the RUVNIL process described above. In particular, Figure 4.7a shows a cross-sectional SEM image of a bilayer woodpile-like structure, where there is no underflow of the second polymer layer on the first imprinted layer. In this case, 650 nm lines have been

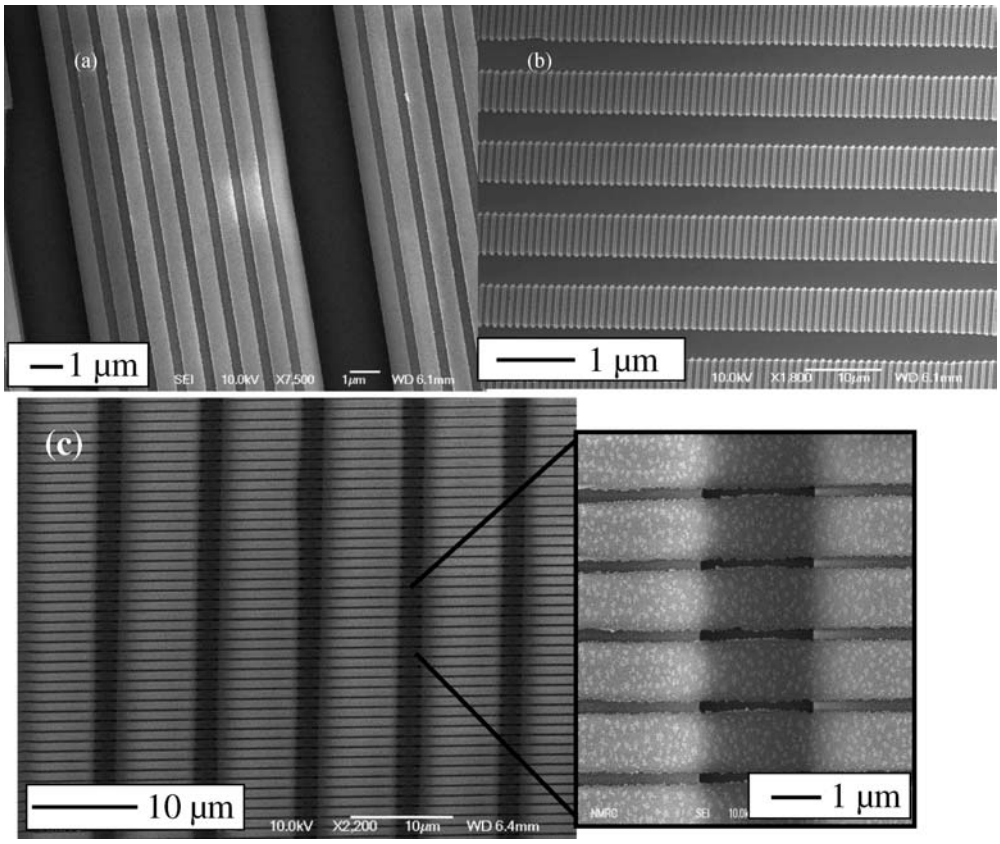


Figure 4.5 Examples of reverse-imprinted features in the whole-layer transfer mode. A 1 μm period grating is reverse-imprinted on a 10 μm period grating, with the lines (a) parallel or (b, c) perpendicular to each other. Depending on the features separation on the substrate, the transfer can be effective only on the protrusions (b) or on the whole surface, forming air-bridged structures (c).

RUVNIL-imprinted on 1 μm grating structures. A top-view SEM image of the same bilayer structure is shown in Figure 4.7b.

4.2.4 The Prospects for 3-D NIL

The mastering of 3-D NIL is expected to unlock an even wider range of applications, including supramolecular ordering and artificial tissues, while diffractive optical elements and complex light-handling curved structures constitute another avenue which is being actively pursued. However, such advances depend heavily on the stamp design and on material developments, as well as monitoring by suitable metrology.

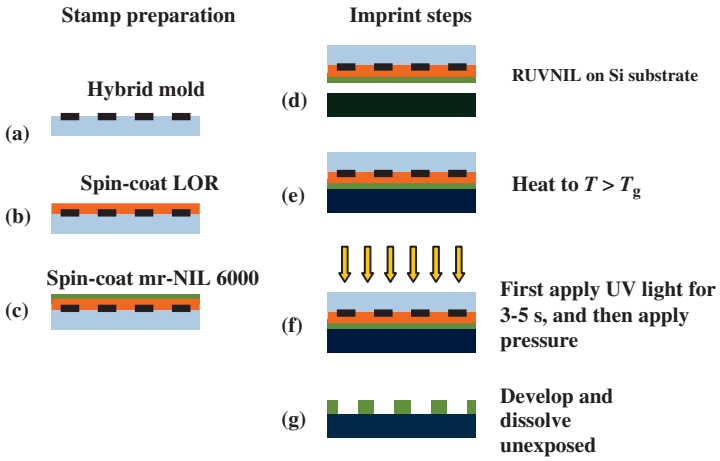


Figure 4.6 (a–c) Schematics of the stamp preparation steps; (d–g) Imprint steps of the RUVNIL technique.

The selective surface functionalization of horizontal or vertical or slopes represents a new challenge to be met in the development of artificial mesomaterials and micromaterials, with nanometer control of their properties – and, therefore, also of their functions.

4.3 Metrology for Nanoimprinting

4.3.1 Introduction

The metrological techniques used for NIL are similar to those that have been developed previously and are currently in use for nanofabrication generally, such as for deep UV optical lithography or electron beam lithography (EBL). Well-established techniques such as SEM [38, 39], atomic force microscopy (AFM) [40] and scatterometry [41] all find good use in characterizing surfaces structured by nanoimprinting. However, there are requirements to measure features that are particular to NIL, such as residual thickness and a complex line profile, and this influences the choice of established techniques which are used. This has also encouraged new developments, such as real-time scatterometry to measure line shape evolution due to reflow or annealing, and diffraction to monitor stamp-filling and detect defects. Polymer physical properties are critical to the nanoimprint process, and these include rheological properties, elastic modulus, and the T_g of the resist. Much effort has gone in to characterizing these at the nanometer scale, to detect changes due to the reduction of size, using methods including nanoindentation, Brillouin scattering, and photoacoustic metrology.

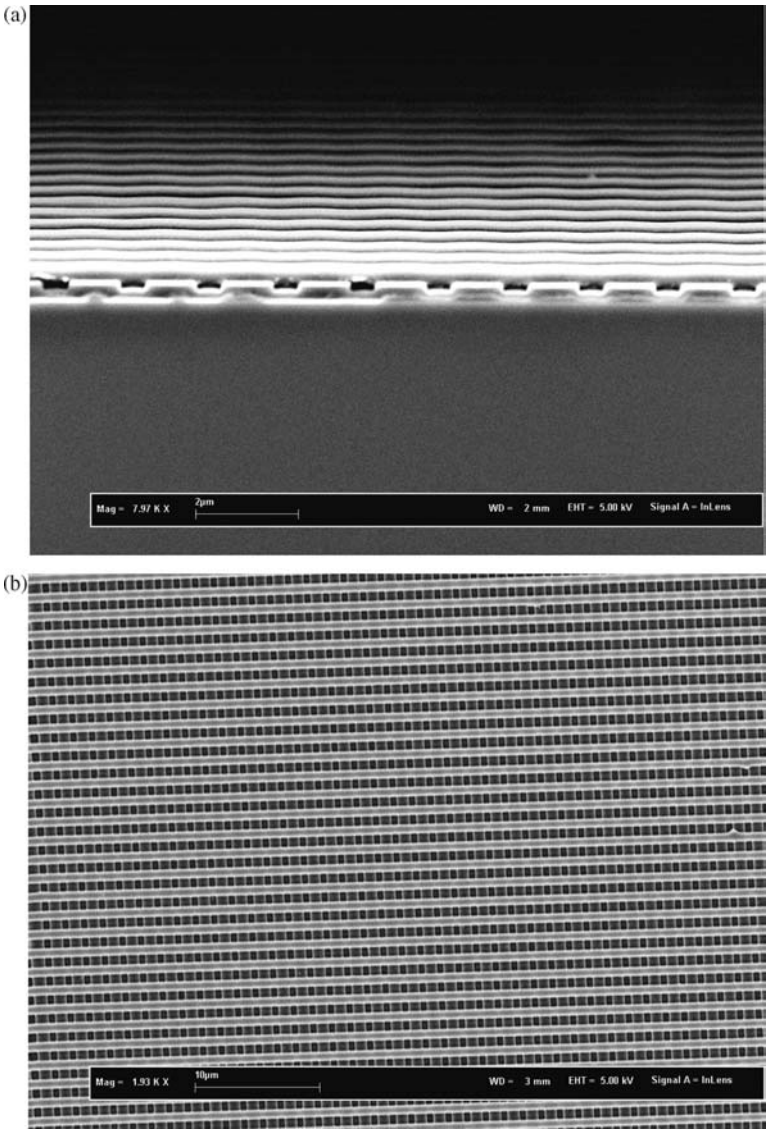


Figure 4.7 (a) Cross-sectional SEM image of a polymer woodpile-like structure; (b) Top view SEM image of the same 3-D cross-bar structure (see text for details).

4.3.2 Scanning Electron Microscopy

Scanning electron microscopy is the standard "workhorse" metrology technique for NIL, as it is for all nanoscale structures and fabrication methods, whether biological, chemical, or physical. The latest commercially available field-emission SEM instru-

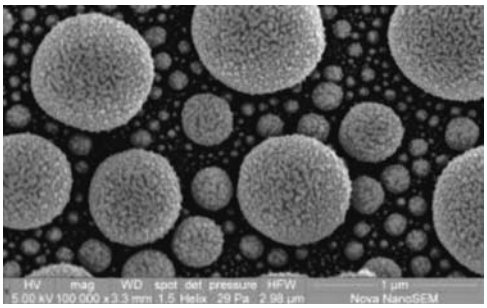


Figure 4.8 A test structure of tin particles, imaged using environmental SEM [43].

ments are capable of a resolution of less than 1 nm when used at high voltage, above 15 kV [42, 43]. High-voltage operation requires that the samples have a conducting surface, so as to remove the relatively large current deposited by the electron beam; consequently, such resolution is only possible when imaging stamp materials such as doped silicon or nickel, or resists made from conducting polymers, as might be used for EBL. The imaging of insulating materials such as quartz or glass (or indeed most of the polymers used as nanoimprint resists) at such high resolution generally requires a conductive coating layer of gold or other metal [44]. Uncoated small polymer structures can also be damaged by the electron beam current [45]; hence, in order to avoid damaging or altering the sample, the insulating surfaces can be imaged directly by using a low accelerating voltage for the electron beam, of approximately 1 kV or less [46], and using a partial vacuum with a low water vapor pressure to help remove the charge. This method, referred to as "environmental SEM" (Figure 4.8) [47], is the most common method by which nanoimprinted structures and processes are characterized [48].

The measurement of width-critical dimensions with SEM (or CD-SEM) requires the use of image-processing algorithms to assign a definite boundary to the gray edges of imaged features [49]. This is especially difficult if the edges themselves are sloped, although it is possible to assign a top and bottom width to features [50].

In order to measure thicknesses or 3-D profiles with SEM, it is necessary to make destructive cross-sections through samples [51]. Whilst this is performed routinely to make use of the excellent resolution of SEM, it leads inevitably to the waste of sample materials.

By using methods developed in the semiconductor fabrication industry, a large number of individual imprints and wafers have been tested, and statistical analyses applied to characterize the reliability of NIL as a large-scale nanofabrication method. This was typified by a recent study conducted at Sematech, the semiconductor industry research and development agency of the USA [52]. Results obtained using a Molecular Imprints Inc. Imprio300 instrument are summarized in Table 4.5. In this case, 300 mm wafers were imprinted with a stamp with a patterned area, or field, of 32×26 mm, and measurements made at five locations within each field (upper and lower left and right, and center), of the

Table 4.5 SEM results of thickness measurements of imprinted lines.

		Line 2	Line 4	Line 6
Lower Left	Average	32.4	33.7	31.9
	Min	30.1	30.9	30.2
	Max	33.3	34.7	32.8
	Range	3.2	3.8	2.6
	3 Sigma	1.5	1.6	1.3
Upper Left	Average	33.2	34.5	32.5
	Min	31.4	32.6	31
	Max	33.9	35.5	33.4
	Range	2.6	2.9	2.5
	3 Sigma	1.2	1.2	1.1
Center	Average	30.8	33	32.3
	Min	29	31.5	30.9
	Max	31.6	33.7	32.8
	Range	2.6	2.2	1.9
	3 Sigma	1.1	1	1
Lower Right	Average	32.1	33.7	32
	Min	30.7	32.1	30.4
	Max	32.8	34.4	32.6
	Range	2.1	2.3	2.2
	3 Sigma	1	1.1	0.9
Upper Right	Average	31	30.3	32.8
	Min	29.6	29	31.5
	Max	31.7	31	33.6
	Range	2.1	2	2.1
	3 Sigma	0.9	1	0.9

same three lines (Lines 2, 4, and 6). Of 2400 individual lines measured, across five wafers, the average width recorded was 32.4 nm, with three standard deviations (3σ) of 3.5 nm.

Whilst this represents an excellent degree of uniformity in fabrication, it must be considered in light of the very high standards required for the semiconductor industry. The International Technology Roadmap for Semiconductors (ITRS) sets a requirement of 3σ uniformity of less than 12% of the minimum feature size; hence, for the 32 nm node (which is due to be achieved by 2011 for Flash gate half pitch [53]) there is a uniformity requirement that 99.7% of all structures must be within ± 3.8 nm of the 32 nm designed critical dimension. By this criterion, the Sematech NIL results are only just acceptable.

The ability to measure such critical dimension uniformity (0.32 nm) places a requirement on any metrology technique for a measurement repeatability of 10% of this (i.e., 0.32 nm) [54], which is approximately at the limit of what can be achieved by using SEM and scatterometry at present [55, 56]. The need for a metrology tool

uncertainty of less than 0.3 nm remains an ongoing challenge, with metrology being pushed closely by the demands of reducing lithography feature sizes.

4.3.3 Atomic Force Microscopy

Atomic force microscopy is probably the most inexpensive method used to obtain nanoscale measurements, with instrument prices beginning at about €20 000 for the simplest designs. Hence, since its invention at IBM in 1986 [57], AFM has become a ubiquitous feature of laboratories. AFM is a powerful method that can be used to map the surface topography of a sample by scanning a probe across the sample's surface. The probe is held by a flexible cantilever, the deflection of which can be monitored, and this allows the relative height of the surface to be recorded. Since the displacement of the probe from the surface is known through the atomic forces between them, if the interaction with the surface is limited to one atom at the apex of the probe tip, then the atomic resolution of the surface can be determined [58].

The method is nondestructive and can be used on any type of surface, from conducting to insulating; moreover, atomic-scale resolution is possible even when used in air [59]. Because the atomic force between the probe and surface can be monitored, and is characteristic of the material present, it is possible to obtain physical, chemical, and dimensional information about the surface. As well as chemically distinguishing individual atoms of a certain material on the surface [60], it is also possible to measure frictional forces, by dragging the probe across the surface, lateral to the cantilever [61]; moreover, by pushing the probe tip into the surface and then withdrawing it, the viscoelastic properties of the material can be calculated from the force versus displacement curve [62]. This is of particular relevance to thermal nanoimprinting and nanoindentation (these are discussed further in Section 4.3.7).

While AFM is capable of extremely high spatial resolution, one drawback of the method is that it is relatively slow, and so is normally limited to imaging only a small part of the sample surface. The speed of the method is limited by the fact that a probe tip must be physically scanned across the surface while maintaining good contact; this allows an image to be built up serially by rastering back and forth across the surface. However, this cannot be achieved as quickly as scanning with an electron beam, or when optical methods, which sample an area of the surface in parallel. As a consequence, AFM has often taken the role of a reference metrology, and used to calibrate and improve the accuracy of other methods (e.g., SEM or scatterometry) being used to characterize a whole wafer [63].

One weakness of AFM is its limited ability to measure the width of high-aspect ratio structures. This stems from the shape of the probe tip, which is roughly pyramidal, and means that whilst at the probe apex one atom may interact with a flat surface, the increasing width of the probe away from the tip prevents access to narrow gaps and corners of features. Currently, this problem can be overcome by extending the probe tip with very high-aspect ratio structures such as carbon nanotubes (CNTs)

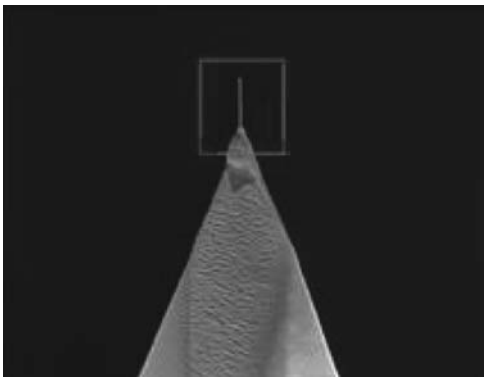


Figure 4.9 A carbon nanotube attached to the tip of an AFM probe [65].

(Figure 4.9) [64]. Curved and flared tips have also been used to improve access to sidewalls [65], while tips can also be tilted with respect to the surface [66], so that accurate 3-D profiles can be generated.

4.3.4
Transmission Electron Microscopy

Although transmission electron microscopy (TEM) is another extremely powerful technique, capable of atomic resolution [67], it requires samples which are thin enough to allow electrons to pass through (this is normally $<$100 nm, depending on the material). For this reason, the sample preparation – which normally involves cutting a slice of material with a focused ion beam – means that the technique takes a long time, and is destructive [68]. For this reason, TEM is also mainly used in nanofabrication to provide cross-sections as a reference metrology for quicker methods, such as SEM, scatterometry [69], and AFM [70].

4.3.5
Optical Critical Dimension Metrology: Scatterometry

The diffraction limit of light sets a lower bound on the size of what can be imaged optically to approximately half the wavelength of the light used. However, if the properties of the incident light are changed by the structure being investigated, then an image can be recreated from the reflected or scattered light, with resolution far below the diffraction limit [71]. Scatterometry uses the optical signal from an ellipsometer, applied to periodic structures, in which the change of the polarization of the scattered light depends upon the size and shape of the lines or other structures comprising it.

Ellipsometry is a one-dimensional method, used to measure the thickness and optical properties of planar layers, with a resolution of Angstroms, by analyzing the change of polarization of light reflected from the sample. Two curves are generated,

based on the S and P polarizations of the reflected light (oscillating perpendicular and parallel to the plane of incidence) as a function either of wavelength, angle, or incident polarization. The curves generated by optical modeling are fitted to these in order to extract the data of layer thicknesses and complex refractive indices [72]. Ellipsometry is often used to establish the refractive index of flat samples of a material, to be used as input data for scatterometry measurements of patterned samples of the same material.

Reflectometry is another optical metrology, in which the light is incident normally on the sample, and the thickness is calculated from the intensity of the reflected light, which depends upon constructive or destructive interference, depending on the layer thickness [73]. Most optical metrology tool manufacturers provide versions of all three techniques [74, 75].

As with optical methods generally, *scatterometry* is both noncontact and nondestructive in nature. It is capable of measuring line widths as small as 20 nm with a resolution of less than 1 nm, depending on how many independent, or floating, variables there are in the analysis [76]. Scatterometry is an indirect measurement technique, in two important ways. Since it analyzes diffracted light, it is usually necessary to fabricate periodic test structures, alongside the structures to be characterized, which usually are not sufficiently periodic themselves [77], and dimensional data are taken from these test structures. Also, the structure is not imaged directly, rather, the dimensional parameters defining it are calculated by a comparison of the measured signal with simulated signals. This usually involves matching the measured data to a library of previously generated data sets, which can take from several seconds to minutes to complete [78]. However, in an effort to speed up the matching process, a variety of mathematical solutions, including neural networks, have been developed [79].

Scatterometry is a genuinely 3-D measurement method, in that the width and height critical dimensions, as well as line shape profile can be measured at the same time, without altering or damaging the sample. This has led to it becoming a reliable method for measuring the width of lines below 50 nm, because at this scale, fabrication processes produce lines the degree of sidewall slope or curvature of which is significant compared to their average width, and so must be properly modeled to produce an accurate width result. In this way, more than five floating parameters can be solved simultaneously, allowing complex 3-D structures to be characterized in a single measurement [80] (Figure 4.10).

4.3.6
Other Methods

Other techniques under development include scatterfield and through-focus optical metrology, both of which use an optical microscope as a nanometrology tool. Scatterfield metrology can be described as a version of variable angle scatterometry, in which an angular scan is achieved by moving an aperture in the back focal plane of the microscope, so that light is incident and re-collected over a small angle through the objective lens [81]. This method is attractive not only

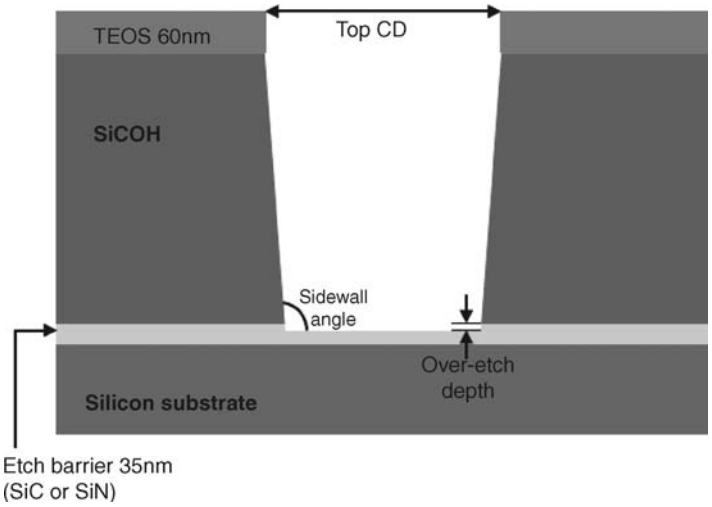


Figure 4.10 Multilevel structure measured using scatterometry [81]. See text for details.

because it opens the possibility of using alignment optics for metrology, but also because it enables the use of smaller grating test structures than are required for conventional scatterometry.

In the *through-focus method*, a series of images are recorded, using an optical microscope, of features below the diffraction limit, at different focal distances. The intensity profiles of these are combined to create a 2-D intensity map characteristic of the structures, with a sensitivity to line width changes of 2 nm [82]. One strength of this method is that structures do not have to be periodic in order to be measured.

The short wavelength of X-rays, from approximately 10 to 0.01 nm, can be used in a number of ways for metrology on the nanometer scale. X-ray diffraction (XRD) can be used to provide both 3-D and physical information, such as density and porosity [83]. In critical dimension-small-angle X-ray scattering (CD-SAXS), a beam of X-rays from a synchrotron source is incident normal to the sample, and the transmitted, diffraction pattern analyzed. The line width, pitch, height and line profile can be recreated, as well as the material density, upon which the diffraction of the X-rays also depends [84]. In specular X-ray reflectivity (SXR), X-rays from a Cu Kα source are incident at a grazing angle and the reflected diffracted signal is measured over a range of angles, using a goniometer [85]. With SXR, although the line to space ratio can be measured, in order to calculate the actual line width value it is necessary to have measured the period using another method.

X-ray imaging is currently being investigated for metrology, but is challenging as it requires a good collimated X-ray beam (typically from a synchrotron), and because the resolution is limited not by the wavelength but rather by half the minimum separation in the Fresnel zone plate lenses used to focus the beam. To date, the smallest separation to have been achieved is 20 nm [86].

Optical sub-wavelength diffraction is a technique which can be used to optically characterize structures as small as 50 nm by analyzing the far-field diffraction pattern of line gratings composed of sub-wavelength-sized features. Information can be obtained about the critical dimension, height, and the presence of any defects in the structures. The method uses grating test structures, which have a period greater than the diffraction limit. However, within each period there are features below the diffraction limit which do not affect the angle of diffracted orders, but do affect the relative diffraction efficiency of each order [87].

4.3.7 Physical Properties

A vital part of controlling the nanoimprint process is knowing the physical properties of the polymers used, including viscosity, Young's modulus, and the T_g. Since it is particularly important to measure these parameters at the nanoscale, where physical properties are often different to the bulk values, this has led to the development of existing techniques to measure such small samples.

Elastic modulus measurement at the nanometer scale can be performed by nanoindentation using an AFM tip on polymers [88], as well as other materials [58]. There is some uncertainty regarding this method, due to the difficulty of controlling the tip shape and radius; hence, it has not been used to characterize films of thickness less than 100 nm, and neither has any difference been recorded for Young's modulus measured by nanoindentation from films thicker than 100 nm, and bulk values.

Force and displacement measurements of polystyrene films on the nanometer scale have been made using nanoindentation techniques [89, 90]. These have been used to study the viscoelastic flow of polymers above their T_g under flat and patterned punches, which press the polymer from an initial thickness of 170 nm to a residual thickness of 20 nm, closely recreating the conditions of NIL. The viscoelastic flow of the polymer under a flat punch was found to behave according to bulk models in most cases. One effect of confinement was found for high-molecular-weight polymers (9 000 kDa), which occurs when the thickness of the film is reduced to less than the gyration radius of the polymer, which in this case was approximately 84 nm [91]. For such a thin layer, it was found that deformation of the polymer was accelerated, and this has been termed "confinement thinning."

The most commonly observed effect due to nanoscale dimensions in polymers is a change in the T_g. Typically, the T_g is indicated by a sudden increase in the rate of thermal expansion, taken from thickness measurements made using ellipsometry. For free-standing polystyrene films, a decrease in T_g of up to 80 K has been measured for 20 nm-thick films, compared to the bulk value [92]. The onset in the reduction of T_g begins at a thickness less than approximately 90 nm.

The elastic modulus of polymer structures as thin as 80 nm has been measured by studying the acoustic modes via Brillouin scattering [93]. This is a version of Raman scattering, but with a much higher resolution, achieved by using a Fabry–Perot cavity to resolve peaks within 1 nm of the excitation wavelength, which are characteristic of

the energies of phonons confined in nanoscale structures. The Brillouin spectra of lines of thickness 180 to 80 nm, created by deep UV lithography on silicon, were analyzed to derive the speed of sound in the structures, and from this the elastic modulus. No change in the physical properties was measured with respect to bulk values for thicknesses above 80 nm.

The *ultrashort laser pulse photoacoustic method* has been used to characterize the physical properties of layers of PMMA of thicknesses ranging from 586 to 13 nm, spin-coated onto Si wafers. Acoustic speeds, cp, calculated from time of flight and film thicknesses as measured by ellipsometry, were found to increase below approximately 80 nm, with an increase of 20% for a 13 nm sample, compared to the bulk value. This corresponds to an increase in Young's modulus of 44% [94]. The implications of this result for the NIL process are currently under investigation.

4.3.8
Residual Layer Thickness in NIL

Nanoimprint lithography has some features which are particular to it, and occur as a result of the process, including an underlying residual resist layer, and complex line shape profiles, due to the physical forces which deform the resist during and after imprinting.

In order to use the nanoimprinted resist layer for pattern transfer into the substrate, the residual layer must first be etched away. It is essential that the residual layer is uniform across the wafer in order for the fabricated structures to retain the same critical dimensions after this etching step. The most appropriate method to measure residual layer thickness is scatterometry, as it is nondestructive and enables characterization underneath the patterned layer through the transparent resist [95]. Alternatively, samples can be cleaved and a cross-section image made by using SEM, although as the imprinted samples are clearly then sacrificed this is most often used for process development, including methods to ensure a uniform residual layer, such as designing stamps with a uniform array of imprint features [96]. Where large (μm) and small (nm) features are together on the same stamp, the drop dispensing of a precise volume of resist to match the stamp feature sizes [97] has been developed.

Coarse-grain finite difference simulation software has been developed to model the nanoimprinting process over an extended area of several square millimeters, to predict the residual layer thickness, and a very close agreement was found with measured thicknesses [98].

The line shape profile produced by nanoimprinting has been studied using a variety of metrologies, although the most commonly used is scatterometry [99], due to its 3-D measurement capability. In order to enhance the effect of the residual stresses which remain in the polymer structures after imprinting, and which may cause them to deform over time after their release from the mold, one commonly used experiment has been to deliberately cause the lines to reflow by heating them outside of the mould. The resultant line shapes that evolve with time have been measured

using AFM [100], while optical diffraction [101] has been used to characterize when the onset of deformation occurs.

4.3.9 Towards An Integrated Metrology

Real-time measurements of polymer line reflow *in situ* of the heating stage have been performed using scatterometry, and compared with *ex situ* measurements of the frozen line shapes by using AFM and SXR, producing very similar results [102]. Each scatterometry measurement takes less than 12 s; however, similar scatterometry measurements, using an optimized library matching method to assign line profile shapes, reduced the individual measurement acquisition times to less than 10 s [103].

These measurements represent progress towards achieving an *in situ* monitoring of the nanofabrication process, which is part of the goal of an integrated metrology. As outlined in the ITRS Roadmap [104], with the reduction in minimum features below 32 nm, and the associated reduction in tolerances to a few nanometers, it is increasingly important to control small variations, or drift, (due to the environmental, instrument or materials) in the fabrication process parameters. This calls for close monitoring of the process, or Advanced Process Control (APC), which requires that metrology be performed either in line with the production, or *in situ*, in the process chamber, so that the time taken for corrective actions can be reduced, and the yield maintained. This is especially relevant to nanoimprinting, as any defect which may occur in the stamp will be replicated in every imprint, and so must be corrected as soon as possible.

Most work in this area has been focused on using optical techniques, as these are nondestructive, can be performed in any atmosphere, and are relatively fast. Scatterometry provides the most complete picture of the fabricated structure [105], while optical diffraction has been used to monitor mold-filling, and to provide a timescale of the imprinting process. However, this is only possible with line widths greater than the diffraction limit – usually approximately 200 nm, with pitches over 400 nm [106]. Recently, scatterometry has been used in optical lithography to make combined measurements of overlay, for double-patterning and critical dimensions [107]. What is a common theme in making use of the simplicity of optical diffraction, or of using the more complex scatterometry for combined measurements, is the need to ensure the maximum value for money from measurements. As integrated metrology requires the deployment of more instruments to monitor each step of the process, the cost is inevitably increased, and so this too requires careful measurement to ensure an efficient fabrication process.

From the above it is clear that new methods are needed, or existing ones need to be developed further, to be used in-line with suitable accuracy and speed. One major issue will inevitably relate to data handling with decreasing feature sizes under production conditions. Furthermore, nanometrology will have to be extended to applications which have tolerances of a few nanometers, and not necessarily in flat

but rather on curved substrates. Much remains to be developed in nanometrology in general, and in NIL in particular.

4.4
Two-Dimensional Nanopatterned Polymer Components for Photonic Applications

4.4.1
Overview on Applications Realized by NIL

Optical devices, such as displays, light-emitting components, polarizers, and anti-reflective coatings are used in everyday life in a variety of applications. Currently, NIL offers the capability of a cost-efficient large-scale patterning for polymer photonic components with the requested high resolution and high throughput for such structures. As with common lithography techniques, NIL provides the same tremendous number of applications, which could be separated in two categories: (i) the use of the nanoimprinted polymers as an etch mask to be transferred into the substrate; and (ii) the direct use of the nanoimprinted polymer as devices or components. In both cases, high throughput and high resolution in the nanoscale range are requested over large areas.

With regards to the first category of applications, when using standard NIL techniques it becomes necessary to etch the thin residual layer before transferring the nanostructures into the substrate. This additional step is not critical, and the number of process steps is usually reduced by the use of 3-D imprinting in comparison with standard lithography techniques (see Section 4.2). Moreover, NIL enables potentially relatively cost-efficient nanopatterning if compared to next-generation optical lithography technologies such as deep UV or extreme UV lithography. The latter requires extremely high manufacturing volumes to be economically viable, and it was for this reason that NIL was included in the semiconductor industry roadmap as a possible next-generation lithography to deliver the 32 nm node and beyond [108]. This alternative fabrication method of nanoimprinted mask transfer has already provided applications examples in the fabrication of surface-acoustic-wave generators and filters for mobile phones [109], patterned media for hard-disks [110], and as sub-wavelength polarizers for displays [111]. Although, NIL is currently not limited in resolution for such applications, the main technological issues are the overlay accuracy, defectivity, and critical dimension control (cf. Section 4.3).

The second category of applications concerns the direct patterning in a single step of polymer structures as end products. Nanoscale-patterned polymer films could be tentatively organized into sub-applications categories: the fabrication of organic light-emitting diodes (OLEDs) and plastics electronics; the realization of templates and polymer stamps; biotechnology, including tissue engineering lab-on-a-chip and cell studies; surface modification combined with self-assembled techniques; and 3-D polymer surfaces and the generation of optical components.

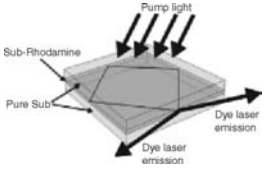
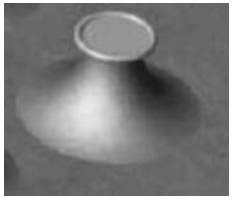
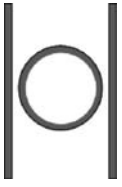
In this section, the recent advances in the realization of nanopatterned polymer devices for light control are highlighted. Following by a review of the different optical resonators produced with NIL, examples are provided of optical resonators for lasing applications. Attention is then focused on the use of NIL to pattern functionalized polymers [112]. Active materials can be directly patterned to create lasers [113–115], organic light-emitting devices [116, 117] and conductive organic polymers to realize cost-efficient organics electronics [118]. Specific examples of the direct imprint of nanocomposite polymers to control the emission of light from polymer films will be provided, after which a brief description of nanoimprinted polymer devices will conclude this overview of nanoimprinted photonic components.

4.4.2 Nanoimprinted Optical Resonators

The NIL technique is particularly suitable for the fabrication of integrated polymer optical devices, due to its high resolution and parallel processing of polymer layers. Furthermore, NIL can deliver a surface roughness compatible with the demands of light guiding. Polymer waveguide-type wavelength filters based on a Bragg grating [119, 120], waveguides [121], microring resonator [122, 123], Mach–Zehnder interferometers [124], lasers [125–129], plasmonic components [130, 131], and photonic crystals [132–136] have been recently realized using NIL, and demonstrating its potential as a high-volume and cost-effective patterning technique with sub-10 nm resolution. Optical resonators represent the key components for the spatial confinement and control of light, and the different approaches for creating polymer optical resonators via NIL, associated with their high-quality factors (when applicable), are listed in Table 4.6. (Note: A high-quality factor indicates a long lifetime of the photons in the resonator, leading to a sharp wavelength selection.)

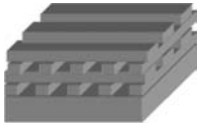
Different optical resonator types can be allocated roughly to three types to achieve optical feedback: (i) specular reflection [137–141]; (ii) ring resonators [142]; and (iii) periodic structures [143–146] (gratings/photonic crystals). Light is coupled in/out to optical resonators in several ways: evanescent field coupling via random scattering from roughness (specular reflection); via a closely spaced waveguides (specular reflection) and standard input/output coupling through gratings (photonic crystals); or by photo-pumping of the imprinted active media, and re-emission. The feedback of resonators based on specular reflection is ensured by reflections on the sidewalls, with incidence angles above the critical angle for total internal reflection. This type of feedback is based on plane waves propagation to minimize losses on reflection, which limits the minimum resonator size to $100 \times 100\,\mu m^2$. There is no obvious input/output coupling mechanism for such resonators, except by evanescent field coupling which can achieved by bringing a waveguide in close proximity ($<\sim$200 nm) to the resonator. The incident angle on one of the sidewalls can be reduced below the critical angle by using a trapezoidal shape instead of a square; coupling-out of the light is then realized, showing multimode laser emission, although the magnitude of

Table 4.6 Polymer optical resonators in polymer fabricated by NIL.

Method	Typical structures	In/Out coupling	Wave-length selectivity	*Q* factor	Advantage	Drawback
Specular reflection	 Refs [137–139]	Evanescent	No	NA	Relatively easy fabrication	No wavelength selectivity
	 Ref. [140, 141]	Evanescent	Yes	5×10^6	Relatively easy fabrication	No wavelength selectivity
Wave-guide	 Ref. [142]	Evanescent	Yes	5800	Relatively easy fabrication	Difficult to optimize Wavelength selectivity

(*Continued*)

Table 4.6 (*Continued*)

Method	Typical structures	In/Out coupling	Wave-length selectivity	*Q* factor	Advantage	Drawback
1-D grating	 Refs [143, 144]	Built in	Yes	-[a]	Relatively easy fabrication Wave-length selectivity	Light control: 1-D
2-D grating	 Ref. [145]	Built in	Yes	1020[a]	Relatively easy fabrication	Light control: 2-D
3-D grating	 Refs [146, 147]	Built in	Yes	—	Total control of the light	Difficult fabrication Alignment issue

a) Limited by the resolution of the spectrometer.

the output coupling is difficult to control. A ring resonator coupled to waveguides [144] acts as Fabry–Pérot showing a *Q*-factor of 5800, and the device size is limited by the bending loss of the waveguide. The quality of the imprints between the ring and the waveguides is critical to achieve high *Q*-factors. In order to achieve high *Q*-values with specular resonators, polymer-imprinted toroidal resonators have been successfully created by using extremely smooth polymer surfaces. Material-limited *Q*-values of up to 5×10^6 have been measured by bringing a waveguide within close proximity to imprinted toroidal resonators elevated above the substrate. Unfortunately, however, exposure of the surfaces to ambient conditions may cause the resonators to be degraded over time; consequently, it is usually necessary to shield the resonators.

Photonic crystals can be used to produce compacter resonators, with optical feedback being achieved with only a few tens of periods. The wavelength selection is realized by periodically varying the refractive index of the patterned materials (typically a fraction of the optical wavelength); such variation in turn induces a modification of the dispersion relation of propagating modes in the material, and this may lead to the generation of slow group velocity modes and to the opening of photonic band gaps (as the well-known Bragg grating). Photonic bandgaps result in optical feedback in 1-, 2- and 3-D directions. A typical cross-section of a nanoimprinted Bragg grating with 350 nm pitch is shown in Figure 4.11a. Features from the silicon stamp are usually very well reproduced in polymer layers with adequate imprinting parameters; for example, a lines separation of 60 nm is shown in Figure 4.11, and the residual layer is about 120 nm thick in this case. Such structures, known as "Bragg gratings," can be used efficiently to create nanoimprinted polymer optical resonators and to build in the coupling of the light. These optical resonators can be integrated with waveguides to couple in/out light of the resonators in planar geometries. In fact, planar polymer-based resonators can operate by using waveguiding in the polymer. Imprinted waveguides, beam splitters and bent waveguides can be easily obtained using standard imprinting processes [148]. Figure 4.11b and c show current optical microscope images of such structures, where the uniform Newton colors indicate a uniform residual layer. Polymer interferometers can be made with 800 nm separation between the two waveguides (Figure 4.11d and e).

In order to achieve 2-D optical feedback, 2-D periodic structures – called 2-D photonic crystals (PhCs) – are formed, for example, with holes in the imprinted polymer [127]; the light is then controlled in the plane of the substrate. Other advantages of 2-D defect-free PhCs over conventional 1-D feedback gratings include the potentially highly directional vertical emission and a lower lasing threshold. Ideally, 3-D periodic structures which allow the control of the light in the three spatial directions are required, but such structures are particularly complex to fabricate using either holographic lithography [149] or serial layer stacking [150] or self-assembled techniques.

At present, nanoimprinted optical resonators that combine the best control of light in/out coupling, wavelength selection and easiest fabrication by NIL are the 1-D and 2-D grating resonators. Smooth polymer surfaces are required to avoid random

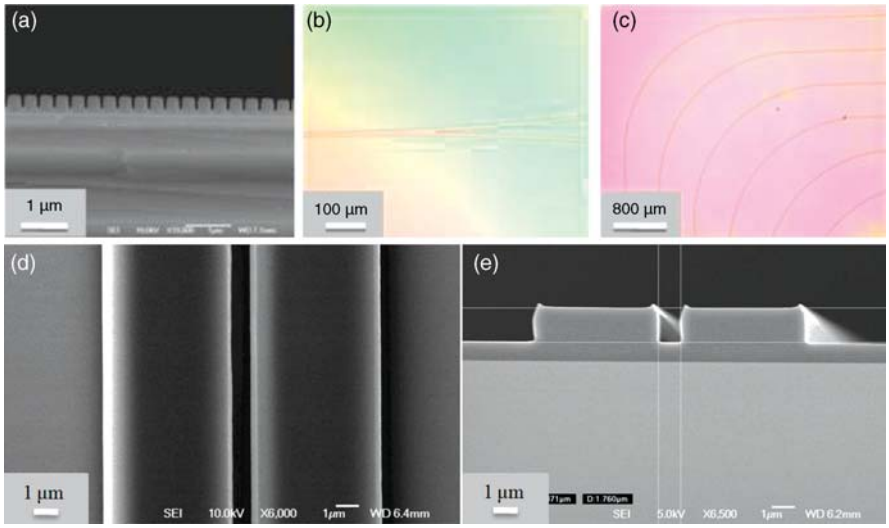


Figure 4.11 (a) SEM cross-section of imprinted polymer grating (pitch: 350 nm, lines separation: 60 nm, residual layer: 120 nm); (b) Optical microscope image of bent waveguides in mr-NIL 6000; (c) Optical microscope image of imprinted beam splitter in mr-NIL 6000; (d) Top-view; (e) Cross-sectional SEM images of the two interfering waveguides.

scattering and to reach respectable *Q*-factors. Nanoimprint lithography is well suited for the creation of such polymeric structures, since the low degree of roughness of silicon stamps is transferred to the polymers (the transferred roughness can be reduced even further by a control reflow of the polymer [151].).

4.4.2.1 **Nanoimprinted Polymeric Band-Edge Lasers**

Polymer dye lasers might provide compact and inexpensive coherent light sources for microfluidics and integrated optics in the visible range. The advantages of 2-D defect-free PhCs over the conventional 1-D feedback gratings are, potentially, a less-directional vertical emission and lower lasing thresholds. Such PhCs have been studied to create band-edge lasers in organic media [152, 153], and have shown their potential of integration in planar optical circuits [154]. However, one reason why 2-D PhCs have not yet been produced in mass numbers is that expensive techniques, such as EBL or laser interference lithography, are required for their fabrication. The latter approach offers a higher throughput than the former, but does not yet meet the requirements for mass production [155]. A series of PhCs was directly fabricated in a printable polymer loaded with dye-emitting molecules, and showed lasing oscillation at different photonic band-edge frequencies. For these nanoimprinted band-edge lasers, a dye-doped polymer was used, composed of rhodamine 6G (R6G; from Sigma-Aldrich) directly dissolved at a concentration of 2.5×10^{-3} mol l^{-1} in the polymer mr-NIL 6000 (from Microresist Technology). It is known that organic emitters degrade when exposed to air at high temperatures, and this results in low light emission efficiency. To minimize the thermal degradation, mr-NIL 6000 was

chosen because of its relative low T_g (ca. 45 °C). The active films were spun on a glass substrate at 1500 r.p.m. for 1 min and then baked at 115 °C for 5 min to remove the residual solvent. The measured polymer thickness was 400 nm.

The imprinting process was performed in a 6 cm (2.5 inch) Obducat nanoimprinting machine at 90 °C and under 60 bar pressure for 5 min. After sample cooling, the stamp was separated from the patterned polymer at 40 °C. A reduction of less than 3% in the photoluminescence (PL) intensity was found after patterning with an unstructured silicon stamp, while the refractive index of the doped polymer was measured by ellipsometry to be 1.614 at 550 nm. Stamps with the two lattice constants (460 and 500 nm) have been produced using EBL (Jeol 6000) with a dose of 130 μC cm^{-2} and a beam current of 100 pA, using a 150 nm-thick layer of ZEP 520 resist (Zeon) that had been pre-baked at 120 °C and developed for 30 s in a solution of ZED N50 (Zeon). The silicon stamp was etched 350 nm deep using an inductively coupled plasma (ICP) reactive ion etching system (Surface Technology Systems), with a mixture of SF_6 and C_4F_8 gases. The stamp was subsequently coated with an anti-adhesive monolayer (tridecafluor-1,1,2,2-tetrahydro-octyl trichlorosilane) deposited from the vapor phase, which results in a very low surface energy to facilitate detachment of the stamp from the polymer. The nanoimprinted samples were optically pumped in a vacuum cell with a 0.7 ns frequency-doubled Q-switched Nd : YO4 laser light at 532 nm focused to a 50 μm-diameter spot on the sample surface.

Figure 4.12a shows the SEM image of a silicon stamp and a nanoimprinted photonic crystal in the dye-chromophore-loaded polymer matrix. The 2-D pillar arrays of the silicon stamp were faithfully transferred onto the modified polymer. Figure 4.12b shows the measured emission spectrum of the 2-D PhC with 460 and 500 nm lattice constants. The reduced frequencies were measured as 0.765 ± 0.001 and 0.876 ± 0.001 for the 460 and 500 nm PhCs, respectively. Using a plane-wave algorithm, the reduced frequencies of the three band edges, Γ_1, X_4, were calculated as 0.766 and 0.877, respectively, yielding a very good agreement between the numerical and experimental band edge frequencies. The laser with the 460 nm lattice constant operated at 601.4 ± 0.3 nm, approximately 40 nm away from the maximum of the spontaneous emission peak, and indicating a strong optical feedback provided by the

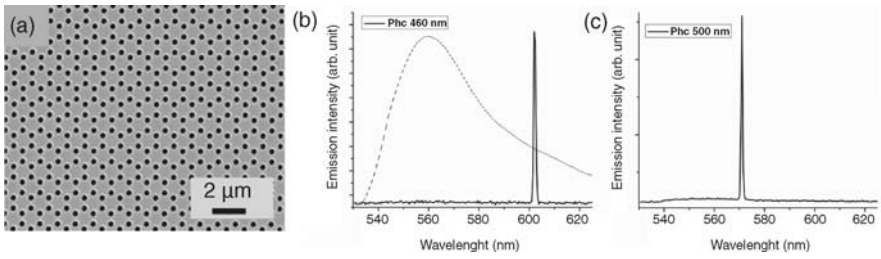


Figure 4.12 (a) SEM image of an imprint in the composite polymer mr-NIL 6000 with rhodamine 6G; (b) Emission spectra of a band edge laser with lattice constants of 460 nm and 500 nm. The dotted line shows the emission spectra of rhodamine 6G in mr-NIL 6000 below the threshold.

Table 4.7 Comparison of 1-D nanoimprinted R6G-doped polymer lasers.

Reference	Threshold ($\mu J\,mm^{-2}$)	Area (μm^2)
[000] (Li *et al.*, 2006)	8	200×4
[156]	8	250×1000
[000] (Pisignano *et al.*, 2004)	6.5	200×200

PhC. The laser surface was 125-fold smaller than those measured for the 1-D nanoimprinted organic feedback grating, made from R6G-doped SU-8 polymer [156]. A laser threshold of $3\,\mu J\,mm^{-2}$ was obtained. A 1-D nanoimprinted laser doped with R6G is shown for comparison in Table 4.7. No lasing oscillation was observed for areas without a PhC pattern under the same pumping conditions. These results indicated that the direct imprint of a spin-coated, dye-doped polymer with a PhC stamp could be used as a band-edge laser with a relative low degradation of the gain property of the dye.

The lifetime of the lasers as function of the number of excitation pump pulse at $8.5\,\mu J\,mm^{-2}$ (corresponding to 2.8-fold the laser threshold) was also measured. The laser emission fell exponentially to 10% of its initial values after 8200 pulses, which was comparable to solid-state dye lasers emitting in the visible wavelengths. In order to improve the lifetime of the lasers, one solution might be to replace the dye molecules with semiconductor nanocrystals with optical gain embedded in a printable polymer [137].

4.4.3
Patterning in Functionalized Polymers

NIL presents the unique advantage to pattern polymers in a single step to realize, for example, optical components. In addition, unsurpassed size- and shape-dependent electronic properties of semiconductor and metal nanocrystals (NCs) are extremely attractive as novel structural building blocks for the construction of a new generation of innovative materials and solid-state devices. Recent advances in chemical synthesis have resulted in colloidal NCs with a wide range of compositions, combined with an excellent control of size, shape, and uniformity. As the surfaces of the NCs can be easily engineered by ligand exchange and surface functionalization, they can be placed in almost any chemical environment, although their use in devices for photonic and sensing applications normally requires them to be incorporated into a polymer matrix, in order that their properties can be exploited. A strategy for incorporating NCs into printable polymers, to allow their homogeneous distribution in the polymer matrix, is shown in Figure 4.13a. In this case, the pre-synthesized TOPO-coated (CdSe)ZnS core-shell luminescent nanocrystals are incorporated into a PMMA homopolymer and into a PMMA-based co-polymer. The functionalized polymer acts as a stabilizing and protective layer surrounding every single NC (see Figure 4.13b and c). The

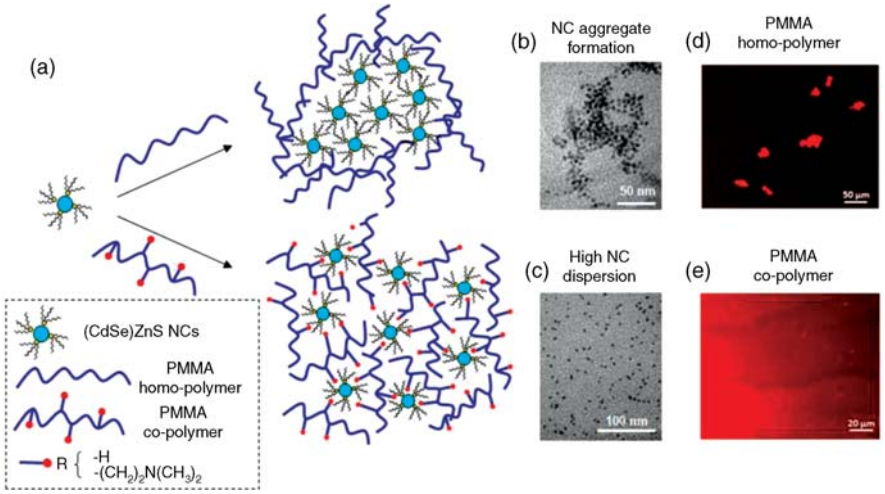


Figure 4.13 (a) Schematic illustration of TOPO-coated (CdSe)ZnS nanocomposites (NCs) in PMMA homopolymer and in functionalized PMMA based co-polymers; (b) TEM images of NCs, revealing aggregation in the case of PMMA homopolymer; (c) High NC dispersion when embedded in functionalized co-polymers; (d, e) Fluorescence microscopy images of (CdSe)ZnS NCs incorporated in (d) PMMA homopolymer and (e) PMMA block co-polymer ($PMMA_{70}$-*co*-$DMEAMA_{30}$). Reproduced from Ref. [134].

fluorescence microscopy image (Figure 4.13d) with the PMMA homopolymer shows a strong aggregation of the NCs whilst, in the case of the PMMA co-polymer block, a homogeneous emission over the whole surface of the polymer film is obtained, showing a high compatibility of the NCs with the host polymer. (The experimental details of these studies can be obtained from Ref. [135].)

The fabrication of 2-D patterned light-emitting structures in this luminescent nanocomposite, using NIL, has been demonstrated. A similar imprint process as that described above has been used to pattern the nanocomposite polymer. Figure 4.14a

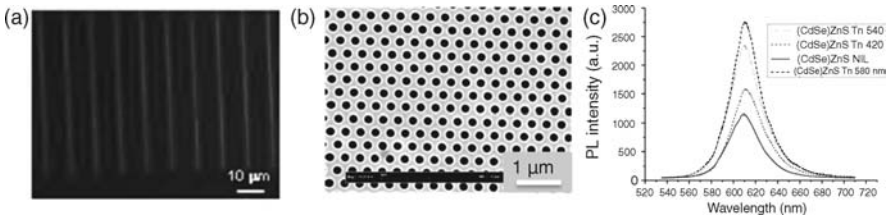


Figure 4.14 (a) Fluorescence image of patterned waveguides obtained by NIL onto the nanocomposite thin films; (b) SEM images of nanoimprinted photonic crystals in a PMMA-based polymer matrix, into which (CdSe)ZnS NCs have been incorporated; (c) PL spectra of unpatterned PMMA-based polymer matrix with (CdSe)ZnS NCs on a Pyrex substrate (solid line) imprinted with a flat stamp, and PL spectra of a 2-D photonic crystal with a 580 nm lattice (dash-dot line). Panels (a) and (b) reproduced from Ref. [134].

shows the fluorescence image of nanoimprinted waveguides, and Figure 4.14b the SEM images of nanoimprinted photonic crystals onto the nanocomposite polymer. The imprinted pattern transfer did not suffer from any unusual roughness or degradation, which meant that the incorporation of (CdSe)ZnS had not caused any deterioration of the duplication fidelity of the stamp in the photoluminescent polymer. The nanoimprinted PhCs could then be optically excited using a CW Ar^+ laser at a wavelength of 514.5 nm, with a power of 240 μW focused down to a 10 μm spot, and with the incident beam normal to the surface. The collection cone was defined by a 10× microscope objective with a numerical aperture (NA) of 0.4. The photoluminescence of samples was collected through the same objective of microscope, and analyzed spectrometrically. First, no degradation of the photoluminescence intensity of NCs was observed after patterning by NIL. The ratio between the PL intensities (integrated over 530–750 nm) of the patterned structure and the unpatterned imprinted surface has been measured: the maximum enhancement factor obtained was equal to 2.4 at room temperature for the lattice constant of 580 nm. This extraction enhancement demonstrated that the nanoimprinted photonic crystal slab structure could represent a potential candidate for high-efficiency light-emitting diodes (LEDs), based on polymers.

Whilst it is clear that one solution to solving the light-trapping problem consists of using 2-D photonic crystals fabricated by NIL, another approach would be to increase the spontaneous recombination rate of the emitters. This can be based on the energy transfer between light emitters and surface plasmons (SPs). The results of several investigations of enhanced light emission via SPs have been described [157, 158]. Recently, comparable studies have been conducted with dye-doped organic films [159] and with conjugated polymer films [160] in close proximity to metal surfaces. Similar coupling processes have been observed in both organic and inorganic structures, with important enhancements in the spontaneous emission intensity of the emitters. The two approaches mentioned above can be combined to enhance the light-emission efficiency of organic thin films. As an example, the fabrication of printed nanostructures using a thermoplastic polymer was studied, into which R6G and gold nanoparticles were incorporated, using a previously described strategy [134] to incorporate the nanoparticles. The water-soluble gold nanoparticles were synthesized via a non-seeding method [161], the gold nanoparticle size having been chosen to be close to the plasmon resonance wavelength of the nanoparticles with the emission wavelength of the dye. The R6G was diluted in 1 ml of mr-I PMMA 75 k 300, with a concentration of dye molecules of 5×10^{-4} M. The peak emission of R6G in the polymer was measured at 550 nm, and small amounts of gold nanoparticles in toluene solution were added to the polymer–dye emitting composite (Figure 4.15a). Several nanoparticles concentrations were tested, from 0 to 9.76×10^{-5} M.

Samples of the mixtures of dye with Au nanoparticles in mr-I PMMA (now called the "functionalized polymer") were placed in quartz cuvettes, and the spontaneous emission spectra (obtained by exciting the functionalized polymer at 450 nm) were collected perpendicular to the excitation. In order to pattern the functionalized

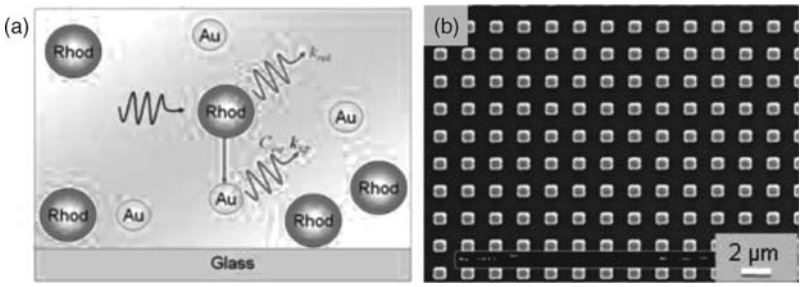


Figure 4.15 (a) Schematics of the coupling between metallic nanoparticles and dye emitters; (b) SEM image of a nanoimprinted grating in mr-I PMMA, in which R6G and Au nanoparticles have been incorporated.

polymer by NIL, the same process as described above was followed. The stamp and polymer were brought into contact at a temperature of 170 °C, and a pressure of 60 bar was applied for 5 min. Separation of the stamp and substrate was performed at 90 °C.

The coupling of SPs to emitters has been tested with metallic nanoparticles. For this, small amounts of Au nanoparticles were added to the solution of R6G in mr-I PMMA, increasing the concentration of nanoparticles in the polymer from 0 to 9.76×10^{-5} M. The PL intensity was then recorded from samples with different nanoparticle concentrations in the mixture. Measurements of the emission intensity, integrated over the range 500 to 750 nm, indicated an increase of up to 175% in the emission of the dye when the Au nanoparticles concentration reached 3.82×10^{-5} M. This enhancement could be attributed to an increased absorption and emission of R6G in the presence of metallic nanoparticles. For an Au nanoparticles concentration below 9×10^{-5} M, the number of absorbed R6G molecules per metallic nanoparticle was expected to exceed one, whereby the gain of the dye compensated for the loss in the localized SPs.

At higher concentrations of Au nanoparticles in the functionalized polymer, the PL intensity decreased due to a too-high number of Au nanoparticles in solution, the losses of which – due to the localized SPS – were no longer compensated. However, the good printability of PMMA as a thermoplastic material confirmed the use of prepared nanocomposite materials for the fabrication of devices for plasmonic applications. In these studies, grating structures and plain films were patterned in the functionalized polymer by using NIL. The SEM image of a nanoimprinted grating in the nanocomposite thin film is shown in Figure 4.15b. The results obtained may lead to a new approach towards new plasmonics devices fabricated using NIL.

4.4.4
Outlook on Nanoimprinted Polymer Devices

It has been seen that NIL is well-suited process to the fabrication of polymeric optical resonators, challenging photonic structures, and polymer devices, combining the

unique advantages of patterning with high resolution on large areas, and with a high throughput. Nonetheless, further developments are required in this direction to produce full photonic integrated circuits at low cost, with high degrees of functionality, and also to monitor the lifetime of such imprinted optical components. These studies have been conducted in part to demonstrate how NIL might be developed on 200 mm wafers to allow the fabrication of two optical devices, namely optical encoders [162] and double-sided organic LEDs with enhanced light extraction efficiency [3]. Both, photonic devices [136, 137] and plasmonic [3] devices can be directly fabricated by NIL, using the same materials during the same fabrication step with different functionalities [163]; alternatively, they may be separately fabricated and post-assembled. In addition, roll-to-roll, step-and-flash and step-and-stamp techniques have demonstrated huge potential for mass production, with the roll-to-roll approach having recently been used to fabricate fluidic platforms and organic LEDs [3]. One important area of activity that has developed during recent years has involved electronic printing devices, the target being to create printing methods that can be used for large-area flexible electronic devices [164]. High-resolution patterning techniques have been developed to provide a nanometer-scale separation between the source and drain electrodes of transistors in plastic circuits [165], while NIL methods have proven their ability in the fabrication of polymer devices and photonic components. Yet, there remain three main issues to be solved before NIL can meet the stringent requirements of mass manufacturing: (i) to minimize the number of defects per square centimeter; (ii) to improve the overlay accuracy (a 50 nm overlay accuracy was achieved by using interferometric *in situ* alignment techniques [166]); and (iii) to create a production throughput that would permit many hundreds of wafers to be printed on an hourly basis.

4.5 Conclusions

In reviewing the status of 3-D NIL, the associated nanometrology, and its optical applications, 3-D NIL will surely become increasingly important as novel and highly functional nanostructures are achieved that combine polymers, optically active, and biological materials. Moreover, these methods are nonexhaustive, since magnetic, nonlinear optical and other properties can be utilized, depending on the doping of the printable polymer.

The crucial role of nanometrology in NIL has also been recognized, and how it can be used to unravel nanorheological aspects when polymer thicknesses are on the order of only a few tens of nanometers. The need for new methods was clearly illustrated, as was the need not only to integrate nanometrology into a manufacturing environment, but also to take into consideration the vast amount of data that will be generated via in-line monitoring.

Clearly, the application of NIL to photonics is simply the "tip of the iceberg." By using periodic patterning, it should be possible to enhance certain interactions,

such as the slow modes of a photonic crystals and the emitted light from dye chromophores. It is also likely that the interplay between nanophotonic concepts and dispersion relationships in solids for practical polymer photonics structures will be clarified in the very near future.

Acknowledgments

The authors gratefully acknowledge the support of the EC-funded project NaPaNIL (NMP2-LA-2008-214249). The content of these studies is the sole responsibility of the authors.

Mads Brokner Christiansen and Anders Kristensen are sincerely acknowledged for their helpful insight on nanoimprinted optical resonators.

References

1 Sotomayor Torres, C.M. (ed.) (2003) *Alternative Lithography: Unleashing the Potentials of Nanotechnology*, Kluwer Academic Plenum Publishers, New York.
2 http://pubict.itrs.net.
3 Ahopelto, J. and Schift, H. (2008) Library of Processes, www.NAPANIL.org.
4 Chou, S.Y., Krauss, P.R., and Renstrom, P.J. (1995) *Appl. Phys. Lett.*, **67** (21), 3114.
5 Chou, S.Y. and Krauss, P.R. (1997) *Microelectron. Eng.*, **35**, 237–240.
6 Schift, H., Saxer, S., Park, S., Padeste, C., Pieles, U., and Gobrecht, J. (2005) *Nanotechnology*, **16**, 171–175.
7 Austin, M.D., Ge, H., Wu, W., Li, M., Yu, Z., Wasserman, D., Lyon, S.A., and Chou, S.Y. (2004) *Appl. Phys Lett.*, **84**, 5299.
8 Perret, C., Gourgon, C., Lazzarino, F., Tallal, J., Landis, S., and Pelzer, R. (2004) *Microelectron. Eng.*, **73–74**, 172.
9 Brandrup, J., Immergut, E.H., and Grulke, E.A. (1999) *Polymer Handbook (database)*, John Wiley & Sons.
10 Colburn, M., Johnson, S., Damle, S., Bailey, T., Choi, B., Wedlake, M., Michaelson, T., Sreenivasan, S.V., Ekerdt, J., and Willson, C.B. (1999) *Proc. SPIE*, **3676**, 379.
11 Miller, M., Schmid, G., Doyle, G., Thompson, E., and Resnick, D.J. (2006) S-FIL Template Fabrication for Full Wafer Imprint Lithography, in Proceedings NNT 06, San Francisco, US, 15–17 November.
12 Resnick, D.J., Schmid, G., Thompson, E., Stacey, N., Olynick, D.L., and Anderson, E. (2006) Step and Flash Imprint Lithography Templates for the 32 nm Node and Beyond, in Proceedings NNT 06, San Francisco, US 15–17 November.
13 Austin, M.D., Ge, H., Wu, W., Li, M., Yu, Z., Wasserman, D., Lyon, S.A., and Chou, S.Y. (2004) *Appl. Phys. Lett.*, **84**, 5299.
14 Gourgon, C., Perret, C., Tallal, J., Lazzarino, F., Landis, S., Joubert, O., and Pelzer, R. (2005) *J. Phys. D: Appl. Phys.*, **38**, 70.
15 Miller, M., Schmid, G., Doyle, G., Thompson, E., and Resnick, D.J. (2006) Proceedings NNT 06, San Francisco, US 15–17 November.
16 Wu, T.W., Best, M., Kercher, D., Dobisz, E., Bandic, Z., Yang, H., and Albrecht, T.R. (2006) Proceedings NNT 06, San Francisco, US 15–17 November.
17 Sreenivasan, S.V., Schumaker, P., McMackin, I., and Choi, J. (2006) Nano-Scale Mechanics of Drop-On-Demand UV Imprinting, in Proceedings NNT 06, San Francisco, US 15–17 November.

18 Hershey, R., Miller, M., Jones, C., Subramanian, M.G., Lu, X., Doyle, G., Lentz, D., and LaBrake, D. (2006) *Proc. SPIE*, **6337**, 6337M.

19 Vlasov, Y.A., Bo, X.Z., Sturm, J.C., and Norris, D.J. (2001) *Nature*, **414**, 289–293.

20 Seekamp, J., Zankovych, S., Helfer, A.H., Maury, P., Sotomayor-Torres, C., Boettger, M., Liguda, G., Eich, C., Heidari, M., Montelius, B., and Ahopelto, L. (2002) *J. Nanotechnol.*, **13**, 581–586.

21 Cui, B., Zhaoning, Y., Ge, H., and Chou, S.Y. (2007) *Appl. Phys. Lett.*, **90**, 043118.

22 Palmieri, F., Stewart, M.D., Wetzel, J., Hao, J., Nishimura, Y., Jen, K., Flannery, C., Li, B., Chao, H.L., Young, S., Kim, W.C., Ho, P.L., and Willson, C.G. (2006) *Proc. SPIE*, **6151**, 61510J/1–61510J/9.

23 Stewart, M., Wetzel, J., Schmid, G., *et al.* (2005) *Proc. SPIE-Microlithography*, **5751**, 210.

24 Schmid, G.M., Stewart, M.D., Wetzel, J., Palmieri, F., Hao, J., Nishimura, Y., Jen, K., Kim, E.K., Resnick, D.J., Liddle, J.A., and Willson, C.G. (2006) *J. Vac. Sci. Technol. B*, **24** (3), 1283.

25 Romanato, F., Businaro, L., Vaccari, L., Cabrini, S., Candeloro, P., De Vittorio, M., Passaseo, A., Todaro, M.T., Cingolani, R., Cattaruzza, E., Galli, M., Andreani, C., and Di Fabrizio, E. (2003) *Microelectron. Eng.*, **479**, 67–68.

26 Tormen, M., Businaro, L., Altissimo, M., Romanato, F., Cabrini, S., Perennes, F., Proietti, R., Sun, H.-B., Kawata, S., and Di Fabrizio, E. (2004) *Microelectron. Eng.*, **73**, 535–541.

27 Jeon, S., Malyarchuk, V., Rogers, J.A., and Wiederrecht, G.P. (2006) *Opt. Express*, **14**, 2300–2308.

28 Tormen, M., Carpentiero, A., Vaccari, L., Altissimo, M., Ferrari, E., Cojoc, D., and Di Fabrizio, E. (2005) *J. Vac. Sci. Technol. B*, **23**, 2920–2924.

29 Kehagias, N., Zelsmann, M., Pfeiffer, K., Ahrens, G., Gruetzner, G., and Sotomayor Torres, C.M. (2005) *J. Vac. Sci. Technol. B*, **23** (6), 2954–2957.

30 Huang, X.D., Bao, L.-R., Cheng, X., Guo, L.J., Pang, S.W., and Yee, A.F. (2002) *J. Vac. Sci. Technol. B*, **20**, 2872.

31 Borzenko, T., Tormen, M., Schmidt, G., Molenkamp, L.W., and Janssen, H. (2001) *Appl. Phys. Lett.*, **79**, 2246.

32 Bao, L.R., Cheng, X., Huang, X.D., Guo, L.J., Pang, S.W., and Yee, A.F. (2002) *J. Vac. Sci. Technol. B*, **20**, 2881.

33 Kehagias, N., Zelsmann, M., Pfeiffer, K., Ahrens, G., Gruetzner, G., and Sotomayor Torres, C.M. (2005) *J. Vac. Sci. Technol. B*, **23** (6), 2954–2957.

34 Nakajima, M., Yoshikawa, T., Sogo, K., and Hirai, Y. (2006) *Microelectron. Eng.*, **83**, 876–879.

35 Kehagias, N., Reboud, V., Chansin, G., Zelsmann, M., Jeppesen, C., Schuster, C., Kubenz, M., Reuther, F., Gruetzner, G., and Sotomayor Torres, C.M. (2007) *Nanotechnology*, **18**, 175303.

36 Kehagias, N., Chansin, G., Reboud, V., Zelsmann, M., Schuster, C., Kubenz, M., Reuther, F., Gruetzner, G., and Sotomayor Torres, C.M. (2007) *Microelectron. Eng.*, **84**, 921.

37 Kehagias, N., Reboud, V., Chansin, G., Zelsmann, M., Jeppesen, C., Reuther, F., Schuster, C., Kubenz, M., Gruetzner, G., and Sotomayor Torres, C.M. (2006) *J. Vac. Sci. Technol. B*, **24** (6), 3002–3005.

38 Le, Q.T., Claes, M., Conard, T., Kesters, E., Lux, M., and Vereecke, G. (2009) *Microelectron. Eng.*, **86**, 181–185.

39 Klein, M.F.G., Hein, H., Jakobs, P.J., Linden, S., Meinzer, N., Wegener, M., Saile, V., and Kohl, M. (2009) *Microelectron. Eng.*, **86**, 1078–1080.

40 Foucher, J., Pargon, E., Martin, M., Reyne, S., and Dupré, C. (2008) *Proc. SPIE*, **6922**, 69220.

41 Zangooie, S., Sendelbach, M., Angyal, M., Archie, C., Vaid, A., Matthew, I., and Herrera, P. (2008) *Proc. SPIE*, **6922**, 69220.

42 JEOL JSM-7600f, JEOL Ltd., 1-2, Musashino 3-chome Akishima, Tokyo 196–8558, Japan. Available at: http://www.jeol.com/PRODUCTS/ElectronOptics/ScanningElectronMicroscopesSEM/SemiinLensFE/JSM7600F/tabid/519/Default.aspx.

43 Zeiss ULTRA 60, Carl Zeiss NTS GmbH, Carl-Zeiss-Straße 56, 73447 Oberkochen, Germany. Available at:

http://www.zeiss.de/c1256e4600305472/ContentsFrame/b7e0976de51e3013c1256e58004f5177.
44 Otto, M., Bender, M., Zhang, J., Fuchs, A., Wahlbrink, T., Bolten, J., Spangenberg, B., and Kurz, H. (2007) *Microelectron. Eng.*, **84**, 980–983.
45 Austin, M.D., Zhang, W., Ge, H., Wasserman, D., Lyon, S.A., and Chou, S.Y. (2005) *Nanotechnology*, **16**, 1058–1061.
46 Joy, D.C. (2006) *J. Surf. Sci. Nanotechnol.*, **4**, 369–375.
47 Vladár, E., Villarrubia, J.S., Cizmar, P., Oral, M., and Postek, M.T. (2008) *Proc. SPIE*, **6922**, 69220.
48 Sasaki, S., Hiraka, T., Mizuochi, J., Nakanishi, Y., Yusa, S., Morikawa, Y., Mohri, H., and Hayashi, N. (2009) *Proc. SPIE*, **7271**, 72711.
49 Nasu, O., Sasada, K., Ikeda, M., and Ezumi, M.O. (2002) *Hitachi Review*, **51** (4), 125–129.
50 Frase, C.G., Buhr, E., and Dirscherl, K. (2007) *Meas. Sci. Technol.*, **18**, 510–519.
51 Chaix, N., Landis, S., Gourgon, C., Merino, S., Lambertini, V.G., Durand, G., and Perret, C. (2007) *Microelectron. Eng.*, **84**, 880–884.
52 Litt, L.C. and Malloy, M. (2009) *Proc. SPIE*, **7271**, 72711.
53 International Technology Roadmap for Semiconductors, Executive Summary 2007 Ed., p. 14. Available at: http://public.itrs.net/.
54 International Technology Roadmap for Semiconductors, 2007 Ed., Metrology, Table MET3a. Available at: http://public.itrs.net/.
55 Pritschow, M., Butschke, J., Irmscher, M., Parisolib, L., Oba, T., Iwai, T., and Nakamura, T. (2009) *Proc. SPIE*, **7271**, 72711.
56 Ke, C.M., Hu, J., Wang, W., and Huang, J. (2009) *Proc. SPIE*, **7272**, 72720.
57 Binnig, G., Quate, C.F., and Gerber, Ch. (1986) *Phys. Rev. Lett.*, **56** (9), 930.
58 Giessibl, J. (2005) *Mater. Today*, **8**, 32–41.
59 Veeco Dimension Icon, Veeco Instruments Inc., Terminal Drive, Plainview, NY 11803, USA. Available at: http://www.veeco.com/default.aspx.
60 Sugimoto, Y., Pou, P., Abe, M., Jelinek, P., Perez, R., Morita, S., and Custance, O. (2007) *Nature*, **446**, 64–67.
61 Meyer, G. and Amer, N.M. (1990) *Appl. Phys. Letts.*, **57**, 2089–2091.
62 Delobelle, P., Guillon, O., Fribourg-Blanc, E., Soyer, C., Cattan, E., and Remiens, D. (2004) *Appl. Phys. Lett.*, **85** (22), 5185–5187.
63 Ukraintsev, V.A. (2009) *Proc. SPIE*, **7272**, 727205.
64 Park, B.C., Choi, J., Ahn, S.J., Kim, D.H., Lyou, J., Dixson, R., Orji, N.G., Fu, J., and Vorburger, T.V. (2007) *Proc. SPIE*, **6518**, 651819.
65 Muruyama, K., Gonda, S., Koyanagi, H., Terasawa, T., and Hosaka, S. (2006) *Jpn. J. Appl. Phys.*, **45** (7), 5928–5932.
66 Orji, N.G., Dixson, R.G., Bunday, B.D., and Allgair, J.A. (2008) *Proc. SPIE*, **6922**, 692208.
67 Bartela, T.P., Kisielowski, C., Specht, P., Shubina, T.V., Jmerik, V.N., and Ivanov, S.V. (2007) *Appl. Phys. Lett.*, **91**, 101908.
68 Giannuzzia, L.A. and Stevie, F.A. (1999) *Micron*, **30**, 197–204.
69 Sendelbach, M., Zangooie, S., Vaid, A., Herrera, P., Leng, J., and Kim, I. (2008) *Proc. SPIE*, **6922**, 69220.
70 Dahlen, G.A., Liu, H.C., Osborn, M., Osborne, J.R., Tracy, B., and del Rosario, A. (2008) *Proc. SPIE*, **6922**, 69220.
71 Huang, H.T. and Terry, F.L. Jr (2004) *Thin Solid Films*, **455–456**, 828–836.
72 Herzinger, M., Johs, B., McGahan, W.A., Woollam, J.A., and Paulson, W. (1998) *J. Appl. Phys.*, **83** (6), 3323–3336.
73 Clement, T., Ingole, S., Ketharanathan, S., Drucker, J., and Picrauxa, S.T. (2006) *Appl. Phys. Lett.*, **89**, 163125.
74 R. Nanometrics Incorporated, 1550 Buckeye Drive, Milpitas, CA 95035 USA. Available at: http://www.nanometrics.com/products.html.
75 KLA-Tencor Corporation, One Technology Drive, Milpitas, California 95035, USA. Available at: http://www.kla-tencor.com.
76 Germer, T.A., Patrick, H.J., Silver, R.M., and Bunday, B. (2009) *Proc. SPIE*, **7272**, 72720.

77 Zhou, W., l Hsieh, M., Koh, H., and Zhou, M. (2008) *Proc. SPIE*, **6922**, 69223.

78 Pundaleva, H., Nam, D., Han, H., Lee, D., and Han, W. (2006) *Proc. SPIE*, **6152**, 61520.

79 Gereige, T., Robert, S., Thiria, S., Badran, F., Granet, G., and Rousseau, J.J. (2008) *J. Opt. Soc. Am. A*, **25** (7), 1661–1667.

80 Silver, R.M., Zhang, N.F., Barnes, B.M., Zhou, H., Heckert, A., Dixson, R., Germer, T.A., and Bunday, B. (2009) *Proc. SPIE*, **7272**, 727202.

81 Patrick, H.J., Attota, R., Barnes, B.M., Germer, T.A., Dixson, R.G., Stocker, M.T., Silver, R.M., and Bishop, M.R. (2008) *J. Micro/Nanolith. MEMS MOEMS*, **7** (1), 013012.

82 Attota, R., Silver, R., and Barnes, B.M. (2008) *Proc. SPIE*, **6922**, 69220.

83 Lee, H.J., Soles, C.L., Ro, H.W., Jones, R.L., Lin, E.K., Wu, W., and Hines, D.R. (2005) *Appl. Phys. Lett.*, **87**, 263111.

84 Jones, R.L., Hu, T., Soles, C.L., Lin, E.K., Reano, R.M., Pang, S.W., and Casa, D.M. (2006) *Nano Lett.*, **6** (8), 1723–1728.

85 Lee, H.J., Soles, C.L., Liu, D.W., Bauer, B.L., and Wu, W.L. (2002) *J. Polym. Sci., Part B: Polym. Phys.*, **40**, 2170.

86 Jefimovs, L., Vila-Comamala, J., Pilvi, T., Raabe, J., Ritala, M., and David, C. (2007) *Phys. Rev. Lett.*, **99**, 264801.

87 Kehoe, T., Bryner, J., Reboud, V., Kehagias, N., Landis, S., Gourgon, C., Vollmann, J., Dual, J., and Sotomayor Torres, C.M. (2008) *Proc. SPIE*, **6921**, 69210.

88 VanLandingham, M.R., Villarrubia, J.S., Guthrie, W.F., and Meyers, G.F. (2001) *Macromol. Symp.*, **167**, 15–43.

89 Rowland, H.D., King, W.P., Cross, G.L.W., and Pethica, J.P. (2008) *Am. Chem. Soc.*, **2** (3), 419–428.

90 Cross, G.L.W., Connell, B.S.O., Pethica, J.B., Rowland, H., and King, W.P. (2008) *Rev. Sci. Instrum.*, **79**, 013904.

91 Rowland, H.D., King, W.P., Pethica, J.P., and Cross, G.L.W. (2008) *Science*, **322**, 720–724.

92 Dalkoni-Veress, K., Forrest, J.A., Murray, C., Gigault, C., and Dutcher, J.R. (2001) *Phys. Rev. E*, **63**, 031801.

93 Hartschuh, R., Kisliuk, A., Novikov, V., Sokolov, A.P., Heyliger, P.R., Flannery, C.M., Johnson, W.L., Soles, C.L., and Wu, W.L. (2005) *Appl. Phys. Lett.*, **87**, 173121.

94 Kehoe, T., Bryner, J., Reboud, V., Vollmann, J., and Sotomayor Torres, C.M. (2009) *Proc. SPIE*, **71**, 72711.

95 Fuard, D., Perret, C., Farys, V., Gourgon, C., and Schiavone, P. (2005) *J. Vac. Sci. Technol. B*, **23** (6), 3069–3074.

96 Landis, S., Chaix, N., Gourgon, C., Perret, C., and Leveder, T. (2006) *Nanotechnology*, **17**, 2701–2709.

97 Brooks, C., Schmid, G.M., Miller, M., Johnson, S., Khusnatdinov, N., LaBrake, D., Resnick, D.J., and Sreenivasan, S.V. (2009) *Proc. SPIE*, **7271**, 72711.

98 Kehagias, N., Reboud, V., Sotomayor Torres, C.M., Sirotkin, V., Svintsov, A., and Zaitsev, S. (2008) *Microelectron. Eng.*, **85** (5–6), 846–849.

99 Al-Assaad, R.M., Regonda, S., Tao, L., Pang, S.W., and Hu, W. (2007) *J. Vac. Sci. Technol. B*, **25**, 2396–2401.

100 Leveder, T., Landis, S., Davoust, L., Soulan, S., and Chaix, N. (2007) *Proc. SPIE*, **6517**, 65170.

101 Ding, Y., Ro, H.W., Germer, T.A., Douglas, J.F., Okerberg, B.C., Karim, A., and Soles, C.L. (2007) *ACS Nano*, **1** (2), 84–92.

102 Patrick, H.J., Germer, T.A., Ding, Y., Ro, H.W., Richter, L.J., and Soles, C.L. (2009) *Proc. SPIE*, **7271**, 727128.

103 Soulan, S., Besacier, M., Leveder, T., and Schiavone, P. (2007) *Proc. SPIE*, **6617**, 661713.

104 International Technology Roadmap for Semiconductors, 2008 Update, Overview, Table MET1. Available at: http://public.itrs.net/.

105 Levin, T., Livne, M., and Gillespie, R.M. (2007) *Proc. SPIE*, **6518**, 651855.

106 Yu, Z., Gao, H., and Chou, S.Y. (2007) *Nanotechnology*, **18**, 065304.

107 Li, J., Liu, Z., Rabello, S., Dasari, P., Kritsun, O., and Volkman, C. (2009) *Proc. SPIE*, **7272**, 727207.

108 International Technology Roadmap for Semiconductors (April 4 2005). Available at: http://public.itrs.net/.

109 Cardinale, G.F., Skinner, J.L., Talin, A.A., Brocato, R.W., Palmer, D.W., Mancini, D.P., Dauksher, W.J., Gehoski, K., Le, N., Nordquist, K.J., and Resnick, D.J. (2004) *J. Vac. Sci. Technol. B*, **22**, 3265–3270.

110 McClelland, G.M., Hart, M.W., Rettner, C.T., Best, M.E., Carter, K.R., and Terri, B.D. (2002) *Appl. Phys. Lett.*, **81**, 1483–1485.

111 Ahn, S.W., Lee, K.-D., Kim, J.S., Kim, S.H., Lee, S.H., Park, J.D., and Yoon, P.W. (2005) *Microelectron. Eng.*, **78–79**, 314–318.

112 Sotomayor Torres, C.M., Zankovych, S., Seekamp, J., Kam, A.P., Clavijo Cedeño, C., Hoffmann, T., Ahopelto, J., Reuther, F., Pfeiffer, K., Bleidiessel, G., Gruetzner, G., Maximov, M.V., and Heidari, B. (2003) *Mater. Sci. Eng. C*, **23**, 23–31.

113 Nilsson, D., Nielsen, T., and Kristensen, A. (2004) *Rev. Sci. Instrum.*, **75**, 4481–4486.

114 Nilsson, D., Balslev, S., and Kristensen, A. (2005) *J. Micromech. Microeng.*, **15**, 296–300.

115 Reboud, V., Lovera, P., Kehagias, N., Zelsmann, M., Reuther, F., Gruetzner, G., Redmond, G., and Sotomayor Torres, C.M. (2007) *Appl. Phys. Lett.*, **91**, 151101.

116 Wang, J., Sun, X., Chen, L., and Chou, S.Y. (1999) *Appl. Phys. Lett.*, **75**, 2767–2769.

117 Cheng, X., Hong, Y., Kanicki, J., and Guo, L.J. (2002) *J. Vac. Sci. Technol. B*, **20**, 2877–2880.

118 Clavijo Cedeno, C., Seekamp, J., Kam, A.P., Hoffmann, T., Zankovych, S., Sotomayor Torres, C.M., Menozzi, C., Cavallini, M., Murgia, M., Ruani, G., Biscarini, F., Behl, M., Zentel, R., and Ahopelto, J. (2002) *Microelectron. Eng.*, **61–62**, 25–31.

119 Seekamp, J., Zankovych, S., Helfer, A.H., Maury, P., Sotomayor Torres, C.M., Bottger, G., Liguda, C., Eich, M., Heidari, B., Montelius, L., and Ahopelto, J. (2002) *Nanotechnology*, **13**, 581–586.

120 Ahn, S.W., Lee, K.D., Kim, J.S., Kim, S.H., Park, J.D., Lee, S.H., and Yoon, P.W. (2005) *Nanotechnology*, **16**, 1874.

121 Kehagias, N., Zankovych, S., Goldschmidt, A., Kian, R., Zelsmann, M., Sotomayor Torres, C.M., Pfeiffer, K., Ahrens, G., and Gruetzner, G. (2004) *Superlattices and Microstructures*, **36**, 201–210.

122 Chao, C.Y. and Guo, L.J. (2002) *J. Vac. Sci. Technol. B*, **20** (6), 2862–2866.

123 Kim, D.H., Im, J.G., Lee, S.S., Ahn, S.W., and Lee, K.D. (2005) *IEEE Photonic. Tech. L.*, **17**, 11.

124 Paloczi, G.T., Huang, Y., Yariv, A., Luo, J., and Jen, A.K.Y. (2004) *Appl. Phys. Lett.*, **85** (10), 1662–1664.

125 Meier, M., Dodabalapur, A., Rogers, J.A., Slusher, R.E., Mekis, A., Timko, A., Muray, C.A., Ruel, R., and Nalamasu, O. (1999) *J. Appl. Phys.*, **86** (7), 3502–3507.

126 Pisignano, D., Persano, L., Visconti, P., Cingolani, R., and Gigli, G. (2003) *Appl. Phys. Lett.*, **83** (13), 2545–2547.

127 Reboud, V., Lovera, P., Kehagias, N., Zelsmann, M., Schuster, C., Reuther, F., Gruetzner, G., Redmond, G., and Sotomayor Torres, C.M. (2007) *Appl. Phys. Lett.*, **91**, 151101.

128 Arango, F., Christiansen, M.B., Gersborg-Hansen, M., and Kristensen, A. (2007) *Appl. Phys. Lett.*, **91**, 223503.

129 Chen, Y., Li, Z., Zhang, Z., Psaltis, D., and Scherer, A. (2007) *Appl. Phys. Lett.*, **91** (5), 051109.

130 Reboud, V., Kehagias, N., Zelsmann, M., Fink, M., Reuther, F., Gruetzner, G., and Sotomayor Torres, C.M. (2007) *Opt. Express*, **15**, **12**, 7190.

131 Reboud, V., Kehagias, N., Striccoli, M., Placido, T., Panniello, A., Curri, M.L., Zelsmann, M., Reuther, F., Gruetzner, G., and Sotomayor Torres, C.M. (2007) *J. Vac. Sci. Technol. B*, **25**, 2642.

132 Schift, H., Park, S., Jung, B., Choi, C.G., Kee, C.S., Han, S.P., Yoon, K.B., and Gobrecht, J. (2005) *Nanotechnology*, **16**, S261–S265.

133 Belotti, M., Torres, J., Roy, E., Pepin, A., Chen, Y., Gerace, D., Andreani, L.C., and Galli, M. (2006) *Microelectron. Eng.*, **83** (4–9), 1773–1777.

134 Tamborra, M., Striccoli, M., Curri, M.L., Alducin, J.A., Mecerreyes, D., Pomposo,

J.A., Kehagias, N., Reboud, V., Sotomayor Torres, C.M., and Agostian, A. (2007) *Small*, **3**, 822.

135 Reboud, V., Kehagias, N., Zelsmann, M., Striccoli, M., Tamborra, M., Curri, M.L., Agostiano, A., Fink, M., Reuther, F., Gruetzner, G., and Sotomayor Torres, C.M. (2007) *Appl. Phys. Lett.*, **90**, 011114.

136 Reboud, V., Lovera, P., Kehagias, N., Zelsmann, M., Schuster, C., Reuther, F., Gruetzner, G., Redmond, G., and Sotomayor Torres, C.M. (2007) *Appl. Phys. Lett.*, **91**, 151101.

137 Sasaki, M., Li, Y., Akatu, Y., Fujii, T., and Hane, K. (2000) *Jpn. J. Appl. Phys*, **39**, 7145–7149.

138 Li, Y., Sasaki, M., and Hane, K. (2001) *J. Micromech. Microeng.*, **11**, **3**, 234–238.

139 Kragh, P. and Kristensen, A. (2003) *Proc. of the 17th Conference on Solid-State Transducers, Eurosensors*, 380.

140 Armani, M., Srinivasan, A., and Vahala, K.J. (2007) *Nano Lett.*, **7**, 6.

141 Armani, K., Kippenberg, T., Spillane, S.M., and Vahala, K.J. (2003) *Nature*, **421**, 925–929.

142 Jay Guoa, L.Y. (2002) *J. Vac. Sci. Technol. B*, **20**, 2862.

143 Balslev, S., Rasmussen, T., Shi, P., and Kristensen, A. (2005) *J. Micromech. Microeng.*, **15**, 2456–2460.

144 Balslev, S., Nielsen, R.B., Petersen, D.H., and Kristensen, A. (2006) *J. Vac. Sci. Technol. B*, **24**, 3252–3257.

145 Reboud, V., Lovera, P., Kehagias, N., Zelsmann, M., Schuster, C., Reuther, F., Gruetzner, G., Redmond, G., and Sotomayor Torres, C.M. (2007) *Appl. Phys. Lett.*, **91**, 151101.

146 Kehagias, N., Reboud, V., Chansin, G., Zelsmann, M., Jeppesen, C., Schuster, C., Kubenz, M., Reuther, F., Gruetzner, G., and Sotomayor Torres, C.M. (2007) *Nanotechnology*, **18**, **17**, 175303.

147 Kehagias, N., Reboud, V., Chansin, G., Zelsmann, M., Jeppesen, C., Reuther, F., Schuster, C., Kubenz, M., Gruetzner, G., and Sotomayor Torres, C.M. (2006) *J. Vac. Sci. Technol. B*, **24** (6), 3002–3005.

148 Kehagias, N., Zelsmann, M., and Sotomayor Torres, C.M. (2005) *Proc. SPIE*, **5825**, 654.

149 Sharp, D.N., Campbell, M., Dedman, E.R., Harrison, M.T., Denning, R.G., and Turberfield, A.J. (2002) *Opt. Quantum Electron.*, **34**, **3**, 3–12.

150 Kehagias, N., Reboud, V., Chansin, G., Zelsmann, M., Jeppesen, C., Reuther, F., Schuster, C., Kubenz, M., Gruetzner, G., and Sotomayor Torres, C.M. (2006) *J. Vac. Sci. Technol. B*, **24** (6), 3002–3005.

151 Chao, C.Y., and Guo, L.J. (2004) *IEEE Photonics Technol. Lett.*, **16**, 1498–1500.

152 Cho, C.O., Jeong, J., Lee, J., Kim, I., Jang, D.H., Park, Y.S., and Woo, J.C. (2005) *Appl. Phys. Lett.*, **87**, 161102.

153 Kwon, S.H., Ryu, H.Y., Kim, G.H., Lee, Y.H., and Kim, S.B. (2003) *Appl. Phys. Lett.*, **83**, 3870.

154 Joannopoulos, J.D., Villeneuve, P.R., and Fan, S. (1997) *Nature*, **386**, 143.

155 Byeon, K.J., Hwang, S.Y., and Lee, H. (2007) *Appl. Phys. Lett.*, **91**, 091106.

156 Christiansen, M.B., Schøler, M., and Kristensen, A. (2007) *Opt. Express*, **15**, 3931.

157 Köck, A., Gornik, E., Hauser, M., and Beinstingl, W. (1990) *Appl. Phys. Lett.*, **57**, 2327–2329.

158 Barnes, W.L. (1999) *J. Lightwave Technol.*, **17**, 2170–2182.

159 Reboud, V., Kehagias, N., Zelsmann, M., Fink, M., Reuther, F., Gruetzner, G., and Sotomayor Torres, C.M. (2007) *Opt. Express*, **15**, **12**, 7190.

160 Neal, T.D., Okamoto, K., Scherer, A., Liu, M.S., and Jen, A.K.-Y. (2006) *Appl. Phys. Lett.*, **89**, 221106.

161 Jana, N.R. (2005) *Small*, **1**, 875.

162 Merino, S., Retolaza, A., Schift, H., and Trabadelo, V. (2007) *Microelectronic Eng.*, **84**, 848.

163 Hu, W., Lu, N., Zhang, H., Wang, Y., Kehagias, N., Reboud, V., Sotomayor Torres, C.M., Hao, J., Li, W., Fuchs, H., and Chi, L. (2007) *Adv. Mater.*, **19**, 2119.

164 Rogers, J.A. (2001) *Science*, **291**, 1502–1503.

165 Behl, M., Seekamp, J., Zankovych, S., Sotomayor Torres, C.M., Zentel, R., and

Ahopelto, J. (2002) *Adv. Mater.*, **14** (8), 588–591.

166 Fuchs, A., Vratzov, B., Wahlbrink, T., Georgiev, Y., and Kurz, H. (2004) *J. Vac. Sci. Technol. B*, **22**, 3242.

167 Li, Z., Zhang, Z., Emery, T., Scherer, A., and Psaltis, D. (2006) *Opt. Express*, **14**, 696.

168 Christiansen, M., Schøler, M., and Kristensen, A. (2007) *Opt. Express*, **15**, 3931.

169 Pisignano, D., Persano, L., Gigli, G., Visconti, P., Stomeo, T., De Vittorio, M., Barbarella, G., Favaretto, L., and Cingolani, R. (2004) *Nanotechnology*, **15**, 766.

5
Anodized Aluminum Oxide

Günter Schmid

5.1
Introduction

Aluminum is, like titanium, zirconium, hafnium, vanadium, niobium, tantalum or magnesium, a so-called "gate metal" [1]. Gate metals have special electrochemical properties, and are distinguished by the formation of insulating oxide layers if they are used as anodes in electrochemical redox reactions. The current is based on ion transport, and is closed down at a temperature- and voltage-dependent material-specific limit; this occurs due to a reduction of the electric field by the increasing oxide layer. In contrast, the cathodic use of a gate metal results in the evolution of hydrogen. The formation of oxide layers is, of course, only possible if a solvent is used in which the oxide is not soluble. In contrast to most other metals, gate metals form compact, strongly adhering oxide layers that protect the metal beneath then from further oxidation. Due to its electrode potential of −1.662 V, aluminum is readily oxidized by contact with air to give oxide layers that are 1 to 10 nm thick [2], and this protective oxide layer can be extended up to 1 μm by anodic oxidation in neutral electrolytes [3]. This behavior, combined with its low specific weight, makes aluminum a particularly valuable material, and it is used preferentially in aircraft and automobile construction.

The electrolytic oxidation of aluminum has long been known as the Eloxal process, and this has been applied on a practical basis to generate homogeneous oxide layers on aluminum materials. Aluminum oxide is a typical insulator, with a band gap of 7–8 eV [1]. The electronic charge transport is very small and is, at high field strengths, exceeded by the ion conductivity. In anodically generated layers the charge transport is mainly determined by existing space charges [3–6] that are generated during the anodizing process in the barrier layer between metal and metal oxide, and additionally between the metal oxide and electrolyte. These are generated by Al^{3+} cations and by O^{2-} anions, as well as by electrolyte anions.

Porous aluminum oxide is generated by the anodic oxidation of aluminum surfaces in electrolytes that are able to dissolve the formed oxide. Anodized aluminum oxide (AAO), when equipped with nanopores with diameters varying between a few nanometers up to several hundreds of nanometers, is a most valuable

Nanotechnology, Volume 8: Nanostructured Surfaces. Edited by Lifeng Chi

ISBN: 978-3-527-31739-4

material for many reasons. Although porous alumina surfaces are *per se* interesting nanostructured materials, the pores can also be used as templates to generate nanowires, nanotubes or nanoparticles of different materials such as metals or semiconductors. Moreover, they can also serve as imprinting tools for the fabrication of nanostructured surfaces of various materials, and equipped with nanosticks that correspond to the pore size of the AAO-surface. One very special development during recent years has been the generation of AAOs with ordered pore-structures.

These aspects, relating in particular to nanoporous AAOs, will be discussed in detail in the following sections.

5.2 Fundamentals of Aluminum Oxide Formation

5.2.1 Barrier Oxide Layers

During the anodizing of high-purity aluminum foils or plates in neutral or weak acidic electrolytes, compact, nonporous so-called "barrier oxide" layers are formed. Examples of such electrolytes include boric acid, phosphate, ammonium borate, or tartrate solutions [1–3]. If a voltage is applied between an aluminum anode and an appropriate cathode, the current intensity I drops exponentially with time, corresponding to Eq. (5.1) [1–3].

$$I = I_0 \, e^{\beta E} \tag{5.1}$$

where I_0 and β are temperature-dependent material constants, and E is the electric field strength in the oxide.

Since the electric field strength E decreases with increasing oxide layer thickness d, [corresponding to Eq. (5.2)]:

$$E = \Delta U / d \tag{5.2}$$

where ΔU is the voltage drop and the current I falls exponentially. Hence, oxide formation will be closed down depending on the voltage applied. The ratio of the resultant thickness of the oxide layer and the anodizing voltage lies between 0.8 and 1.4 nm V^{-1} [1, 3, 7, 8]. Depending on the type of electrolyte, transport of the current occurs mainly via Al^{3+} cations and protons (phosphate-containing electrolytes) or by anions such as OH^-, O^{2-} and electrolyte anions (borate-containing electrolytes) [2]. The qualitative relationship between current density and time is shown in Figure 5.1 [3].

The formation of compact barrier layers can be explained by the so-called "High Field Law." This starts out from the fact that charge transport in aluminum oxide layers at high field strengths (up to 10^9 V m^{-1}) is dominated by ion conductivity. The electronic conductivity is very small, and only relevant at low field strengths due to tunneling processes and deposited foreign ions [1, 2, 4–6]. This residual current is the

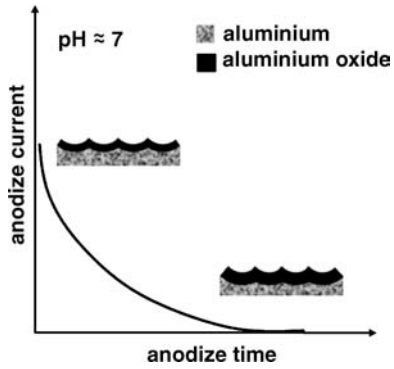


Figure 5.1 Current density–time course of barrier oxide formation.

reason why the current density at the end of the anodizing process is not zero. At high field strengths electronic conductivity can be neglected. The law also presumes that the diffusion of ions in the opposite direction to the electric field at sufficiently high field strengths can be excluded. The ionic conductivity is caused by the ions hopping from regular lattice positions and interstitial positions to neighbored fault points, due to the high electric field strength. These hopping steps require a distinct activation energy W in the field-free space, and W increases with the hopping distance, and can be varied by applying an electric field. The relationship between the activation energy and the electric field is shown graphically in Figure 5.2 [1].

Hence, the activation energy for a hopping process in the opposite direction of the electric field direction increases [Eq. (5.3)] and decreases if hopping occurs in the direction of the electric field [Eq. (5.4)]:

$$W_{\leftarrow} = W + (1-\alpha) a\, z\, \mathrm{F}\, E \tag{5.3}$$

$$W_{\rightarrow} = W - \alpha\, a\, z\, \mathrm{F}\, E \tag{5.4}$$

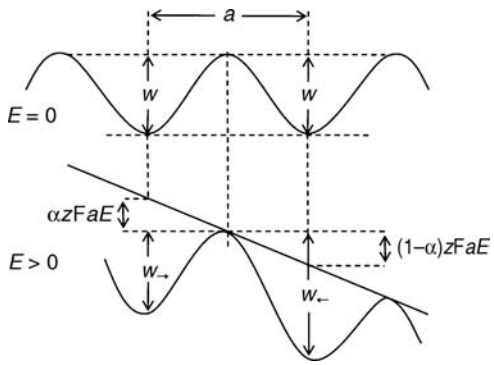


Figure 5.2 The influence of an electric field on the activation energy [1].

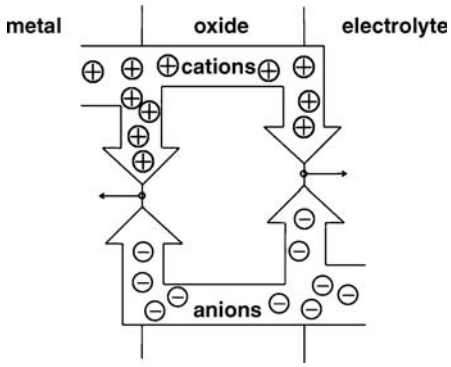


Figure 5.3 Formalized sketch of aluminum oxide growth at both boundary layers [1].

where α is the transfer coefficient which describes the symmetry of the activation barrier, z is the charge number, a the hopping distance, E the electric field strength, and F is Faraday's constant.

Typical values for the activation energies lie between 0.9 and 1.7 eV [1].

Since the transport numbers of the cations during oxide formation lie between 0.2 and 0.4, and may even be 0.5 at high field strengths, both the anions and cations are responsible for oxide formation at the metal/metal oxide boundary layer as well as at the metal oxide/electrolyte boundary layer. This occurs due to the simultaneous migration of the anions and cations [1, 3, 9, 10]. The process is summarized in Figure 5.3.

About one-half of the metal ions generated at the metal/metal oxide boundary layer move to the metal oxide/electrolyte boundary layer, where they react with OH^- and O^{2-} ions to produce a new oxide and, as a consequence the metal oxide/electrolyte boundary layer will be moved in the direction of the electrolyte. The other half of the metal ions react at the metal/metal oxide boundary layer with correspondingly moved O^{2-} ions to form fresh oxide, thus shifting the metal/metal oxide boundary layer in the direction of the metal. The mobile charge carriers are generated exclusively at the boundaries rather than inside the oxides, since the activation energy that is necessary for ion hopping is much lower than the energy needed for the formation of Schottky or Frenkel defects [1].

5.2.2 Formation of Nanoporous Aluminum Oxide

5.2.2.1 Mechanism of Formation

The anodic switching of aluminum foils or plates in alkaline (and especially in acid) electrolytes results in the formation of porous oxide layers that consist of two components – a thin barrier layer, which is comparable to that discussed above, and a porous layer. The thickness of the barrier layer is only 0.8–1.0 nm V^{-1}, due to a continuous dissolution of the oxide by the electrolyte [2, 3, 11]; consequently, a continuous anodic current passage is possible. The conductivity of the barrier layer is

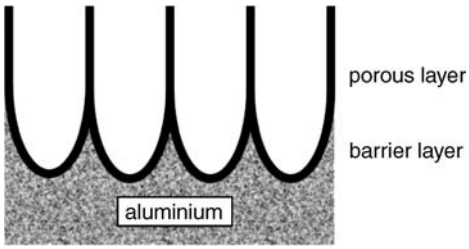


Figure 5.4 The sequence of aluminum, the barrier layer, and the porous aluminum oxide layer.

based on ion transport phenomena (see Section 5.2.1), with the adjacent porous oxide layer being characterized by pores that run perpendicular to the surface. The formation of pores is the consequence of several factors, including the voltage, temperature, the type of electrolyte, and the electrolyte concentration. The typical construction of the barrier and porous layers is shown schematically in Figure 5.4.

There exist two major models that describe the mechanism of porous oxide formation. In the model of O'Sullivan and Wood [11], it is assumed that the aluminum surface has a nanometer-scale roughness, while on top of the heights the current density is increased such that oxide formation at these positions is also increased. Due to the high field strength (up to $10^9\,V\,m^{-1}$), polarization of the oxide lattice sets in, and this is accompanied by a favored dissolution of aluminum oxide by the electrolyte. As the electric field strength at positions with thin oxide layers is higher, the oxide dissolves faster here than at the elevated positions, with the result that concave structures and, finally, pores are formed. The formation of a surface which is saturated with nanometer-sized irregularities can be initiated by a so-called "electropolishing process." To achieve this, an anodization process is carried out at high temperature, high current, and under very low or very high pH conditions, linked with a continuous dissolution of the formed oxide. This results in a surface of nanosized irregularities as the best conditions in which to grow AAOs.

The course of the pore formation, starting with the generation of the barrier layer, is shown schematically in Figure 5.5. Here, the starting situation A changes to B, after which an interplay between the chemical dissolution and the electric field-assisted pore growth begins (C). This ends in a steady-state plateau region as a function of the barrier layer thickness (D) which is, in part, a function of the applied voltage.

The model of Macdonald [8] starts out from the presumption that cationic vacancies in the aluminum oxide layer diffuse to the surface under the influence of the electric field, with interruption of the electric contact and followed by formation of small elevations on the aluminum surface. From here on, the process follows the steps described by O'Sullivan and Wood. Hence, the two models only differ in their very early stages, which are responsible for different current densities linked with different dissolution behaviors.

The pore diameters, pore densities and thicknesses of the membranes have been shown to depend on the experimental conditions (see Table 5.1 for details). Clearly, pore diameters ranging from about 10 nm to several hundred nanometers are possible, depending on the voltage. Logically, the thickness of the oxide layer

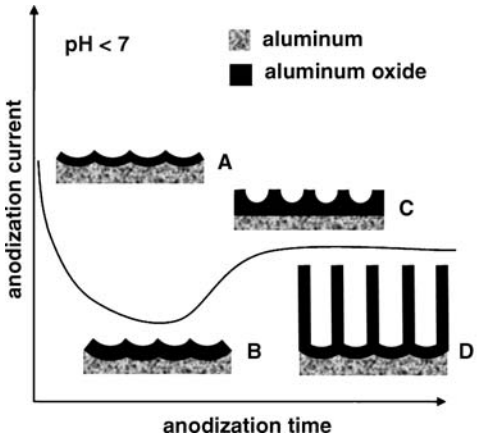


Figure 5.5 Relationship between anodization current and anodization time in acidic electrolytes, with the formation of pores.

corresponds with the anodizing time, but may vary from a few nanometers up to several hundred micrometers. In order to generate pores of a constant diameter through the whole oxide layer, a constant voltage must be maintained during the generation process. The linear correlation between the applied voltage and pore diameter is shown graphically in Figure 5.6.

Based on transmission electron microscopy (TEM) investigations, a linear regression of the measured values and the specific pore diameter d_{TEM} results in [13]:

$$d_{\mathrm{TEM}} = 1.37\,\mathrm{nm\,V^{-1}}\,U + 0.36\,\mathrm{nm} \tag{5.5}$$

Based on other TEM investigations, it also follows that the pore density is linearly proportional to the reciprocal square of the applied voltage (see Figure 5.7). A similar

Table 5.1 Parameters for anodized aluminum oxide (AAO) formation for some frequently used polyprotic electrolytes (Source [12]).

Anodization parameters for common polyprotic electrolytes						
% Electrolyte (w/w)	Temperature (°C)	Potential (V)	Duration (h)	Thickness (μm)	Pore diameter (nm)	Pore density (N cm^{-2})
4% H_3PO_4	0	130	8	50	200	1.3×10^8
1% H_2CO_4	−5	90	2	40	120	2.2×10^9
1% H_2CO_4	0	70	3	30	86	4.3×10^9
1% H_2CO_4	0	40	6	40	60	1.2×10^{10}
4% H_2CO_4	2	30	10	25	52	1.4×10^{10}
10% H_2SO_4	0	20	4	20	32	3.7×10^{10}
10% H_2SO_4	5	15	6	15	22	8.0×10^{10}
15% H_2SO_4	8	10	3	5	10	3.8×10^{11}
20% H_2SO_4	20	2	1	<1	~5	1.5×10^{12}

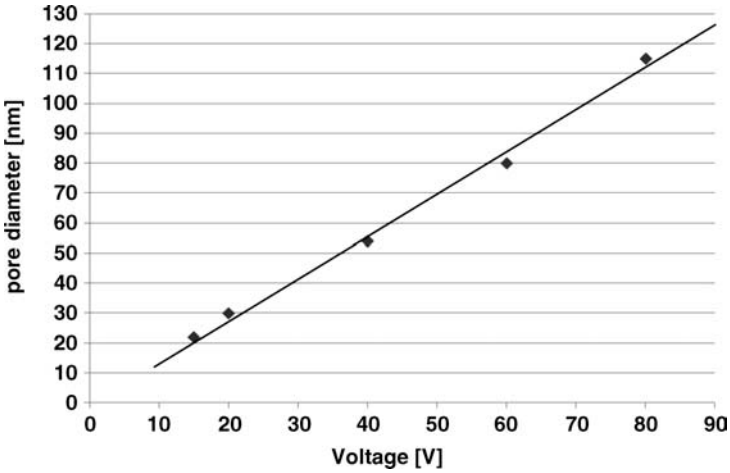


Figure 5.6 Relationship between the applied voltage and the resultant pore diameter.

relationship can be deduced via investigations using atomic force microscopy (AFM) [13].

One important aspect of AAO fabrication is a perfect detachment of the membranes from the aluminum substrates, where they are generated. Among several routes employed to detach the porous layer from the substrate, a stepwise 10% voltage reduction at completion of the process was shown to be the most efficient. Due to the voltage reduction, proportionally smaller pores are formed in the barrier layer, and this is linked with a corresponding thinning of the barrier layer. Finally, hydrogen evolution begins due to an interaction of the acidic electrolyte solution and the

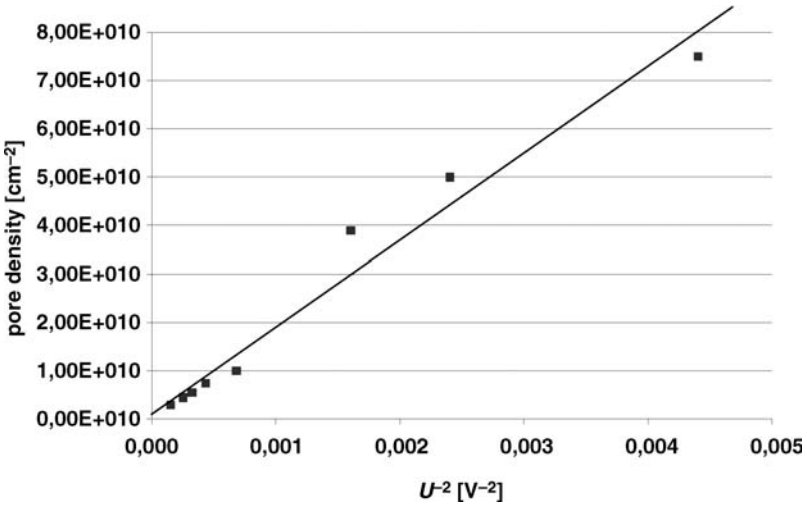


Figure 5.7 Relationship between the pore density and the reciprocal square of voltage [13].

aluminum metal. As the hydrogen gas is formed, it bubbles between the oxide layer and the aluminum, so as to detach the AAO film.

Another frequently applied method of detachment involves the use of mercuric chloride, which reacts with the aluminum support to form an amalgam, causing separation of the AAO from the surface. In both cases, the wet oxide film can be removed carefully from the solution, rinsed, dried and, if necessary, transferred to a crystalline membrane.

One final point with regards to AAO formation concerns the order of the pores. Under the conditions described above, the generated pores are predominantly dense-packed but do not show any real highly ordered structure. However, highly ordered pore structures can be fabricated by applying a two-step process [14] that starts with a 24 h anodization process, followed by dissolution of the oxide layer in a strong acid, such that an aluminum surface is formed that is well prepared for a second anodization procedure. This second anodization results in perfect hexagonally ordered pores that have no branches which can often be observed under "normal" conditions. Another pretexturing process involves the use of a single-crystal silicon carbide mold to imprint aluminum foils before anodization. and by using this method defect-free areas in the millimeter range can be generated [15]. Highly ordered pores can also be obtained by means of a two-dimensional (2-D) array of monodisperse nanoparticles (e.g., Fe_2O_3) as a template [16]. In this case, an aluminum layer is first sputtered onto the particles and then removed, producing a perfect template for the subsequent anodization process.

5.2.2.2 **Chemical Composition**

The as-formed "aluminum oxide" layers do not really consist of pure Al_2O_3; rather, they are characterized by a complex mixture of amorphous hydrated aluminum–oxygen compounds, including incorporated electrolyte ions [2, 8, 11, 17]. Within a porous layer, it is possible to distinguish three main areas (see Figure 5.8). The pores themselves are enclosed by an amorphous electrolyte-containing layer, followed by an aluminum oxide hydrate (AlOOH) layer, which is free of electrolyte. The latter layer

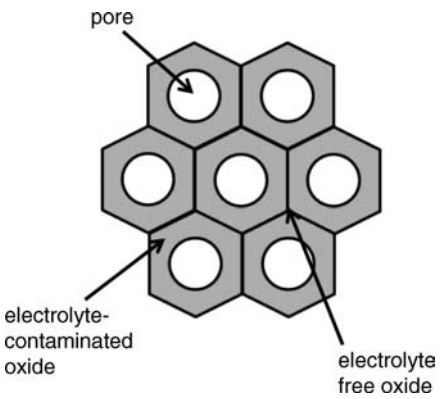


Figure 5.8 The approximate chemical composition of a porous alumina layer.

corresponds in the ideal case to a Böhmit stoichiometry [1–3, 18]. The thickness of the electrolyte-free region varies with the nature of the electrolyte, and increases from sulfuric acid, oxalic acid, phosphoric acid, to chromic acid. The yield of electrolyte in the contaminated layer may be up to 14% in the case of sulfate.

The hydrothermal treatment of such layers results in crystalline Bömit species; thermal treatment from 450 °C upwards produces nonstable γ-aluminum oxide, whilst at temperatures in excess of 1200 °C stable α-aluminum oxide layers are formed [18].

5.3
Characterization of Nanoporous Aluminum Oxide Membranes

During recent years, two main methods have been successfully developed to image and characterize nanoporous aluminum oxide layers on aluminum, or detached as membranes, namely electron microscopy (EM) and AFM. Whether applied as TEM or scanning electron microscopy (SEM), EM can provide valuable complementary information to AFM, and the simultaneous use of both techniques is generally advantageous when investigating membrane samples.

5.3.1
Electron Microscopy

Generally, the layer thickness of AAO membranes does not allow direct imaging by TEM, and samples must often be prepared using an ion-etching process, despite this resulting in very thin layers. A TEM image of the surface of a well-ordered membrane with a pore size of 53 ± 10 nm is shown in Figure 5.9. In this case, the fabrication was carried out at 40 V in 4% oxalic acid over 46 h, and the pore density (p) was $10^{10}\,\mathrm{cm}^{-2}$ [19].

As can be seen from Figure 5.9a, highly ordered hexagonal areas were apparent over the entire image, while the membrane structure, with its pores, electrolyte-containing intermediate layer and dark hexagons of pure oxide (see Figure 5.8) could be clearly seen in a magnified area (Figure 5.9b).

Typically, SEM is also well-suited to imaging not only the surfaces of membranes, but also their cross-sections, from which the pores – which run perpendicular to the surface – can be visualized. A minor disadvantage of SEM is that nonconducting surfaces such as aluminum oxide must first be coated with a thin metal film in order to avoid electrostatic charging. In fact, if this is not very carefully carried out, the pores can become fully or partially become closed and the images falsified. The major advantage of SEM is that the AAOs do not require any previous treatment, by ion-etching. The SEM images of 15 V and 40 V membrane surfaces, equipped with pores of about 20 nm and 50–60 nm, are shown respectively in Figures 5.10 and 5.11 [13]. In Figure 5.11, some of the pores appear either partially or totally closed.

From about 60 V upwards, the pore structures are normally less well developed when compared to the 40 V membranes, with both pore geometry and distribution

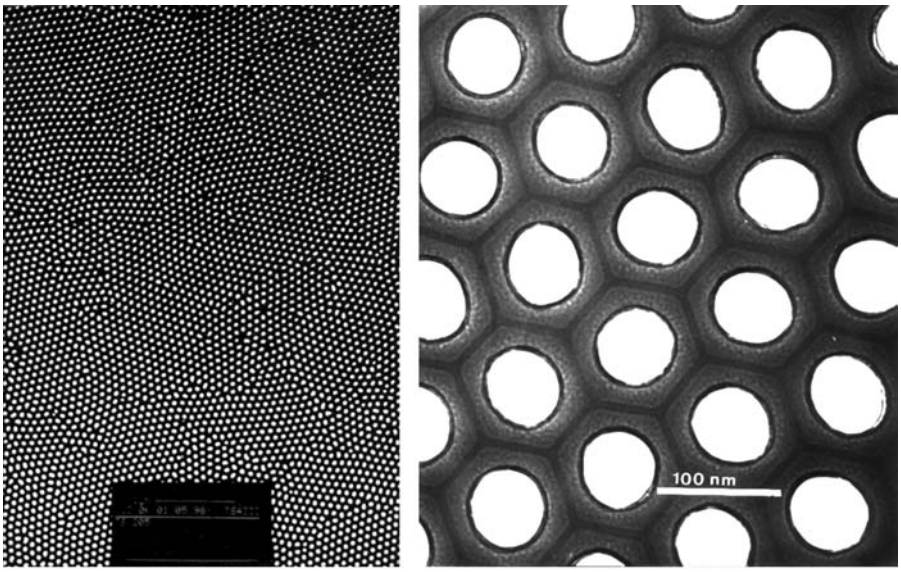


Figure 5.9 Transmission electron microscopy image of a 40 V aluminum oxide membrane. (a) Highly ordered hexagonal areas; (b) Magnified cutout indicating the electrolyte-containing intermediate layer and hexagons of pure oxide (dark regions). Panel (b) was reproduced with permission from Springer.

being of poorer quality. An enlargement of a portion of membrane generated at 150 V is shown in Figure 5.12 [13].

The cross-sectional imaging of membranes with SEM allows an insight into both pore structure and quality. Although, normally, the pores run parallel through the membrane, occasional embranchments may be observed, especially in membranes

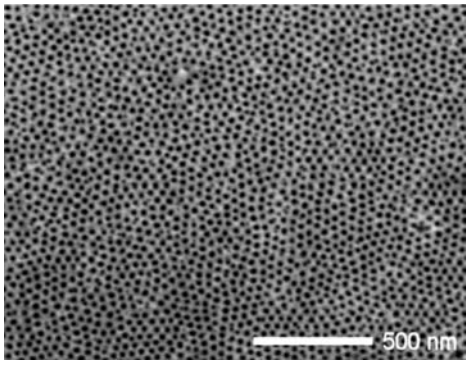


Figure 5.10 Scanning electron microscopy image of a 15 V membrane surface with pores of about 20 nm.

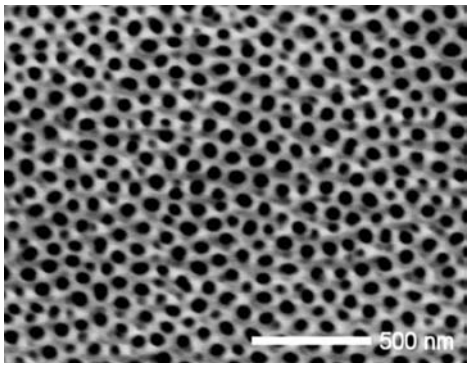


Figure 5.11 Scanning electron microscopy image of a 40 V membrane with 50–60 nm pores.

where the pores are very small. A cross-sectional SEM image of a 60 nm pore membrane with perfect parallel pores of uniform size is shown in Figure 5.13.

SEM has also been applied to demonstrate the existence of the barrier layer beneath the membranes [20].

5.3.2 Atomic Force Microscopy (AFM)

Since its discovery, AFM, together with scanning tunneling microscopy (STM) and other scanning probe microscopy techniques, have undergone continuous development to a point where they now represent unrenouncable tools for the characterization of all types of surface. In contrast to SEM, AFM is based on the forces between a tip and the surface atoms, without the need for any electronic interactions, so that both conductive and nonconductive surfaces can be investigated [21, 22]. The resolution depends firstly on the quality of the tip. At the so-called contact-modus, the forces between the tip and the surface are repulsive, whereas in the noncontact modus technique the forces are of an attractive nature. The best results for AAO

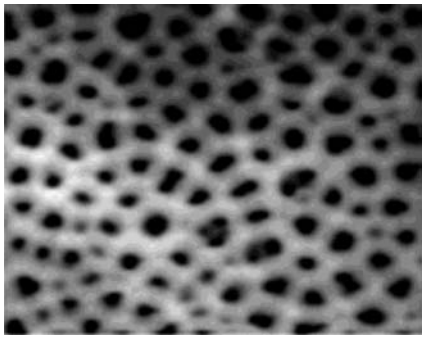


Figure 5.12 Scanning electron microscopy image of a 150 V membrane, indicating the reduced quality compared to 40 V membranes.

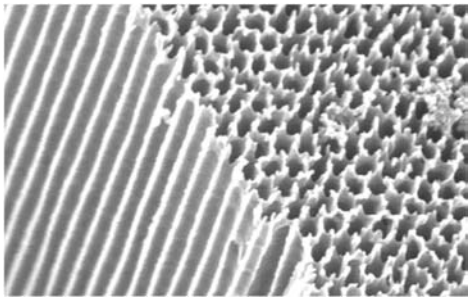


Figure 5.13 Scanning electron microscopy image of a cross-sectioned membrane with 60 nm pores. Reproduced with permission from A. Heilmann, Fraunhofer Institute, Halle, Germany.

surfaces are obtained when they are pretreated by ion beam milling. Comparisons between EM and AFM measurements have indicated a small difference in pore diameter, of about $0.18\,\mathrm{nm\,V^{-1}}$, the reason for this being seen as a method-specific image fault, with nanometer-sized structures on surfaces becoming enlarged while the cavities appear reduced in extension. The reasons for this well-known phenomenon are explained in Figures 5.14 and 5.15, where the numbers 1 to 3 indicate typical positions of the tip's route over the surface. Compared to the size of the object, the curvature radius of the tip is larger. As a consequence, the lateral extension of an isolated object may appear to be too large, whereas the height of the particle can be measured more or less correctly.

The pores in AAOs can be considered as the negatives of particles. Depending on the tip radius, the tip will enter too late into the pore space and be repelled too early, due to contact of the tip wall with the pore wall (see Figure 5.15).

In the case of ion beam-milled surfaces, an additional special effect must be considered since, due to the low angle at which the ion beam meets the surface, the upper areas of the pore walls will be additionally thinned and alter the detail of the AFM (see Figure 5.16) [13].

The scanning curve is characterized by two steps, caused first by the thinner pore wall and then by the original wall; this can be seen from a cross-section through an AFM image of an ion beam-milled surface (Figure 5.17). Occasionally, the step from the thinned upper pore wall to the original wall can be identified by additional shoulders in the cross-section line.

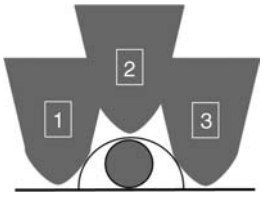


Figure 5.14 Schematic depiction of the scanning line of an atomic force microscopy (AFM) tip over an isolated particle on a smooth surface. The numbers 1 to 3 indicate typical positions of the tip's route over the surface.

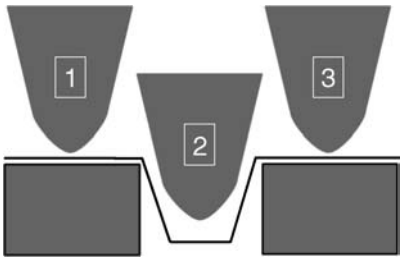


Figure 5.15 Sketch of the AFM tip motion in contact with a nanopore.

Hence, AFM images can become falsified not only by the unavoidable imaging problem of the method, but additionally by any preceding treatment of the surface. An AFM image of a hexagonally ordered 40 V pore structure after ion beam milling is shown in Figure 5.18.

AFM has also been applied to image the reverse side of AAO membranes with an intact barrier layer, and this can be achieved by the oxidative dissolution of the aluminum carrier. An AFM image of a membrane lower side, indicating the polyhedral round forms of the barrier layer (see also Figure 5.4) is shown in Figure 5.19 [13].

The differences in pore diameter for TEM and AFM characterizations and eventual pretreatments are summarized in Table 5.2.

5.4 Aluminum Oxide Membranes as Templates

One of the most attractive applications of nanoporous alumina membranes is in using the ordered pores as templates, to generate correspondingly ordered arrays of 2-D wires, tubes, or particles. As a result, numerous activities have emerged during the past decade, with polymers [23], carbon [24–27], semiconductors [28–30],

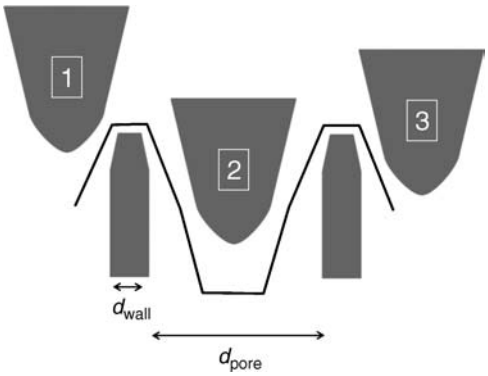


Figure 5.16 Sketch of the AFM tip motion in contact with ion beam-milled surfaces in detail.

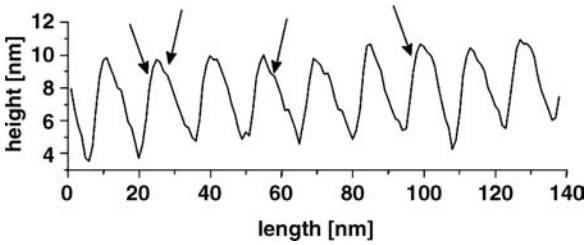


Figure 5.17 Cross-section through an ion beam-milled 40 V anodized aluminum oxide (AAO) membrane. The arrows indicate steps from the thinned upper pore wall to the original wall (cf. Figure 5.16.

metals [31–33] and various other materials having been arranged in the pores of AAOs [34]. In general, the tubes or wires generated can be separated from the alumina template by dissolving the latter with an acid or a base.

The variability of the pore diameters and pore lengths is of enormous advantage when tuning the products. However, if membranes with barrier layers are used as templates, then the resultant wires or tubes cannot be positioned directly on a surface in an upright position. However, this does become possible if membranes that are open on both sides are used, and this can be achieved either by removing the oxide barrier layer with a phosphoric acid solution [35], or by ion beam milling. In the following sections, examples are provided where tubes, wires or dots are placed on supports, leading to the creation of novel nanostructured surfaces (as the title of this book might suggest!).

5.4.1 Carbon Nanotubes and Nanowires

Highly ordered arrays of carbon nanotubes (CNTs) can be fabricated by the pyrolysis of acetylene on cobalt catalysts, inside hexagonally ordered pores of aluminum oxide membranes at 650 °C [25]. In this way, large-scale arrangements of densely packed CNTs become available, with diameters ranging from 10 nm to several hundred nanometers, and lengths of up to 100 μm. CNT arrangements of such quality are of

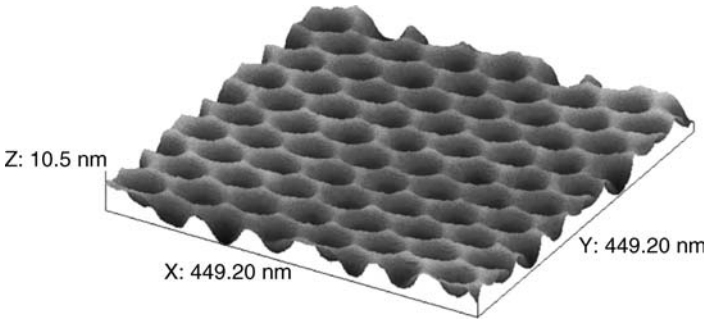


Figure 5.18 Atomic force microscopy image of a perfectly ordered hexagonal AAO surface.

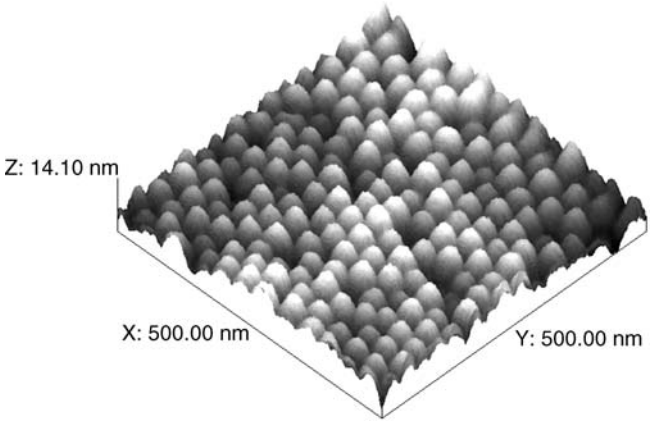


Figure 5.19 Atomic force microscopy image of the reverse side of a barrier layer, generated at 25 V.

general interest for future applications in nanoelectronics, such as data storage, field emission displays, or sensors.

In contrast to the above-described bodies, CNTs as well as wires of diamond structure may become available if microwave plasma-assisted chemical vapor deposition (CVD) is used with acetone as the carbon source, instead of the simple thermal decomposition described above [27]. Diamond tubes and wires exhibit totally different properties compared to CNTs consisting of graphitic structures, with the high stability and the negative electron affinity allowing novel applications in both nano- and microelectronics. The creation of low-field emitters [36–38] or cold cathode flat-panel displays [39, 40] has been proposed by using such diamond species.

Both, aligned polycrystalline diamond nanowires and diamond-like nanotubes can be created using AAO membranes as templates, the main advantage being their variability. The anodization of a 0.15 mm Al sheet in 0.3 M H_3PO_4 at 1 °C under a voltage of 190 V for 70 min resulted in pores which were about 7 μm long and 300 nm wide. In order to open the down-side of the membrane, the aluminum is first dissolved with $HgCl_2$, after which the barrier layer is removed using phosphoric acid [35]. Such through-hole membranes are perfectly suited for the generation of diamond tubes or wires, the individual production steps of which are depicted in Figure 5.20.

Table 5.2 Observed diameters (*d*) depending on the method of investigation and pretreatment by ion beam milling [13].

Method	Diameter (*d*) (nm)
TEM	$1.37\,\mathrm{nm\,V^{-1}} \cdot U + 0.36\,\mathrm{nm}$
AFM, untreated	$0.72\,\mathrm{nm\,V^{-1}} \cdot U + 0.07\,\mathrm{nm}$
AFM, untreated (ion beam)	$1.19\,\mathrm{nm\,V^{-1}} \cdot U + 0.02\,\mathrm{nm}$

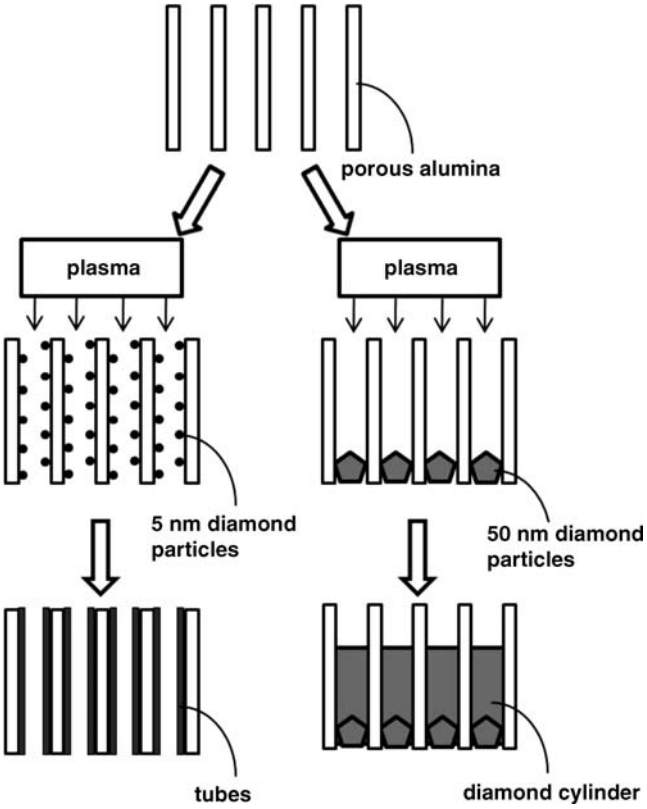


Figure 5.20 Schematic of the generation process of diamond-like nanotubes (left) and nanowires (right).

The first step consists of depositing ultrasonically dispersed 5 nm diamond particles on the pore walls to create tubes, and of about 50 nm diamond particles at the bottom of the pores to create wires. In both cases, the deposited diamond particles serve as seeds for the growth of further particles, resulting in tubes and cylinders, respectively. The as-prepared alumina membranes are then placed into a microwave plasma chamber whereby, at a microwave power of 3 kW, 80 Torr with hydrogen as carrier gas and temperatures of about 1000 °C are achieved. A reaction time of 8 h leads to an arrangement of parallel tubes or wires with an outer diameter of 300 nm, corresponding to the pore diameters, with both the tubes and cylinders standing upright and parallel on the substrate after dissolution of the alumina with concentrated phosphoric acid. The substrate also consists of a diamond film formed during the deposition process on the membrane. The SEM image of diamond-like nanotubes is shown in Figure 5.21, whereas Figure 5.22a shows a cross-sectional field of parallel diamond nanowires, and Figure 5.22b shows the wire arrangement from the top, indicating a clear hexagonal arrangement [27].

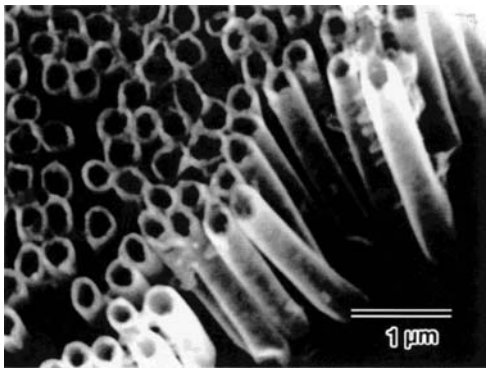


Figure 5.21 Scanning electron microscopy image of diamond-like nanotubes of 300 nm diameter and about 7 μm length [27].

As already mentioned, diameters and lengths of tubes and wires can be varied along with the pore geometries. Based on the Raman spectra of diamond wires, a broad signal at 1440 cm^{-1} indicates the presence of some sp^2 carbon, whereas the diamond peak is observed at 1334 cm^{-1}, as expected.

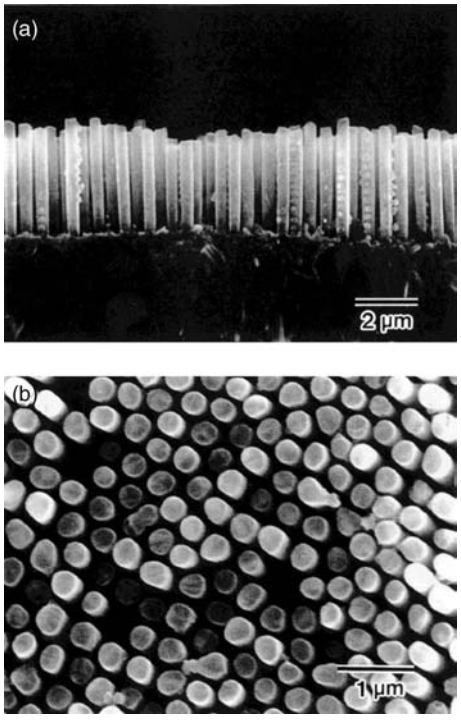


Figure 5.22 Scanning electron microscopy image of diamond nanowires visualized (a) from the side and (b) from the top. The diameter (300 nm) and length (7 μm) correspond to the alumina pores [27].

The procedure also allows doping of the diamond wires, for instance with boron. In this case, the originally colorless and transparent wires become dark blue.

5.4.2
Metal Nanotubes and Nanowires

Whereas, the simple filling of porous aluminum oxide membranes by using electrochemical or chemical techniques is relatively well known, the fabrication of free-standing aligned metal nanotubes on substrates remains very limited. However, by applying appropriate templates and synthesis procedures, several examples have been identified in recent years, some of which are discussed in the following subsections.

Well-ordered, 35 μm-long Ni nanotubes can be generated by electrodeposition using commercially available alumina membranes (Anapore) [41], although to achieve this the pore walls must first be chemically modified with methyl-γ-diethylenetriaminopropyl-dimethoxysilane; this allows nickel to be deposited preferentially onto the walls, as a result of strong interactions between the metal and the amino groups. Without such pretreatment of the pore walls, Ni nanowires would be formed instead of tubes. At current densities of 0.6 mA cm^{-2} and Ni^{2+} ions from $NiSO_4$ in a boronic acid-containing aqueous solution, Ni nanotubes of outer diameter 160 ± 20 nm can be generated. However, as the length of the tubes and their wall thickness will depend on the experimental conditions employed, and can vary over a wide range, it is possible to obtain pore wall thicknesses ranging from 30 nm to 60 nm. Moreover, if the membrane material is removed by aqueous NaOH, then arrays of Ni nanotubes are obtained, as shown in Figure 5.23.

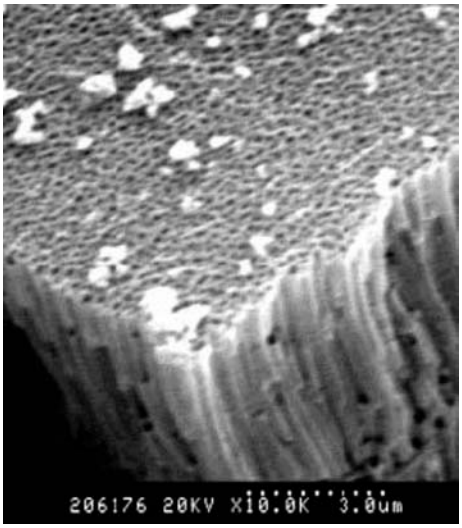


Figure 5.23 Scanning electron microscopy image of free-standing, densely packed Ni nanotubes of 160 ± 20 nm outer diameter [41].

Magnetic measurements of these Ni nanotube arrangements with aspect ratios of ~200 indicate enhanced coercivities compared to the bulk nickel. Strong inter-tube interactions are also observed.

Various metal nanotubes can also be synthesized by a different procedure in AAOs [33]. In contrast to the above method with chemically modified pore walls, the alternative consists of the pre-deposition of CNTs onto the pore walls of the alumina membranes via the pyrolytic decomposition of ethyne [42]. By using a gold film on one side of the membrane as an electric contact, the electrodeposition of Ni first results in the formation of Ni wires with CNT walls. The next step comprises a thermal treatment of these AAO/CNT/Ni systems at 400 °C in air, so as to oxidize the Ni wires to NiO; an increase in temperature to 600 °C then burns up the CNTs. As a result of this procedure, highly ordered nanochannels of the former CNT dimension are yielded. In this way, various metals such as Pt, Au, Bi, In, and Ni can be electrodeposited into the annular nanochannels, and this will result in a coaxial sandwich system with NiO cores and metal walls inside the aluminum oxide membrane. The NiO and Al_2O_3 can then be chemically dissolved to leave the free-standing metal nanotubes. The different reaction steps, from the alumina membrane to the nanotube array, are shown schematically in Figure 5.24.

By having CNT diameters of 50 nm and a wall thickness of 10 nm, it is possible to produce metal nanotubes of the same dimensions; an SEM image of such Pt nanotubes of 30 μm length is shown in Figure 5.25.

The deposition of metal nanowires in the pores of alumina membranes usually occurs if a direct current (DC) is used. However, the first step involves a detachment

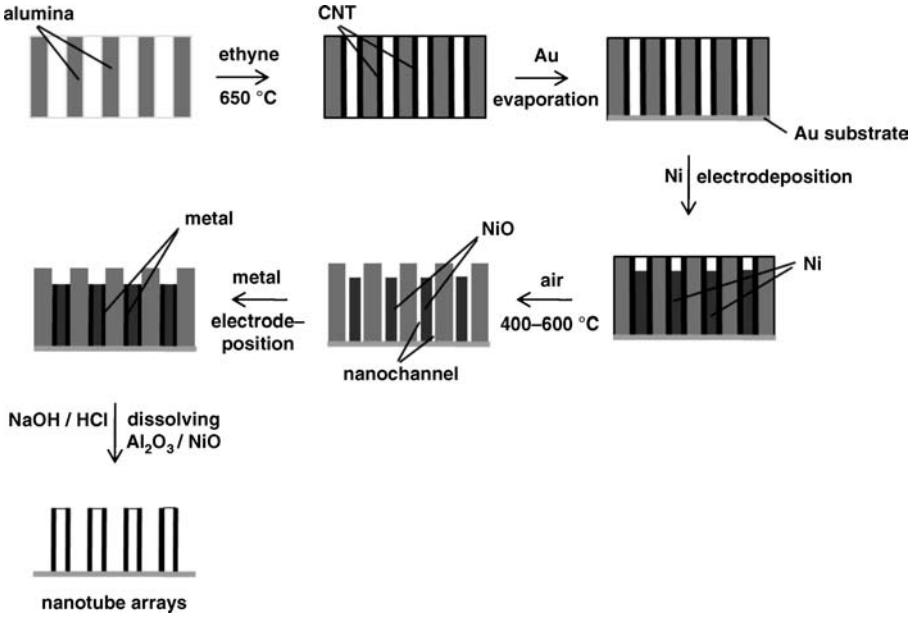


Figure 5.24 Schematic of the different steps used to fabricate nanotubes from various metals.

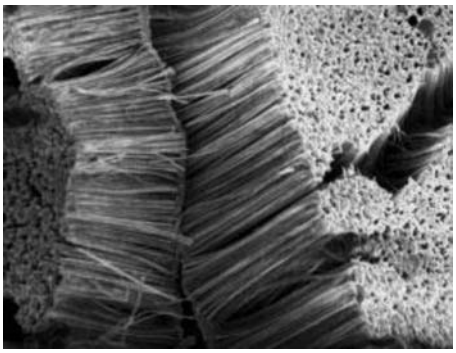


Figure 5.25 Scanning electron microscopy image of aligned Pt nanotubes, with wall thickness ca. 10 nm, and total diameter ca. 50 nm [33].

of the membrane from the aluminum support, and etching of the barrier layer (as discussed above). Since a metallic contact must be created on one side of the membrane, the latter must be thick enough to be free-standing [43, 44]. Another procedure involves the use of an alternating deposition potential, which causes the metal nanowires to be deposited on the barrier layer of the membrane, thus avoiding any membrane pretreatment [45–48]. This method has recently been adopted by using highly ordered alumina membranes [31]. Pore arrays of a high aspect ratio are fabricated by using an appropriately long anodization [35, 49] whereby, after generating the membrane, the barrier layer and pore walls are thinned by isotropic chemical etching, followed by two current-limited anodization steps. The remaining aluminum support can then be used as an electric contact for the electrodeposition of metals from metal salt solutions [31].

Pulsed electrodeposition under a modulated voltage control has been shown to produce the best results if compact metal nanowires are fabricated in the membranes [31, 50–53]. Despite the creation of numerous metal nanowires inside the aluminum oxide membranes, only a limited number of examples exists, where parallel and free-standing wires on a support have been obtained after detachment of the membrane material, representing a nanostructured surface. Rather, in most cases randomly oriented wires, lying on the support, are produced. On the other hand, numerous other methods to generate ordered metal or semiconductor nanowires on surfaces have been developed, albeit via other methods that are not the object of this chapter (e.g., see Ref. [54]). Two examples of metal nanowires plated substrates by means of AAO membranes are discussed in the following subsections.

Perfect 2-D arrays of Cu nanowires on silicon supports have been prepared by the anodization of an Al film sputtered onto a Si/SiO_2 substrate (step a $\rightarrow$ b in Figure 5.26). The aluminum oxide membrane was then etched away with phosphoric acid/chromic acid so as to produce an array of nanoholes in the Al film (step b $\rightarrow$ c in Figure 5.26). In a second anodization step, carried out at 40 V until the aluminum was totally oxidized, the final perfect pore structure was formed [55] (steps c $\rightarrow$ d in Figure 5.26). Deposition of Cu in the pores proceeds best by nonselective electroless plating (step d $\rightarrow$ e).

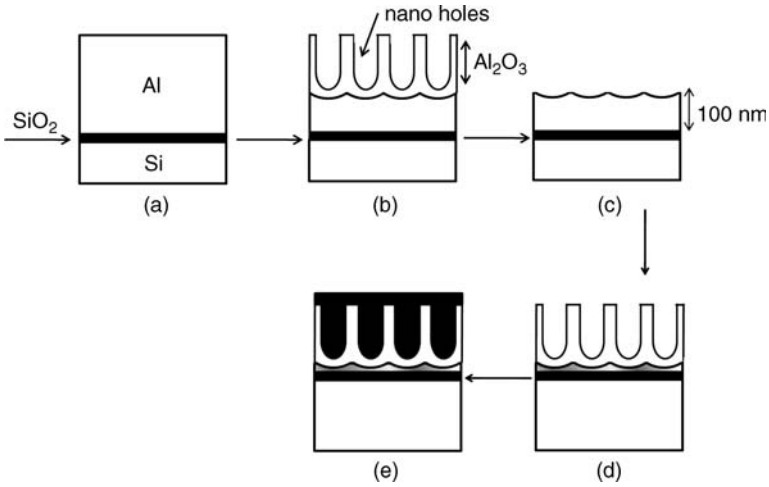


Figure 5.26 The generation of copper nanowires on a Si support. See text for details.

The deposition of Cu in the pores succeeds best via nonselective electroless plating. In this case, after having activated the alumina surface by $PdCl_2$, copper is deposited from $CuSO_4 \cdot 5H_2O$, while the pores are completely filled with Cu if the aspect ratio is 2.5. As the Cu nanowires are connected by a copper film on top, this film can be delaminated with the wires, using a Scotch tape and pulling from the Si substrate.

A similar generation of free-standing copper nanowires has succeeded via nanoporous AAO membranes on top of a metallic substrate, followed by the deposition of copper inside the pores from a plating bath of specific formulation and dissolution of the AAO template [56].

Antimony nanowires can be fabricated in anodic alumina membranes by pulsed electrodeposition under modulated voltage control at 40 V [57]. In this case, the aluminum support and barrier layer could be removed chemically, followed by the sputtering of a gold film onto one side of the membrane as the working electrode. As an antimony precursor $SbCl_3$ was used, the upper part of the alumina was dissolved with NaOH solution following electrodeposition of the Sb wires.

5.4.3
Semiconductor Nanowires

The application of nanoporous alumina membranes is, of course, not limited to the fabrication of carbon or metal nanotubes and nanowires, but has been extended to semiconductors and to other materials (see Section 5.4.4). Some examples of typical semiconductor nanowires are provided in the following subsections.

The generation of silicon nanowires can, among others, be based on a so-called vapor–liquid–solid (VLS) growth [58]. In this case, a gaseous Si source (SiH_4 or $SiCl_4$) is thermally decomposed at an appropriate temperature by contact with a metal catalyst (such as Au), and then diffuses into the metal to produce a liquid alloy. On

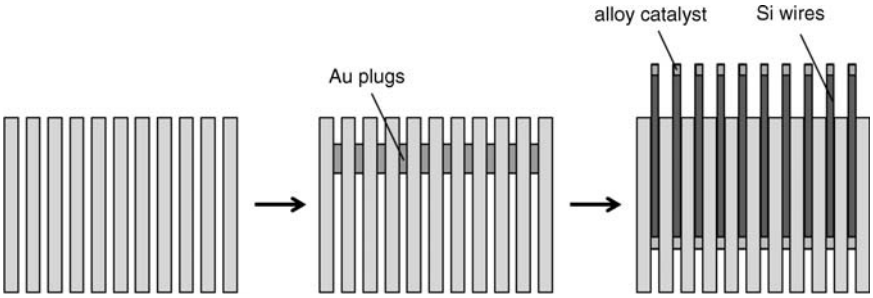


Figure 5.27 Schematic representation of the fabrication of Si nanowires on top of an alumina membrane.

reaching supersaturation, the Si nanowires begin to precipitate and grow underneath the liquid alloy droplet. Whilst the principle of this process has long been known [59–64], it can also be applied by using nanoporous alumina membranes as templates [58], whereby membranes of various pore diameters, equipped with a thin Ag film on the top side, are used to electrodeposit Au plugs inside the pores, and to act as catalysts for the VLS process. When the VLS process is started the Si wires grow first inside and finally outside of the pores, such that the membrane is decorated with Si nanowires. The principle of the system is depicted in Figure 5.27, and an example of such a nanostructured surface is shown in Figure 5.28.

Currently, TiO_2 is of increasing interest with regards to its semiconductor properties, most recently in photochemical solar cells. As TiO_2 nanowires are believed to play a special role in these developments, the routine production of TiO_2 nanowires *per se* and nanowires decorating different supports continues to attract increasing attention. Indeed, nanoporous aluminum oxide membranes might again play an important role in routine fabrications, with 40 V membranes being applied to TiO_2 production via a sol–gel process [65]. In this case, a TiO_2 sol can be prepared from tetrabutyl titanate, ethanol, acetic acid, and water. A through-hole 40 V membrane, generated by etching the Al support by $HgCl_2$ and dissolving the barrier layer with phosphoric acid, is then immersed gradually in the gel, followed by heat treatment. Temperatures of up to 650 °C finally cause the TiO_2 to be transferred to anatase quality. Previously, free-standing nanowires of up to 20 μm have been created by dissolving part of the membrane material from one side with phosphoric acid.

Cadmium chalcogenides belong to the most attractive – and therefore increasingly investigated – group of materials in nanotechnology; CdS is representative of this group. Whereas, CdS films and nanoparticles are well-known objects of research, CdS nanowire arrays have undergone much less development, due mainly to their more complicated fabrication. Yet, aluminum oxide membranes represent a good opportunity to close that gap. By following a long-recognized method for preparing CdS from Cd^{2+} and elemental sulfur [66] in dimethyl sulfoxide (DMSO), CdS nanowires can be generated in AAO pores by electrochemical deposition [28]. For this, membranes with pore diameters between 9 and 35 nm, together with the original Al support, are used as the working electrode, with graphite as the

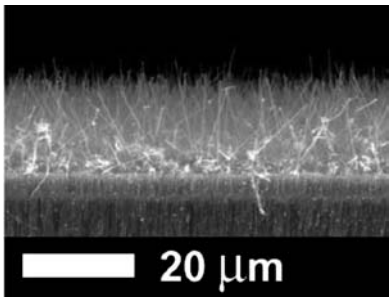


Figure 5.28 Scanning electron microscopy image of Si nanowires grown from the pores of a 50 V alumina membrane.

counter-electrode, using an alternating current. However, if the Al support and the barrier layer are removed by the above-mentioned methods, then a metal electrode can be positioned, after which a direct current electrodeposition would become possible [67, 68]. Detachment of the alumina material by NaOH has resulted in free-standing wires with diameters that corresponded to the pore diameters, and lengths of up to 1 μm.

5.4.4
Other Materials

Besides carbon, metals and semiconductors (as discussed above), a series of other materials can be used to create nanotubes or nanowires in the pores of alumina membranes. Boron nitride (BN), a typical nonconductor, is briefly described as an example in the following subsection. By using trichloroborazine, polycrystalline BN nanotubes were prepared, using CVD, when gaseous borazine was pyrolyzed at 750 °C over the open pores of commercially available membranes (Anodisc) [29]. In the case of borazine (which has the advantage of already having the correct B : N ratio), the nanotubes were able to reach lengths of up to 20 μm with, as usual, parts of the membrane material being dissolved in NaOH to produce free-standing nanotubes on the alumina support.

Coaxial C/BN/C nanotubes can be generated by the sequential pyrolysis of acetylene and trichloroborazine. The filling of BN nanotubes with metals has also been shown possible via an electrochemical deposition of copper into BN tubes that resulted in the production of BN-coated Cu nanowires [68].

5.4.5
Nanoparticles

Clearly, in addition to fabricating nanotubes or nanowires of various materials, it is also possible to generate assemblies of nanoparticles by the use of nanoporous alumina membranes. Although a recent report has described much of the detail relating to this subject [69], the discussion here will relate only to certain principles, as

the technical procedures for preparing nanoparticle arrangements rather than nanowires are only marginally different. For example, separately fabricated ultra-thin through-hole alumina membranes can be used equally well as can membranes generated directly on the smooth surfaces of interest and coated in advance with a thin aluminum film [70]. Consequently, semiconductor, metal or metal oxide nanoparticles can each be deposited through the short pores on the substrates, followed by dissolution of the alumina masks (as per usual). Generally, the arrangement of pores in pre-prepared membranes is of a better quality than that produced by connected Al films. In fact, the best results with connected membranes have been achieved using rather thick Al films (20 μm), though this is a clear disadvantage for the subsequent deposition of nanoparticles. Thinner Al films (0.8 μm) can be used if pre-texturing with SiC or Si_3N_4 is carried out in advance [71, 72], such that the formed concave imprints serve as nucleation points in the subsequent anodization process. The use of such connected membranes means that the removal of any barrier layer between the pores and substrate must first be mastered, although several procedures have been described to overcome such a problem [31, 71–77].

Currently, most deposition procedures are based on vapor-phase techniques such as electron-beam evaporation, sputtering, molecular beam epitaxy, CVD, or pulsed laser deposition [69]. The use of membrane masks, directly fabricated onto a substrate, allows also wet chemical procedures such as electrochemical and electroless deposition to be carried out, due to their strong connection with the underlying substrate.

5.5 Nanoporous Alumina Membranes as Imprinting Tools

Among the principal techniques to fabricate micro- and nanostructured surfaces, which include lithography, self-assembly, controlled deposition, size reduction or replication by physical contact, the lithographic methods play a major role in the semiconductor industry [78–80]. Although lithographic methods are generally limited by the wavelength of the applied irradiation, the so-called "nanolithographic" techniques have meanwhile opened the door to higher-quality nanostructured surfaces by providing the to write structures on surfaces, using an AFM tip [81].

Among these various techniques for fabricating micro- or nanostructured surfaces, imprinting processes with stiff masks play a dominant role. The expression "imprinting" infers that a mask is pressed into a substrate to form an inverse 1 : 1 image. For instance, compact discs (CDs) are produced by imprinting polycarbonate discs by nickel matrices [82]. In another example of a successful imprinting process, a SiO_2-covered Si wafer is structured with EBL, and then used to imprint a poly (methylmethacrylate) (PMMA)-coated Si wafer. The holes produced can then be filled by metals, followed by dissolution of the PMMA, such that a surface with metal islands results [83].

Replication techniques which use nanoporous aluminum oxide membranes as an imprinting tool are details in the following subsections.

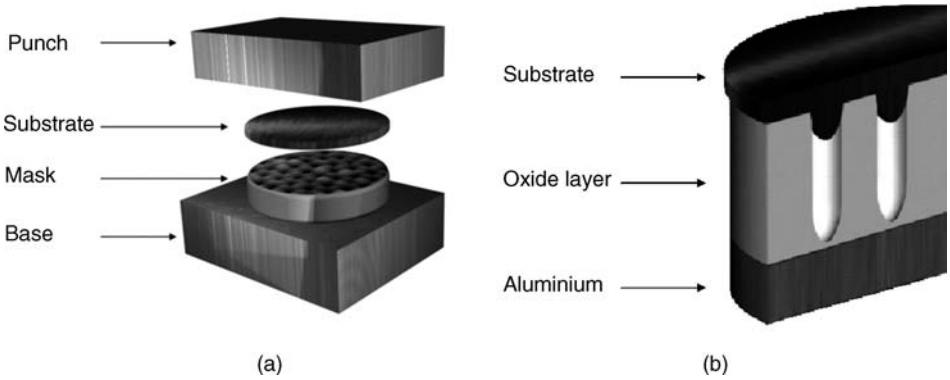


Figure 5.29 (a) Schematic of the imprinting device; (b) Material flow into the nanopores. Note: the pores shown on the mask in panel (a) were enlarged in order to render them visible.

5.5.1
Imprinting Techniques and Conditions

The alumina membranes can be applied as stamps in different forms, with sheets, discs, or foils each suited to imprinting other materials. Whilst one-sided anodized discs can be used directly as stamps, foils must be placed on a stamp of another material. However, foils have the advantage of being more flexible than hard discs, and so cracking of the porous layer can better be avoided; in fact, even re-use is possible. Details of the imprinting process are shown in Figure 5.29a and b [84, 85]. In this case, the alumina mask and substrate to be imprinted are positioned between the punch and the base of the press device. The material flow into the pores is shown in Figure 5.29b. Usually, the surface to be nanostructured is softer than the alumina, but if this is not the case then raising the temperature may help to achieve an appropriate softness.

The perfect 1 : 1 transfer of a porous stamp onto a surface can best be visualized if some irregularities exist, as is the case with the example in Figure 5.30. Here, a 150 V membrane with some fractures and other defects has been used to imprint a Pd

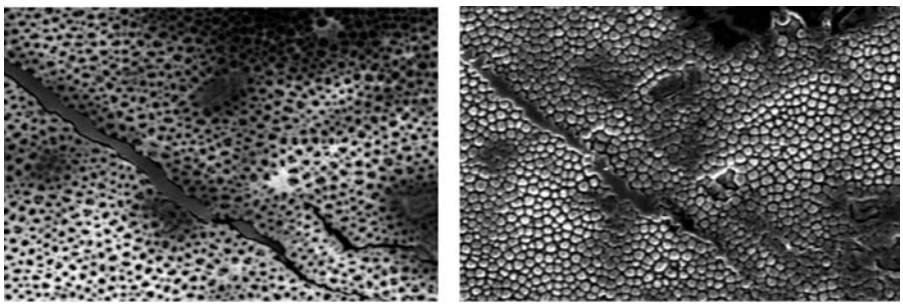


Figure 5.30 Scanning electron microscopy images of (a) a nonperfect 150 V alumina membrane, and (b) an imprinted Pd surface, showing the same defects.

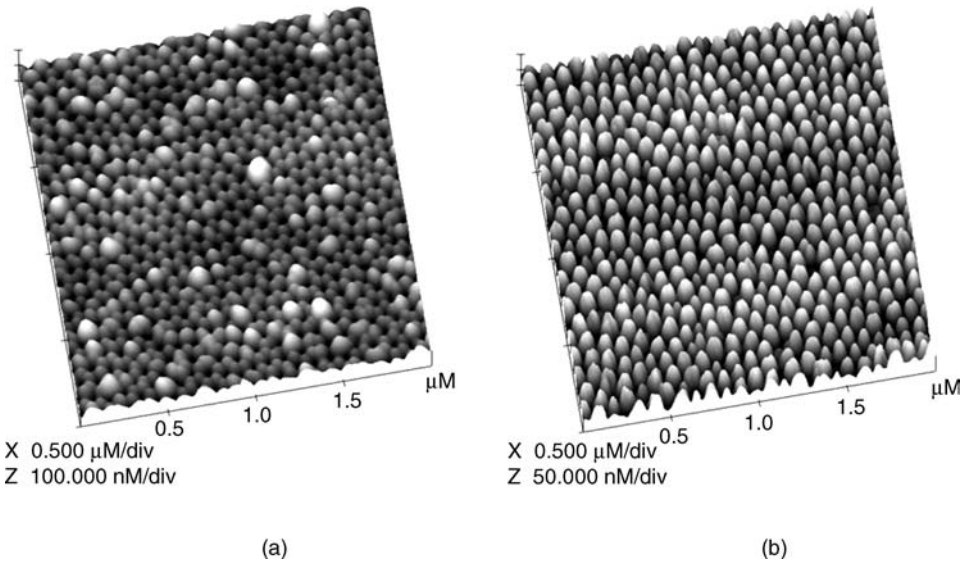


Figure 5.31 Atomic force microscopy image of (a) a 40 V (50 nm) alumina membrane, and (b) the corresponding imprinted PMMA surface.

surface. It can easily be seen that, besides the pores themselves, all defects of the membrane are transferred to the Pd surface [84].

By using a high-quality membrane for imprinting a PMMA surface at 110 °C, combined with a pressure of 80 MPa for 60 s, the result was a near-perfect inverse image of the pore structure (see Figure 5.31), where the hexagonal structure of the pores is seen to be transferred to the PMMA surface. Whereas, the pore and the pillar density, the diameters and the distances between membrane and PMMA surface agree quite well, the heights of the PMMA pillars deviate characteristically from the lengths of the pores. This can be followed from height profiles [84]. This effect is not caused by a break of the PMMA pillars during the separation of the mask from the substrate, but rather is due to an incomplete filling of the pores by the polymer. The average pillar height in this case lies between 40 and 50 nm, giving an aspect ratio of approximately 1.

The results of imprinting depend mainly on the mechanical properties of the masks and substrates. The mechanical properties of metals are given by a series of characteristic values, such as hardness, breaking tension, break constriction, breaking strain, pressure resistance, and bending strength. These values relate to well-known properties such as toughness, ductility, brittleness and wear resistance, and provide information regarding the general properties of a material. If an external force is applied at a metal stick, it is elastically elongated in a reversible process that is described by Hooke's law:

$$\sigma = E \cdot \varepsilon \tag{5.6}$$

where σ is the tension (in Pa), E is the elasticity module (in Pa), and ε is the elongation ($\Delta l/l$) [86].

If the so-called "flow limit" is overcome by an external force, a metal stick would be irreversibly deformed and Hooke's law no longer valid, such that finally the break point would be reached. The special ductility of metals is a consequence of the bonding situation in metals, characterized by the existence of freely mobile electrons (electron gas). The mechanical properties of the metals that have been used for imprinting (see Section 5.5.2) are summarized in Table 5.3 [87–89].

As can be seen from these data, properties such as flow limit, breaking strain or hardness can vary for the same metal, depending on its pretreatment. The metals used in the processes described below for imprinting are classified as "hard" due to their pretreatment. In spite of certain differences in their mechanical properties, the metals used behave in identical fashion with respect to the irreversibility of the imprinting-determined structures. The ductility of all metals used is sufficient for their surfaces to be imprinted at room temperature.

In contrast, the mechanical properties of polymers differ greatly from each other compared with metals, due to their chemical composition. As the polymer chains of thermoplasts are not linked among each other (in contrast to elastomers and duromers), thermoplasts should be better suited for imprinting. Above the glass-point the mechanical stability decreases rapidly; however, as the glass-point depends not only on the molecular masses but also on the crystallinity, technical polymer products usually do not have a distinct glass-point but rather a transition region. This

Table 5.3 Mechanical properties of the metals used for imprinting with alumina membranes.

Metal	Flow limit (MPa)	Breaking strain (MPa)	Modulus of elasticity (GPa)	Hardness (Vickers)
Aluminum (99.999%)	10–35 (soft) 110–170 (hard)	50–90 (soft) 130–195 (hard)	70.6	21 (soft) 35–48 (hard)
Aluminum (99.5%)	—	68–127	—	—
Aluminum (99.0%)	—	98–157	—	—
Lead	5.5	12	16.1	5
Iron	120–150	180–210	211.4	—
Copper	54 (soft) 270 (hard)	224 (soft) 314 (hard)	129.8	49 (soft) 87 (hard)
Brass	300–700	—	110–115	—
Nickel	150 (soft) 480 (hard)	400 (soft) 600 (hard)	199.5	75
Palladium	34.5 (soft) 205 (hard)	140–195 (soft) 325 (hard)	121	40 (soft) 100 (hard)
Platinum	14–35 (soft) 185 (hard)	125–150 (soft) 200–300 (hard)	170	40 (soft) 100 (hard)
Silver	—	172 (soft) 330 (hard)	82.7	25 (soft) 95 (hard)
Zinc	—	37	104.5	30

Table 5.4 Mechanical properties of the polymers used for imprinting.

Polymer	Elongation at break (%)	Breaking strain (MPa)	Modulus of elasticity (MPa)
PC	100–150	55–75	2.3–2.4
PE	—	15–40	0.5–1.2
PMMA	1.5–4.0	80	2.4–3.3
PTFE	400	7–20	0.3–0.8

PC = polycarbonate; PE = polyethylene; PMMA = poly(methylmethacrylate); PTFE = polytetrafluoroethylene.

can be seen from the data in Table 5.4, in which the mechanical data of the investigated polymers are collected [87].

Compared to metals, the elasticity moduli and breaking strains of polymers are smaller. Indeed, there are considerable differences in the plasticity of polymers at room temperature; for example, polytetrafluoroethylene (PTFE) can be formed at room temperature, whereas PMMA will break unless it is warmed up.

Besides the mechanical properties of the materials to be imprinted, those of the imprinting mask (in this case, aluminum oxide) are equally important. The hardness of nanoporous alumina layer lies between 320 and 360 (Vickers), which is much higher than that of the materials to be imprinted. In contrast, the brittleness of ceramics is generally high, and therefore oxide layers may be destroyed under pressure if this is not carefully adjusted.

Finally, the behavior of alumina stamps is also dependent on their mechanical constitution. For example, anodized aluminum discs, when used at 100 MPa for imprinting PTFE over a 20 s period will become deformed to some extent, thus reducing the pressure to 70 MPa [84]. As a consequence, the numerous fractures that have formed will be transferred to the substrate (see Figure 5.32).

Bursting of the oxide layer can occur for different reasons, one of which is a lack of perfect coplanarity at any position if extended stamps and substrates are used. A

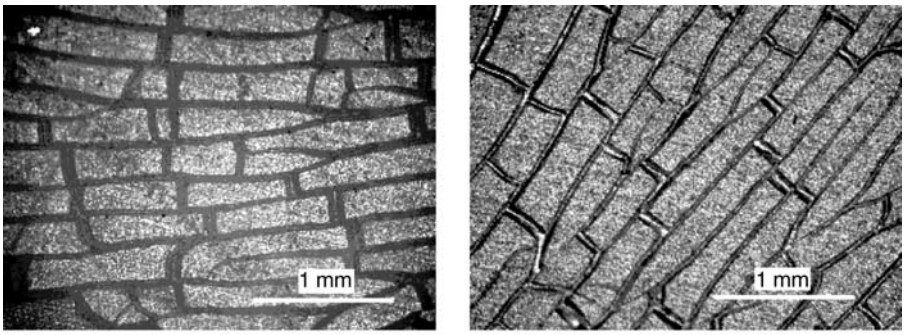


Figure 5.32 Light microscopy images of (a) an anodized aluminum disc (70 V) after imprinting, and (b) and an imprinted PTFE surface.

material which is capable of flowing can equilibrate such unevenness, but not in the case of a thick disc. The main reason for such roughness is the use of insufficiently pretreated technical aluminum surfaces. However, even electropolishing cannot fully prevent such defects, and in fact even if perfectly smooth surfaces are used, crack formation may occur if the flow limit of aluminum is exceeded. The consequent flow of the metal beneath the oxide layer causes tensions to be built up, and this results in crack formation.

Such disadvantages of aluminum discs can be prevented to a large extent if aluminum foils, oxidized on one side, are used for imprinting [84]. In this case, not only will the flow processes be reduced, but so too will the tendency towards crack formation of the oxide layer.

Alumina membranes that have been detached from the aluminum substrate and then used as stamps are not suited for imprinting processes, as the membrane usually fractures under pressure.

5.5.2
Imprinting of Metal Surfaces

The hardness of alumina allows the imprinting of many metals and alloys and, as noted above, the quality of the structured surfaces depends heavily on their pretreatment.

Aluminum can be imprinted by its own oxide. The surface quality is best pre-prepared by electropolishing, with Al sheets and discs being polished by anodization at 75 °C in highly concentrated electrolyte solutions for 10 min, whereas for Al foils (75 μm) only 1 min at 10 V is required. Due to the elevated temperature, the oxide layers are removed continuously from the surfaces, after which any traces of aluminum oxide are dissolved using chromic acid and phosphoric acid [84, 85]. Despite this careful smoothening of the Al surfaces to be imprinted, a nanosized roughness can still be observed. The AFM images of an as-prepared Al surface, before and after imprinting at 250 MPa, are shown in Figure 5.33. Here, the aluminum pillars have an average height of 200–230 nm and an aspect ratio of about 1. The pillars on the imprinted surface (Figure 5.33b) clearly show an additional fine structure, caused by the nanosized roughness of the electropolished surface. In fact, for applications where large surfaces are advantageous (as in catalysis), this effect might be very welcome.

This example demonstrates how the substrate surface influences the imprinting process. Although the flow limit of the metals has been exceeded, the locally different distributions of pressure have resulted in different degrees of transformation.

Lead surfaces may also be prepared, although problems in obtaining smooth surfaces have produced poor results. Due to the special ductility of lead, its surfaces may become partially cracked under high pressure, as identified from SEM images [84].

The transition metals *iron*, *nickel*, *palladium*, and *platinum* can be nanostructured at pressures of between 1000 and 1500 MPa, while *copper* and *silver* have been successfully nanostructured by 50 nm and 150 nm pore masks at 985 MPa [84, 85].

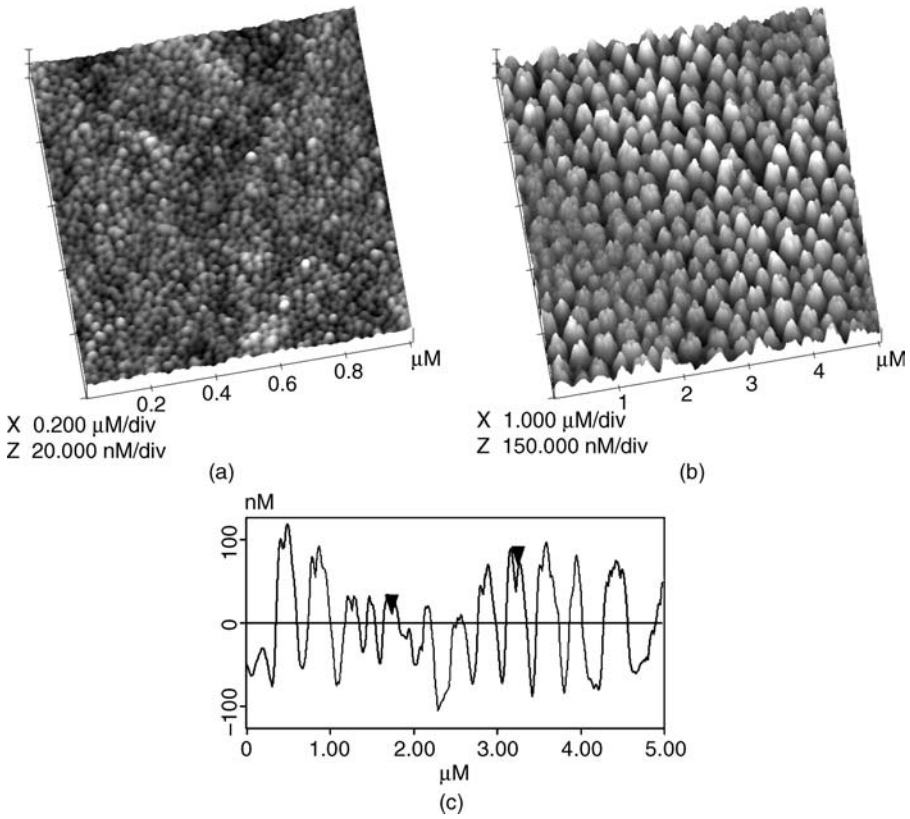


Figure 5.33 Atomic force microscopy images of an electropolished Al surface (a) before and (b) after imprinting with about 180 nm pores, at a pressure of 250 MPa; (c) A cross-section, indicating the multiple tips of the Al pillars.

An AFM image of a silver foil, imprinted with a 50 nm pore alumina mask, is shown in Figure 5.34, where this has resulted in Ag pillars with aspect ratios of between 0.8 and 1.5.

In the case of *zinc* and *brass*, Zn can be easily imprinted at 500 MPa, whereas brass (which is much harder) requires a pressure of 985 MPa to be imprinted with 200 nm pore masks. An AFM image and the corresponding gray-scale shading image of an imprinted brass surface with pillars of an aspect ratio of 0.8–1.4 are shown in Figure 5.35a and b [84]. Here, the gray-scale image (Figure 5.35b) results from the deviation S_x of the height h in the scanning direction x (arrow). The 3-D impression results from the fact that the structure heights, which increase in the scanning direction, appear light (right side), whereas decreasing structure heights on the opposite side become dark (left side). Although, from such images the topography can better be perceived, this type of imaging does not provide any information regarding the structure heights.

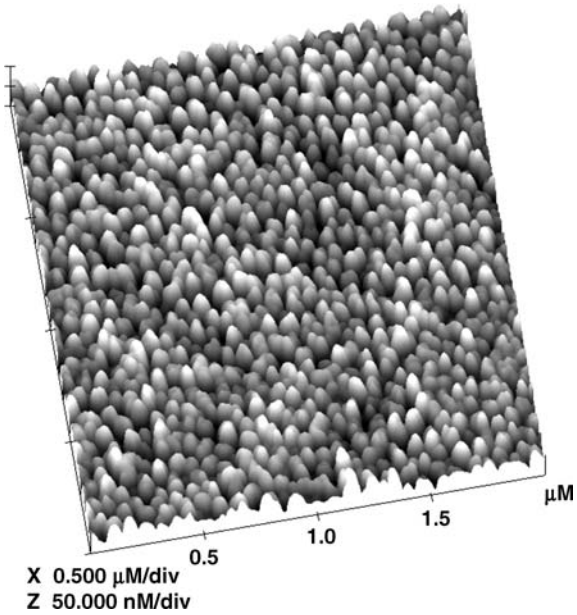


Figure 5.34 Atomic force microscopy image of a nanostructured Ag surface (mask: 50 nm pores).

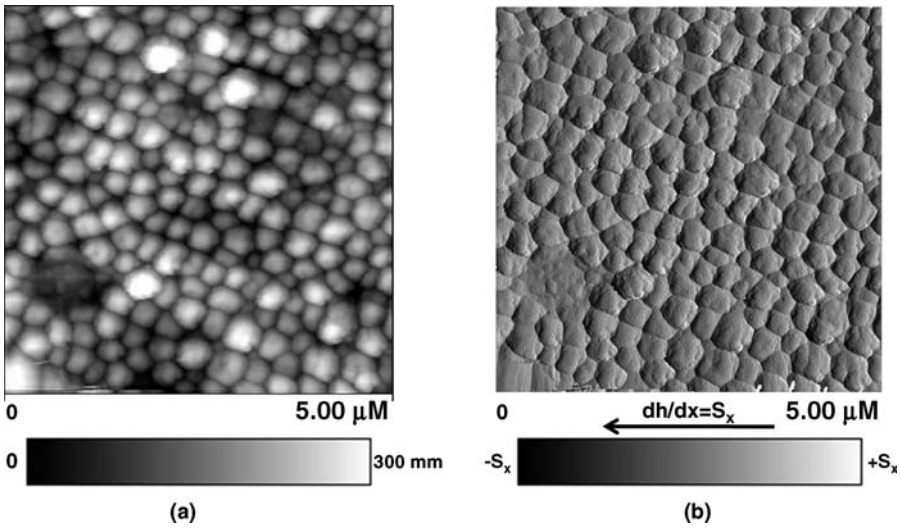


Figure 5.35 Atomic force microscopy image of (a) an imprinted brass surface, and (b) the gray-scale shading image deduced from it.

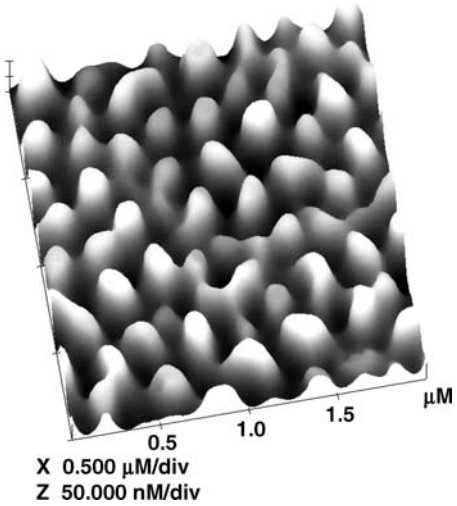


Figure 5.36 Atomic force microscopy image of a polycarbonate surface, imprinted by a 180 nm pore mask.

5.5.3 Imprinting of Polymers

Polymers can usually be imprinted at lower pressures compared to metals, and an image of a perfect transfer of the mask structure on PMMA was shown in Figure 5.31.

Polycarbonate (PC) can be imprinted using 180 nm pore masks at 200 MPa at room temperature [84, 85]; the result is shown in Figure 5.36, and heights of 80–90 nm and aspect ratios of about 0.5 result under these conditions.

Polyethylene (PE) gives corresponding results.

Polytetrafluoroethylene (PTFE) is a technically important, high-temperature-resistant thermoplastic with a low surface energy and friction coefficient. It is easily deformable, tends to creep, and is mechanically only scarcely loadable; hence, it is often used as anti-adhesive material. These properties become visible during imprinting experiments, and for imprinting alumina foils are better suited than either sheets or disks. An AFM image of a PTFE surface, imprinted by a 50 nm pore mask at 130 MPa and at room temperature, is shown in Figure 5.37. In this case, the superimposed roughness can be traced back to the material's properties.

Additional damages may also occur during separation of the mask from the substrate, and this is especially the case at high temperatures. Variations in the imprinting time, between 10 and 300 s, had no significant influence on the quality of the imprinted surfaces.

Polymer-coated metals are of major practical relevance for many reasons, but notably for hydrophobizing the surfaces (this effect may be improved by nanostructuring, and is discussed in Section 5.6). Metal sheets, when coated with two different commercially produced silicon polyesters, Silikoftal HTL 2 and Silikoftal NS 60 (Degussa), were used for imprinting with nanoporous alumina. As these polymers

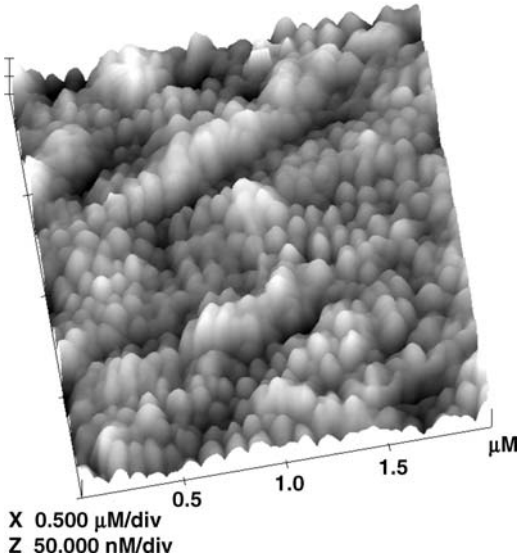


Figure 5.37 Atomic force microscopy image of a PTFE surface, imprinted by a 50 nm pore mask at 130 MPa, at room temperature.

become very hard and brittle if fully polymerized, imprinting is best carried out before polymerization. Although four different states were tested [84], imprinting of the softest coating (10 min at 100 °C) with about 200 nm pore masks (foils) produced good results for both polymers at a pressure of 130 MPa, with aspect ratios of 0.5 to 0.8. An increased hardness (10 min at 160 °C and 10 min at 250 °C) still allowed imprinting, but with decreasing aspect ratios. After the fourth polymerization step (2 min at 300 °C) the somewhat softer material NS 60 could still be structured, and with good results (see Figure 5.38).

5.5.4
Special Techniques to Imprint Hard Materials

The metals and polymers discussed above are each softer than the imprinting material, aluminum oxide. The problem of nanostructuring surfaces which are harder than alumina can be solved by treating via indirect routes. This can be demonstrated by a process which results in the production of a nanostructured silicon, which cannot be used directly for imprinting. The various steps in nanostructuring a silicon surface by using a 50 nm alumina membrane are shown schematically in Figure 5.39 [90]. In this case, aluminum discs of 30 mm diameter and 6 mm thickness (step A2) are fabricated at 40 V and 0 °C for 60 min, after preanodization. The Si surface to be nanostructured is coated with a 70 nm-thick PMMA film by spin-coating a PMMA solution in chlorobenzene (step A1). Subsequently, imprinting occurred at 10^7–10^8 Pa, at temperatures between 140 and 160 °C (step B), and a thin film of PMMA remained after imprinting (step C) which was removed by

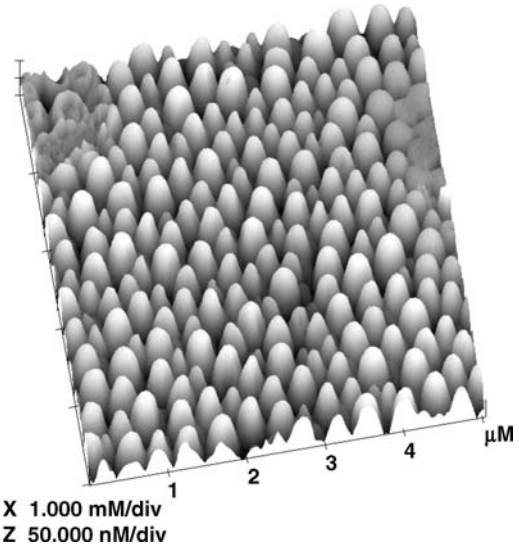


Figure 5.38 Imprinting result (AFM) of Silikoftal NS 60-coated metal sheet after hardening at 300 °C for 2 min. The aspect ratio was 0.3–0.6.

an anisotropic oxygen plasma (step D). Structuring of the Si surface was then followed by reactive ion etching (RIE) with CF_4/H_2 plasma (step E). AFM images of the imprinted PMMA surface before and after RIE are shown in Figure 5.40a and b, respectively.

As can be seen from Figure 5.40b, the plasma attacks not only the residual PMMA film but also to some extent the PMMA pillars, the original height of which (160 nm) was reduced to about 80 nm, although the base diameter was unchanged. The result of the final etching step can be seen in Figure 5.41, where the PMMA pillar structure was transferred into the Si surface, resulting in Si pillars of about 50 nm in height. The SEM image in Figure 5.41a shows a larger area of the structured surface, whereas an AFM image of the magnified cut-out (top right-hand corner) is shown in Figure 5.41b.

5.6 Properties of Nanoimprinted Polymer Surfaces

The altered properties of nanoimprinted surfaces can be illustrated by means of polymers.

5.6.1 Optical Properties

The light transmission of transparent polymers is of considerable practical relevance. The minimum reflection of light (i.e., maximum transparency) is a situation where polymer windows can be used, for example, to cover electronic devices in

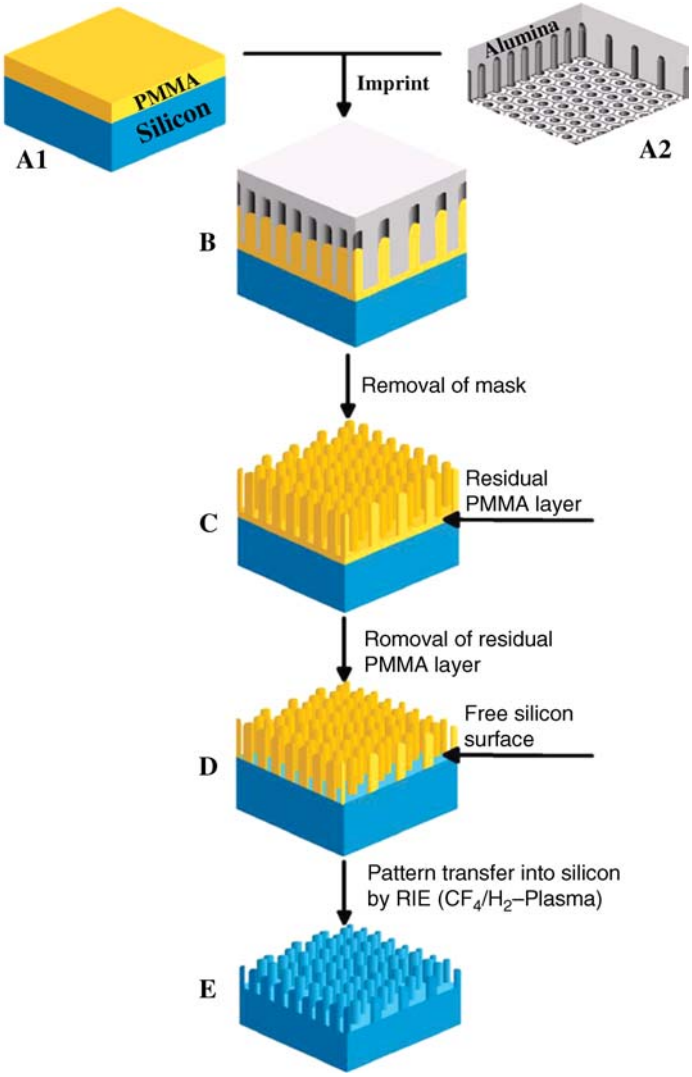


Figure 5.39 Schematic of the procedure to nanostructure a Si surface. Reproduced with permission from Ref. [95]; © 1996, Royal Chemical Society, London.

automobiles, mobile telephones, and solar cells. The elimination of light reflection has also been developed by nature; the so-called "moth-eye effect" causes the invisibility of night-active insects and, due to their high transparency, an optimized light efficiency for animals. The process is based on the nanostructured surface of the eyes where, due to the existence of building blocks that are smaller than the wavelength of visible light (200–300 nm) on the surface, the classical reflection of light when it is transmitted between air and a more dense medium is almost eliminated [91, 92]. Rather, the sharp transition between air and the material on a

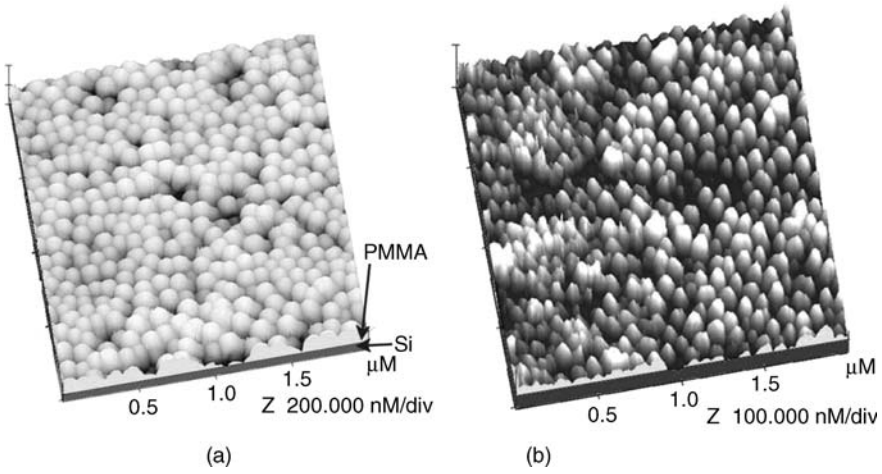


Figure 5.40 (a) Atomic force microscopy image of an imprinted PMMA film; (b) After reactive ion etching to remove the residual film on the Si surface. Reproduced with permission from Ref. [95]; © 1996, Royal Chemical Society, London.

smooth surface is substituted by a series of reflections, and this results in a gradient with a continuously increasing refractive index between the air and the surface.

PMMA, as a representative of the imprinted polymers, has been studied with respect to its transparency change, from a nonimprinted to an imprinted state [85]. In

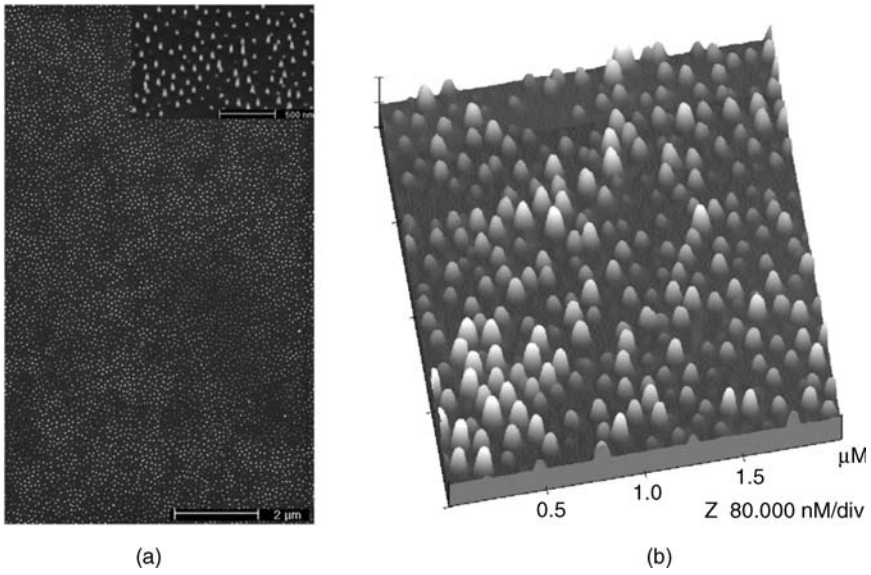


Figure 5.41 (a) Scanning electron microscopy and (b) and atomic force microscopy images of the final nanostructured Si surface. Reproduced with permission from Ref. [95]; © 1996, Royal Chemical Society, London.

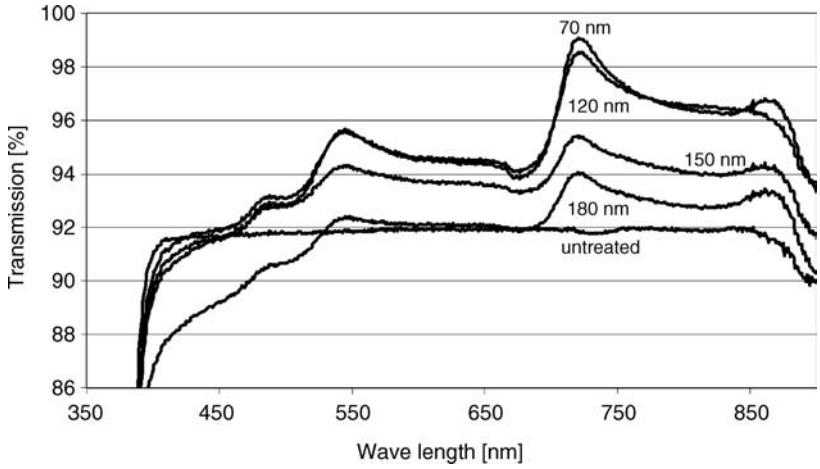


Figure 5.42 UV-visible transmission spectra of PMMA windows, imprinted with 180, 150, 120, and 70 nm pore masks. An untreated sample is shown for comparison.

a series of nanostructured probes with decreasing structure size, transmission was shown to increase in line with decreasing structure units. The UV-visible transmission spectra of PMMA windows imprinted with 180 to 70 nm pore alumina masks, are shown in Figure 5.42 [85]. Here, compared to a nonimprinted sample, there was a clear increase in transparency, especially at 520, 720, and 870 nm, all in the visible region. In the case of the 70 nm PMMA pillars, the transparency reached 99%. Whereas, the moth-eye effect functions best with structure units of about 300 nm, in this case the maximum value was observed with 70 nm pillars, which in turn suggests the need for a somewhat modified explanation, namely the formation of a surface layer with a low refractive index.

5.6.2 Wetting Properties

The wettability of a surface depends on the chemical nature of the material, and of its surface structure. Again, nature demonstrates this phenomenon in the so-called "Lotus effect", which keeps clean not only Lotus leaves but also numerous other biological surfaces, such as butterfly wings [93–95]. The effect is based on the increased hydrophobicity of a material with a low surface tension by a micro- and/or nanostructured surface. Whereas, metals exhibit high surface tensions, linked with a good wettability, glass and plastics show low surface tensions. The wettability of a surface is measured via the contact angle θ, which is the angle between a solid surface and the tangent, set on a liquid droplet. It can be calculated by using the Young equation [96]:

$$s_{sg} - \sigma_{sl} = \sigma_{lg} \cdot \cos(\theta) \tag{5.7}$$

where σ_{sl} is the interfacial tension solid/liquid, σ_{sg} is the interfacial tension solid/gas, σ_{lg} is the interfacial tension liquid/gas, and θ is the contact angle solid/liquid.

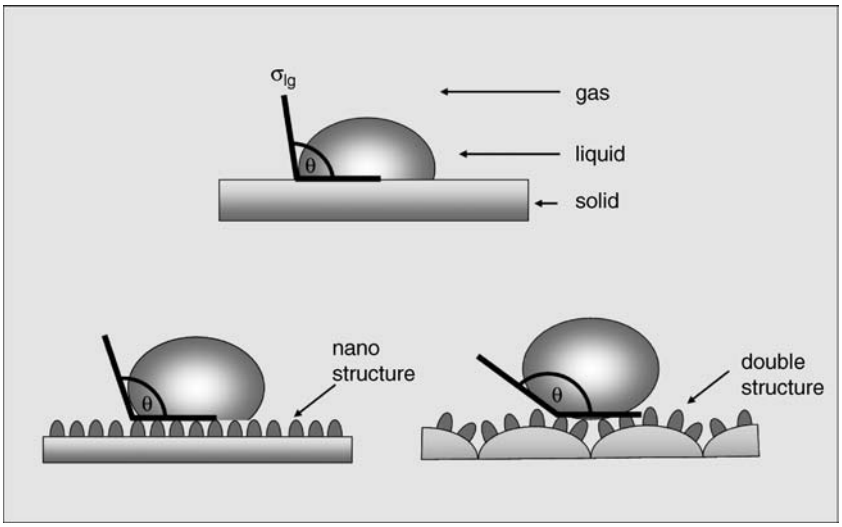


Figure 5.43 Increase of the contact angle (θ) by a micro/nanostructured surface.

Both, microstructured and/or nanostructured hydrophobic surfaces decrease wettability due to an additional reduction in the surface tension, and an increase of the contact angle. The differences in contact angles between a flat structure, and between nanostructured and a micro/nanostructured surfaces, are shown schematically in Figure 5.43. Here, the combination of micro- and nanostructured surfaces results in large contact angles, a situation which is especially realized in Lotus leaves. In this case, a contact angle of 0° corresponds to a perfect wettability, and an angle of 180° to a perfect hydrophobicity; however, neither value is reached in practice.

PTFE surfaces show increasing contact angles for water droplets with increasing structure size [85]. The results obtained with pore diameters ranging between 50 and 200 nm are listed in Table 5.5. Compared to nontreated surfaces, the contact angle was increased from 112° to 146°.

The reality can be shown via light microscopic images of water droplets on variously structured PTFE surfaces (Figure 5.44). Here, the contact surface corre-

Table 5.5 Measured contact angles (θ) depending on structure size on PTFE surfaces.

Pore diameters of the masks (nm)	Calculated structural distances (nm)	θ (°)
—	—	112 ± 3
50	106 ± 6	128 ± 5
75	159 ± 9	126 ± 3
120	212 ± 12	132 ± 2
120	265 ± 15	140 ± 5
180	398 ± 23	145 ± 2
200	451 ± 26	146 ± 3

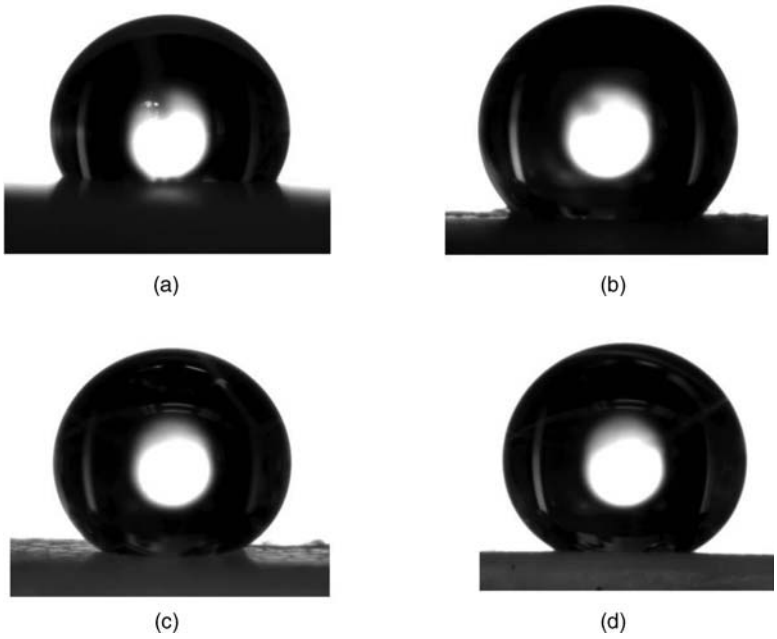


Figure 5.44 Light microscopy images of water droplets on different imprinted PTFE surfaces. (a) Nonimprinted; (b) 50 nm pores; (c) 120 nm pores; (d) 170 nm pores.

sponds with the real wetted area, which can be calculated from the measured contact angle [97]. The proportions of these areas vary from 50% (50 nm) to 25% (200 nm).

5.7 Surfaces with Nanoholes

To date, the use of nanoporous aluminum oxide membranes has been targeted exclusively towards the creation of nanoparticles, nanowires or nanotubes on surfaces. However, nanoporous alumina membranes can also be applied to fabricate surfaces with nanohole arrays. Of course, many other techniques have been developed to structure surfaces with nanoholes where, again, the use of (very thin) alumina membranes offers considerable advantages. This is because, in addition to a high pore regularity, it is also possible to create extended arrays. Diamond, semiconductor and metal surfaces can each be treated by using alumina membranes as masks for various etching processes. Nanohole-structured surfaces are of practical interest for similar reasons, such as nanowire- or nanodot-structured materials. Here again, the thickness of the membranes is a decisive factor for success since, if they are too thick, then shadowing effects of the pore walls will reduce their quality. As the membranes are also etched, too-thin membranes can partially be destroyed, with membranes of 300–700 nm thickness eventually providing the best results [98, 99].

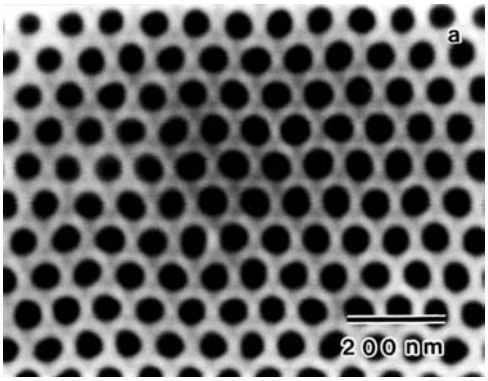


Figure 5.45 Scanning electron microscopy image of a diamond surface with ordered nanoholes, produced by polyethylene (PE) [104].

In any case, the etching rate of the membrane should be less than that of the corresponding surface.

The dominating etching technique is RIE, followed by plasma etching (PE) and fast atom beam etching (FAB), while ion milling (IM) plays a secondary role. Previously, RIE techniques have been applied for etching semiconductor surfaces such as Si [100], GaAs [98, 101, 102], GaN [101] and InP [99], whilst PE is more suited to the etching of Si [103], diamond [104–106], GaN [73, 107] and Al [102]. In the case of FAB, etching has been applied to Si [108], diamond [109], and GaAs [110].

Typically, a variety of etching gases can be used. For RIE processes, Cl_2 and Br_2 as well as BCl_3 and $SiCl_4$ are common, whereas for PE gases such as CF_4, $CBrF_3$, O_2 and also Cl_2 can be applied. In addition to O_2 and Cl_2, SF_6 has been used for FAB techniques.

Due to the slow simultaneous etching of the pore walls, the holes on the substrates will have a more or less expressed conical structure, depending on the etching time. As an example, a diamond surface with ordered holes generated by O_2 plasma etching through an alumina membrane is shown in Figure 5.45 [104].

Similar to surfaces decorated with wires, dots or tubes, surfaces with nanohole arrangements demonstrate interesting physical properties [69], although these will not be discussed at this point. Nonetheless, these include the antireflection of Si [103, 108], the photonic crystal behavior of GaN [110], and the increase of capacitance of porous diamond films by a factor of 400 [105, 106].

5.8 Summary and Outlook

Aluminum plays a special role among the so-called "gate metals," as it forms the most qualified nanopores during anodization, with variations in pore diameter, ranging from about 10 nm to several hundred manometers, easily being arranged. Furthermore, aluminum foils, which frequently are used as starting materials, are cheap and

easily available. Besides routine anodization, special techniques have been developed to generate highly ordered pore arrays, and the mechanism of such formation is well understood. Previously, many applications have been developed, and many are foreseen for the near future. Two of these (as discussed in this chapter) include: (i) pore filling with very different materials, which results in wires, tubes or particles of variable dimensions; and (ii) the use of nanoporous alumina surfaces for imprinting numerous surfaces of metals and polymers. In both cases, nanostructured surfaces are formed, leading to the creation of materials with novel physical and chemical properties. Yet, such early promise must be developed further to permit imprinting processes for extended surfaces, perhaps in the meter range. Pore walls, chemically modified by catalytically active metals, have been tested in gas-phase catalysis [111], and the first attempts to apply a pore-size specific release of chemical compounds (i.e., drugs) have initiated novel developments in slow but continuous rates of drug delivery [112]. Yet, numerous other future applications can be foreseen, including optics, electronics and magnetism, to name but a few.

Nanoporous alumina membranes are also used as matrices for the fabrication of surfaces with nanoholes, including various etching techniques. Such concave, nanostructured surfaces offer interesting properties compared to nontreated materials.

Taken together, nanoporous alumina membranes represent a unique family of materials which, despite having experienced impressive progress during the past two decades, will surely continue to play major roles in many further scientific and practical developments.

References

1 Lohrengel, M.M. (1993) *Mater. Sci. Eng.*, **6**, 241.
2 Thompson, G.E. and Wood, G.C. (1983) *Treatise on Materials Science and Technology*, **23**, 205.
3 Diggle, J.W., Downie, T.C., and Goulding, C.W. (1969) *Chem. Rev.*, **69**, 365.
4 Parkhutik, V.P. and Shershulskii, V.I. (1986) *J. Physics D: Appl. Phys.*, **19**, 623.
5 Valand, T. and Heusler, K.E. (1983) *J. Electroanal. Chem.*, **149**, 71.
6 Hurlen, T., Lian, H., and Ödergard, O.S. (1984) *Electrochim. Acta*, **29**, 579.
7 Keller, F., Hunter, M.S., and Robinson, D.L. (1953) *J. Electrochem. Soc.*, **100/9**, 411.
8 Macdonald, D.D. (1993) *J. Electrochem. Soc.*, **140/3**, L27.
9 Siejika, J. and Ortega, C. (1977) *J. Electrochem. Soc.*, **124/6**, 883.
10 Thompson, G.E., Furneauax, R.C., and Wood, G.C. (1978) *Nature*, **272**, 433.
11 O'Sullivan, J.P. and Wood, G.C. (1970) *Proc. R. Soc. London*, **317**, 511.
12 Hornyak, G.L., Dutta, J., Tibbals, H.F., and Rao, A.K. (eds) (2008) *Introduction to Nanoscience*. CRC Press.
13 Sawitowski, T. (1999) Neue Nanokomposite. Goldcluster, Goldkolloide und Silizium in Aluminiumoxidmembranen – Struktur und Eigenschaften. PhD thesis, University of Essen, Germany.
14 Asoh, H., Nishio, K., Nakao, M., Tamamura, T., and Masuda, H. (2001) *J. Electrochem. Soc.*, **148**, B152.
15 Masuda, H., Yamada, H., Satoh, M., and Asoh, H. (1997) *Appl. Phys. Lett.*, **71**, 2770.
16 Matsui, Y., Nishio, K., and Masuda, H. (2006) *Small*, **2**, 522.

17 Randon, J., Mardilovich, P.P., Govyadinov, A.N., and Paterson, R. (1995) *J. Colloid Interface Sci.*, **169**, 335.
18 Wefers, K. and Misra, C. (1987) Company Report. Alcoa Laboratories.
19 Schmid, G., Bäumle, M., Heim, I., Kröll, M., Müller, F., and Sawitowski, T. (1999) *J. Cluster Sci.*, **10**, 223.
20 Furneaux, R.C., Rigby, W.R., and Davidson, A.P. (1989) *Nature*, **337**, 147.
21 Binnig, G., Gerber, Ch., Stoll, E., Albrecht, T.R., and Quate, C.F. (1987) *Europhys. Lett.*, **3/12**, 1281.
22 Rugar, D. and Hansma, P. (1990) *Physics Today*, **10**, 23.
23 Lu, Q., Gao, F., Komarneni, S., and Mallouk, T.E. (2004) *J. Am. Chem. Soc.*, **126**, 8650.
24 Che, G.L., Lakshmi, B.B., Fischer, E.R., and Martin, C.R. (1998) *Nature*, **393**, 346.
25 Li, J., Papadopoulos, C., Xu, J.M., and Moskovits, M. (1999) *Appl. Phys. Lett.*, **75**, 367.
26 Li, C., Papadopoulos, J., and Xu, J.M. (1999) *Nature*, **402**, 253.
27 Masuda, H., Yanagishita, T., Yasui, K., Nishio, K., Yagi, I., Rao, N., and Fujishima, A. (2001) *Adv. Mater.*, **13**, 247.
28 Routkevitch, D., Bigioni, T., Moskovits, M., and Xu, J.M. (1996) *J. Phys. Chem.*, **100**, 14037.
29 Shemilov, K.B. and Moskovits, M. (2000) *Chem. Mater.*, **12**, 250.
30 Lei, Y., Zhang, L.D., and Fan, J.C. (2001) *Chem. Phys. Lett.*, **338**, 231.
31 Nielsch, K.M., Müller, F., Li, A.P., and Gösele, U. (2000) *Adv. Mater.*, **12**, 582.
32 Cao, H.Q., Xu, Z., Sang, H., Sheng, D., and Tie, C.Y. (2001) *Adv. Mater.*, **13**, 121.
33 Mu, C., Yn, X.Y., Wang, R.M., Wu, K., Xu, D.S., and Guo, G.L. (2004) *Adv. Mater.*, **16**, 1550.
34 Schneider, J.J., Popp, A., and Engstler, J. (2008) Fundamentals and Functionality of Inorganic Wires, Rods and Tubes, in *Nanotechnology. Principles and Fundamentals*, vol. 1 (ed. G. Schmid), Wiley-VCH, Weinheim, pp. 97–138.
35 Masuda, H., Yada, K., and Osaka, A. (1998) *Jpn. J. Appl. Phys.*, **37**, L1340.
36 Okano, K., Yamada, T., Ishihara, H., Koizumi, S., and Itoh, J. (1997) *Appl. Phys. Lett.*, **70**, 2201.
37 Shiomi, H. (1997) *Jpn. J. Phys. Lett.*, **36**, 7745.
38 Kornienko, O., Reilly, P.T.A., Whitten, W.B., and Ramsey, J.M. (2000) *Anal. Chem.*, **72**, 559.
39 Geis, M., Efremow, N.N., Krohn, K.E., Twichell, J.C., Lyszcarz, T.M., Kalish, R., Greer, J.A., and Tabat, M.D. (1998) *Nature*, **393**, 431.
40 Grill, A. (1999) *Diamond Relat. Mater.*, **8**, 428.
41 Bao, J., Tie, C., Xu, Z., Zhou, Q., Shen, D., and Ma, Q. (2001) *Adv. Mater.*, **13**, 1631.
42 Gao, H., Wang, F., Xu, D.S., Wu, K., Xie, Y.C., Liu, S., Wang, E.G., Xu, J., and Yu, D.P. (2003) *J. Appl. Phys.*, **93**, 5602.
43 Masuda, H., Yotsuy, M., and Ishida, M. (1998) *Jpn. J. Appl. Phys.*, **37**, L1090.
44 Jessensky, O. (1997) Untersuchungen zum Porenwachstum in 6H-Siliziumkarbid und anodischem Aluminiumoxid. PhD thesis, Martin-Luther-University of Halle, Germany.
45 Al Mawlavi, D., Coombs, N., and Moskovits, M. (1991) *J. Appl. Phys.*, **70**, 4421.
46 Li, F., Metzger, M., and Doyle, W.D. (1997) *IEEE Trans. Magn.*, **33**, 3715.
47 Routkevitch, D., Tager, A.A., Haruyama, J., Almawlawi, D., Moskovits, M., and Xu, J.M. (1996) *IEEE Trans. Electron Devices*, **147**, 1646.
48 Sautter, W., Ibe, G., and Meier, J. (1974) *Aluminium*, **50**, 143.
49 Li, A.P., Müller, F., Birner, A., Nielsch, K., and Gösele, U. (1999) *Adv. Mater.*, **11**, 483.
50 Dobrev, D., Vetter, J., Angert, N., and Neumann, R. (1999) *Appl. Phys. A*, **69**, 233.
51 Dobrev, D., Vetter, J., Angert, N., and Neumann, R. (2001) *Appl. Phys. A*, **72**, 729.
52 Choi, K.H., Kim, H.S., and Lee, T.H. (1998) *J. Power Sources*, **75**, 230.
53 Sun, M., Zangari, G., Shamsuzzoha, M., and Metzger, F.R.M. (2001) *Appl. Phys. Lett.*, **78**, 2964.
54 Fan, H.J., Werner, P., and Zacharias, M. (2006) *Small*, **2**, 700.
55 Shingubara, S., Okino, O., Sayama, Y., Sakaue, H., and Takahagi, T. (1999) *Solid State Electron.*, **43**, 1143.

56 Taberna, P.-L., Mitra, S., Poizot, P., Simon, P., and Tarascon, J.-M. (2006) *Nat. Mater.*, **5**, 567.

57 Zhang, Y., Li, G., Wu, Y., Zhang, B., Song, W., and Zhang, L. (2002) *Adv. Mater.*, **14**, 1227.

58 Bogart, T.E., Dey, S., Lew, K.-K., Mohney, S.E., and Redwing, J.M. (2005) *Adv. Mater.*, **17**, 114.

59 Wagner, R.S. and Ellis, W.C. (1964) *Appl. Phys. Lett.*, **4**, 89.

60 Wagner, R.S., Ellis, W.C., Jackson, K.A., and Arnold, S.M. (1964) *J. Appl. Phys.*, **35**, 2993.

61 Cui, Y., Lauhorn, L.J., Gudiksen, M.S., Wang, J., and Lieber, C.M. (2001) *Appl. Phys. Lett.*, **78**, 2214.

62 Westwater, J., Gosain, D.P., and Usui, S. (1997) *Jpn. J. Appl. Phys. Part 1*, **36**, 6204.

63 Westwater, J., Gosain, D.P., Tomiya, S., Usui, S., and Ruda, H. (1997) *J. Vac. Sci. Technol. B*, **15**, 554.

64 Westwater, J., Gosain, D.P., and Usui, S. (1998) *Phys. Status Solidi A*, **165**, 37.

65 Lei, Y., Zhang F L.D., Meng, G.W., Li, G.H., Zhang, X.Y., Liang, C.H., Chen W., and Wange S.X. (2001) *Appl. Phys. Lett.*, **78**, 1125.

66 Baranski, A.S. and Fawcett, W.B. (1980) *J. Electrochem. Soc.*, **127**, 766.

67 Klein, J.D., Herrich, R.D. II, Palmer, D., Sailor, M.J., Brumlik, C.J., and Martin, C.R. (1993) *Chem. Mater.*, **5**, 902.

68 Charkarvarti, S.K. and Vetter, J. (1993) *J. Micromech. Microeng.*, **3**, 57.

69 Lei, Y., Cai, W., and Wilde, G. (2007) *Prog. Mater. Sci.*, **52**, 465.

70 Masuda, H. and Satoh, H. (1996) *Jpn. J. Appl. Phys.*, **35**, L126.

71 Masuda, H., Yasui, K., Sakamoto, Y., Nakao, M., Tamamura, T., and Nishio, K. (2001) *Jpn. J. Appl. Phys.*, **40**, L1267.

72 Choi, J., Sauer, g., Goering, P., Nielsch, K., Wehrspohn, R.B., and Gösele, U. (2003) *J. Mater. Chem.*, **13**, 1100.

73 Wang, Y.D., Chua, S.J., Sander, M.S., Chem, P., Tripathy, S., and Fonstad, C.G. (2004) *Appl. Phys. Lett.*, **85**, 816.

74 Sander, M.S. and Tan, L.S. (2003) *Adv. Funct. Mater.*, **13**, 393.

75 Rabin, O., Herz, P.R., and Lin, Y.M. (2003) *Adv. Funct. Mater.*, **13**, 631.

76 Tian, M.L., Yu, S.Y., Wang, J.G., Kumar, N., Wertz, E., and Li, Q. (2005) *Nano Lett.*, **5**, 697.

77 Choi, J., Sauer, g., Nielsch, K., Wehrspohn, R.B., and Gösele, U. (2003) *Chem. Mater.*, **15**, 776.

78 Xia, Y., Rogers, J.A., Paul, K.E., and Whitesides, G.M. (1999) *Chem. Rev.*, **99**, 1823.

79 Xia, Y. and Whitesides, G.M. (1998) *Angew. Chem., Int. Ed.*, **37**, 550.

80 Wallraff, M. and Hinsberg, W.D. (1999) *Chem. Rev.*, **99**, 1801.

81 Piner, R.D., Zhu, J., Xu, F., Hong, S., and Mirkin, C.A. (1999) *Science*, **183**, 661.

82 Emmelius, M., Pawlowski, G., and Vollmann, H.W. (1989) *Angew. Chem., Int. Ed.*, **28**, 1445.

83 Chou, S.Y., Krauss, P.R., and Renstrom, P.J. (1995) *Appl. Phys. Lett.*, **67**, 3114.

84 Levering, M. (2003) Strukturierung von Oberflächen mit nanoporösem Aluminiumoxid. PhD thesis, University of Essen, Germany.

85 Schmid, G., Levering, M., and Sawitowsky, T. (2007) *Z. Anorg. Allg. Chem.*, **633**, 2147.

86 Beyer, B. (1971) *Werkstoffkunde NE-Metalle*, VEB Deutscher Verlag, Grundstoffindustrie, Leipzig, Germany, p. 340.

87 Goodfellow (2002) *Datenblätter*, Goodfellow, Bad Nauheim, Germany.

88 Merkel, M. and Thomas, K.H. (1998) *VEB Fachbuchverlag Leipzig*, Germany, p. 340.

89 Ross, R.B. (1980) *Metallic Materials Handbook*, 3rd edn, E. & F. N. Spon Ltd.

90 Kruse, M., Frankza, S., and Schmid, G. (2003) *Chem. Commun.*, 1333.

91 Macleod, B. and Sonetz, G. (1999) *Laser Focus World*, **8**, 1–55.

92 Aydin, C., Zaslavsky, A., Sonek, G.J., and Goldstein, J. (2002) *Appl. Phys. Lett.*, **80**, 2242.

93 Wagner, T., Neinhuis, C., and Barthlott, W. (1996) *Acta Zool.*, **77**, 213.

94 Wolter, M., Barthlott, W., Knoch, M., and Noga, G.J. (1988) *Angew. Bot.*, **62**, 53.

95 Barthlott, W., Neinhuis, C., Jetter, R., Bourauel, T., and Riederer, M. (1996) *Flora*, **191**, 169.

96 Israelachvili, I. (1995) *Intermolecular & Surfaces Forces*, Academic Press, London.

97 Chen, W., Fadeev, A.Y., Hsieh, M.C., Öner, D., Youngblood, J., and McCarthy, T. (1999) *Langmuir*, **15**, 3395.

98 Cheng, G.S. and Moskovits, M. (2002) *Adv. Mater.*, **14**, 1567.

99 Nakao, M., Oku, S., Tamamura, T., Yasui, K., and Masuda, H. (1999) *Jpn. J. Appl. Phys.*, **38**, 1052.

100 Crouse, D., Lo, Y.H., Miller, A.E., and Crouse, M. (2000) *Appl. Phys. Lett.*, **76**, 49.

101 Liang, J.Y., Chik, H., Yin, A.J., and Xu, J. (2002) *J. Appl. Phys.*, **91**, 2544.

102 Menon, L., Ram, K.B., Patibandla, S., Aurongzeb, D., Holtz, M., and Yun, J. (2004) *J. Electrochem. Soc.*, **151**, C492.

103 Tian, L., Ram, K.B., Ahmad, I., Menon, L., and Holtz, M. (2005) *J. Appl. Phys.*, **97**, 026101.

104 Masuda, H., Watanabe, M., Yasui, K., Tryk, D.A., Rao, T.N., and Fujishima, A. (2000) *Adv. Mater.*, **12**, 444.

105 Honda, K., Rao, T.N., Tryk, D.A., Fujishima, A., Watanabe, M., and Yasui, K. (2000) *J. Electrochem. Soc.*, **147**, 659.

106 Honda, K., Rao, T.N., Tryk, D.A., Fujishima, A., Watanabe, M., and Yasui, K. (2001) *J. Electrochem. Soc.*, **148**, A668.

107 Wang, J.D., Chua, S.J., Tripathy, S., Sander, M.S., Chen, P., and Fonstad, C.G. (2005) *Appl. Phys. Lett.*, **86**, 071917.

108 Kanamori, Y., Hane, K., Sai, H., and Yugami, H. (2001) *Appl. Phys. Lett.*, **78**, 142.

109 Ono, T., Konoma, C., Miyashita, H., Kananmori, Y., and Esashi, M. (2003) *Jpn. J. Appl. Phys.*, **42**, 3867.

110 Nakao, M., Oku, S., Tanaka, H., Shibata, Y., Yokoo, A., and Tamamura, T. (2002) *Opt. Quantum Electron.*, **34**, 183.

111 Kormann, H.-P., Schmid, G., Pelzer, K., Philippot, K., and Chaudret, B. (2004) *Z. Anorg. Allg. Chem.*, **630**, 1913.

112 Kipke, S. and Schmid, G. (2004) *Adv. Funct. Mater.*, **14**, 1184.

6
Colloidal Lithography

Gang Zhang and Dayang Wang

6.1
Introduction

The advent of nanoscience and nanotechnology has, in recent years, led to a tremendous enthusiasm among the research groups of different scientific disciplines such as physics, chemistry, and biology, their common aim being to utilize nanostructures with the intent of pursuing the innovative properties derived from nanometer dimensions. In this context, the fabrication of nanostructures has today become an increasing demand. Clearly, low-throughput and expensive maskless lithography represents a less-accessible choice for chemists, physicists, material scientists, and biologists. The successful extension of mask-assisted lithography beyond microelectronics workshops have been largely limited by problems of mask design and preparation. Recently, much effort has been expended towards the development of nonconventional lithographic techniques, especially those that are integrated with a bottom-up nanochemical procedure for surface patterning with a low-cost, flexible processing capability, and a high throughput. Most of the nonconventional lithographic techniques developed to date, however, require the assistance of conventional lithographic techniques, such as photolithography, to design and make the masks or masters. Hence, in attempting to address this challenge, an increasing amount of attention has been paid to the use of self-assemblies of molecules and colloidal particles for the development of ingenious, cheap, and nonlithographic methods of masking.

Monodisperse colloidal particles with sizes that range from tens of nanometers to tens of micrometers, can be easily synthesized via wet chemistry approaches such as emulsion polymerization and sol–gel synthesis. Due to the size and shape monodispersity, these particles can self-assemble into both two-dimensional (2-D) and three-dimensional (3-D) extended periodic arrays, which usually are referred to as "colloidal crystals." The latter are usually characterized by a brilliant iridescence arising from the Bragg reflection of light by their periodic structures. Despite such beauty, the iridescent color has recently inspired the explosive study of fabrication of 3-D colloidal crystals or inverse opals – that is, a 3-D inverted replication of the

Nanotechnology, Volume 8: Nanostructured Surfaces. Edited by Lifeng Chi

ISBN: 978-3-527-31739-4

crystals for pursuing a complete energy bandgap to manipulate electromagnetic waves, similar to the situation with electrons in semiconductors. Before being used as photonic materials, both the ordered arrays of solid particles and those of the interstices between the particles of colloidal crystals, have already been used as masks or templates for surface patterning, for example via etching or the deposition of materials. This bottom-up masking methodology has recently attracted more attention for surface patterning due to the processing simplicity, the low cost, the flexibility of extending on various substrates with different surface chemistries (and even curvatures), and the ease of scaling down the feature size to below 100 nm. In this chapter, the various surface-patterning processes based on the use of colloidal crystals as masks – referred to hereafter as colloidal lithography (CL) – will be discussed, the processing principles reviewed, and recent advances in the area surveyed.

6.2
Colloidal Crystallization: The Bottom-Up Growth of Colloidal Masks

The success of using colloidal crystals as masks for surface patterning is determined by the ability to directing the self-assembly of colloidal particles and to manipulate the crystal packing structures. Provided that their size and shape are monodisperse, colloidal particles can be readily self-assembled into long-range ordered arrays with a hexagonal packing, driven simply by entropic depletion and gravity. Subsequent evaporation of the solvent leads to thermodynamically and mechanically stable face-centered cubic (*fcc*) or hexagonally close-packed (*hcp*) extended crystals. Until now, a range of colloidal crystallization techniques – with and without the aid of templates – has been successfully developed to implement colloidal crystallization in a controlled fashion [1–3]. However, due to the vast number of reports made on colloidal crystallization, the immense diversity of the crystallization techniques described, and taking into account the fact that colloidal lithography relies on the masking of single or double layers of colloidal crystals, attention in this section will be centered mainly on the currently available techniques for 2-D colloidal crystallization.

6.2.1
Sedimentation

Sedimentation represents a natural pathway for colloidal crystallization since, when dispersed in a liquid, colloidal particles tend to settle out of the fluid under gravity and to accumulate and precipitate on a wall – a process which can be described by Stokes' law. This sedimentation process can be used to grow colloidal crystals of high quality, while the crystal thickness can be fine-tuned by adjusting the particle concentration. Unfortunately, however, as the sedimentation time may be up to several hundreds of hours, time-consumption represents the major drawback of this technique [4]. It is possible to accelerate the rate of sedimentation by applying centrifugal force, but this is undertaken at the cost of reducing the quality of the colloidal crystals obtained. Neither does sedimentation carried out under centrifugal forces allow the formation

of 2-D colloidal crystals. An additional drawback of sedimentation in colloidal crystallization results from the intermediate stage during solvent evaporation, at which the colloidal particles are not in close contact but rather are interspaced by water necks. If a complete evaporation of the solvent is then carried out, this will cause cracks to form that are difficult not only to prevent but also to manage [5, 6].

During the early 1990s, Nagayama's group began a systematic study of the sedimentation of colloidal particles in the presence of strong attractive capillary forces [7]. By using optical microscopy and a Teflon ring to confine the dispersions of colloidal particles, the particle sedimentation dynamics on a solid substrate could be directly observed. The observations made by Nagayama and colleagues suggested the existence of a two-stage mechanism for 2-D colloidal crystallization:

- *Nucleation*, which led to the particles becoming trapped on the substrate due to attractive capillary forces between the particles and the surrounding solvents. This occurred especially when the solvent layer thickness was comparable to the diameter of the particles during solvent evaporation.
- *Crystal growth*, whereby convective flux caused the particles to be moved to the existing ordered domains, as a result of water evaporation from the meniscus between the particles (Figure 6.1) [7].

Subsequently, Micheletto and coworkers fabricated 2-D colloidal crystals on a solid substrate through sedimentation, by tilting the substrate through about 9° and maintaining a constant system temperature with a Peltier cell [8]. The procedure used by Micheletto *et al.* allowed the growth of 2-D colloidal crystals that were composed of particles less than 100 nm in size, which was not possible via the Nagayama protocol. However, when the 2-D colloidal crystals obtained via

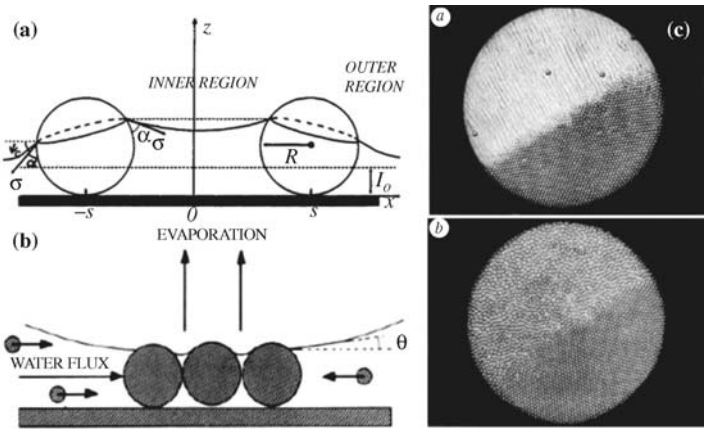


Figure 6.1 (a) Two spheres partially immersed in a liquid layer on a horizontal solid substrate. The deformation of the liquid meniscus gives rise to interparticle attraction; (b) Convective flux toward the ordered phase due to the water evaporation from the menisci between the particles in the 2-D array; (c) Photographs of 2-D crystal growth. Reproduced with permission from Ref. [7].

Micheletto's procedure were compared to those obtained via Nagayama's method, they proved to be of a much poorer quality, especially in terms of their surface coverage and degree of long-range ordering.

6.2.2
Vertical Deposition

When a supporting substrate is held vertically in a suspension of colloidal particles, moving the front of the suspension flow – either by evaporating the solvent or by withdrawing the substrate from the suspension – can cause the colloidal particles to be pinned onto the substrates (the process of nucleation) and also a convective transfer of the particles from the bulk phase to the drying front (the process of crystallization) (Figure 6.2) [9]. The thickness of the colloidal crystals obtained by vertical deposition depends on the ratio between the thickness of the liquid films that remain to support the substrates and the diameter of the colloidal particles [9]. When this ratio is much greater than 1, 3-D colloidal crystals are obtained of high quality, and the crystal thickness can be fine-tuned by adjusting the particle

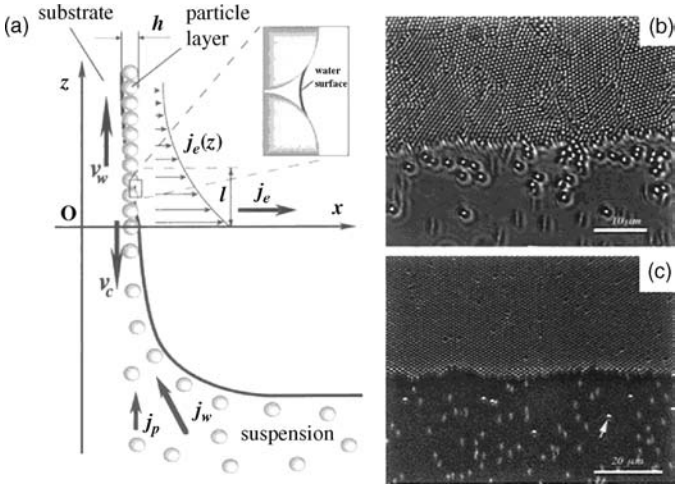


Figure 6.2 (a) Sketch of the particle and water fluxes in the vicinity of monolayer particle arrays growing on a substrate plate that is being withdrawn from a suspension. The inset shows the menisci shape between neighboring particles. Here, v_w is the substrate withdrawal rate, v_c is the array growth rate, j_w is the water influx, j_p is the respective particle influx, j_e is the water evaporation flux, and h is the thickness of the array; (b, c) A part of the leading edge of a growing monolayer particle array. The upper-halves of the photographs show the formations of (b) differently oriented small domains of ordered 814 nm particles, and (c) a single domain of ordered 953 nm particles. The lower halves show particles being dragged by the water flow towards the forming monolayer. Because of the high velocity on a microscale ($v_p = 100\ m\ s^{-1}$), the particles are seen as short, fuzzy lines. The particles (one is indicated by an arrow in panel b) seen as bright spots have a large diameter (compared to average values) and are wedged into the wetting film. Reproduced with permission from Ref. [9].

concentration [10]. However, when the ratio is similar to or less than 1, 2-D colloidal crystals are obtained [9]. Vertical deposition may also allow the formation of crack-free colloidal crystals, provided that the suspensions of colloidal particles wet support the substrates well, that there is no interaction between the particles and the substrates, the suspensions are sufficiently stable, and the solvent evaporation is well controlled [9].

Dip-coating is a rapid and dip-coater assisted variant of vertical deposition [11]. A number of techniques have also been developed to improve the efficiency and quality of colloidal crystallization via vertical deposition, such as variable-flow deposition [12], isothermal heating evaporation-induced self-assembly [13], two-substrate deposition [14], reduction of the humidity fluctuation [15], adjustment of the meniscus shape [16], temperature-induced convective flow [17] and vertical deposition with a tilted angle [18]. The maximal size of the colloidal particles used for vertical deposition is limited by the sedimentation of the colloidal particles; these sizes are typically 400–500 nm for silica particles and 1 μm for polystyrene particles. In aiming to compete with sedimentation, Kitaev and Ozin used a low pressure to accelerate the solvent evaporation, and successfully grew large-area 2-D binary colloidal crystals, where the diameter ratio of the large to small particles was in the range of 0.175 to 0.225 (Figure 6.3) [19].

Vertical deposition has recently been extended to the stepwise growth of 2-D colloidal crystals with large and small colloidal particles on a substrate [20, 21]. In this procedure, the 2-D colloidal crystals of the large particles first formed on the substrate are used as a template to grow 2-D colloidal crystals of the small particles. Then, by deliberately tuning the concentration of the small particle suspension, it is possible to construct binary colloidal crystals with stoichiometric ratios of large to small particle sizes of 1 : 2, 1 : 3, 1 : 4, or 1 : 5 [20, 21].

6.2.3
Spin Coating

Spin coating was the first technique used to grow 2-D colloidal crystal masks for colloidal lithography, due to the fact that it allows easy and quick crystal formation over large areas [22]. The long-range ordering degree of 2-D colloidal crystals obtained by spin coating can be improved by increasing the wetting of the suspensions of colloidal particles on the supporting substrates, for example by adding ethylene glycol to the suspensions [23]. Unfortunately, the spin coating process is far more complicated than it first appears, and the underlying mechanism remains in debate. When Rehg and Higgins conducted a theoretical analysis of the physics governing the spin coating of a colloidal particle suspension on a planar substrate, they proposed that:

- The functional relationship between the suspension viscosity and the particle concentration plays a much more significant role during the spin coating of a colloidal particle suspension, especially in the case of a non-hard sphere suspension, than in the spin coating of polymer solutions.

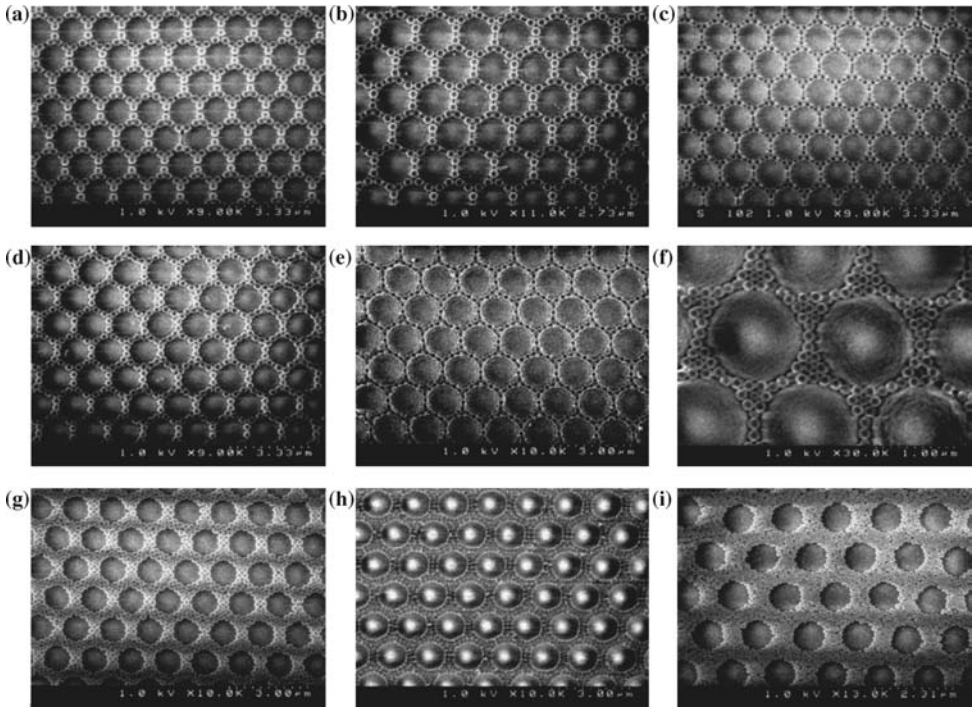


Figure 6.3 Library of surface micropatterns produced by accelerated evaporation coassembly of binary dispersions of monodisperse microspheres with a large size ratio and imaged using field-emission scanning electron microscopy. The larger spheres of all binary dispersions were PS latex of size, $d_L = 1.28\,\mu m$, while their volume fraction (ϕ_L) and the volume fraction (ϕ_S) and size (d_S) of smaller spheres were as follows: (a) $\phi_L = 0.017$, $\phi_S = 3.4 \times 10^{-4}$, $d_S = 290$ nm silica; (b) $\phi_L = 0.014$, $\phi_S = 2.5 \times 10^{-4}$, $d_S = 260$ nm silica; (c) $\phi_L = 0.014$, $\phi_S = 2.1 \times 10^{-4}$, $d_S = 225$ nm silica; (d) $\phi_L = 0.017$, $\phi_S = 3.8 \times 10^{-4}$, $d_S = 260$ nm silica; (e) $\phi_L = 0.014$, $\phi_S = 2.7 \times 10^{-4}$, $d_S = 225$ nm PS latex; (f) $\phi_L = 0.017$, $\phi_S = 3.0 \times 10^{-4}$, $d_S = 145$ nm silica; (g) $\phi_L = 0.017$, $\phi_S = 4.3 \times 10^{-4}$, $d_S = 205$ nm silica; (h) $\phi_L = 0.017$, $\phi_S = 4.1 \times 10^{-4}$, $d_S = 145$ nm silica; (i) $\phi_L = 0.017$, $\phi_S = 5.6 \times 10^{-4}$, $d_S = 145$ nm silica. Reproduced with permission from Ref. [19].

- The time scale associated with the spin coating of colloidal particle suspensions is rather different from that associated with the spin coating of polymer solutions.
- The inter-particle interaction should be taken into account to elucidate the packing ordering of the particles, the porosity of the particle films, and the functional relationship between the coated film thickness and the substrate angular velocity, although this is difficult to model.
- In order to minimize the secondary Marangoni instability for striation-free and uniform films of colloidal particles, a rapid substrate acceleration, high spinning speed, and reduced evaporation speed are needed [24].

Jiang and Mcfarland were successful in fabricating wafer-scale long-range ordered and non-close-packed 2-D and 3-D colloidal crystals by the spin coating of a highly viscous triacrylate suspension of silica particles and the subsequent polymerization of triacrylate, followed by a partial removal of the polymer matrices [25, 26]. Wang and Möhwald have developed a stepwise spin-coating protocol to consecutively deposit large and small colloidal particles in binary colloidal crystals, in which the interstitial arrays in the 2-D colloidal crystal of the large particles are used to template the deposition of small particles, due to their spatial and depletion entrapment (Figure 6.4) [27].

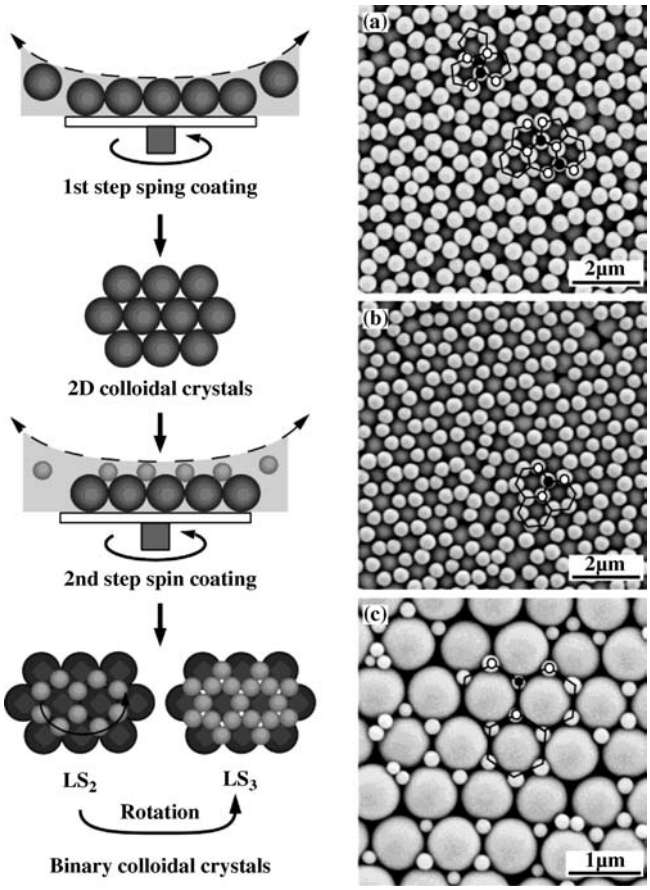


Figure 6.4 Left column: Schematic diagram of the procedure used to fabricate binary colloidal crystals by stepwise spin coating. Right column: Scanning electron microscopy images of the binary colloidal crystals produced by stepwise spin coating at a spin speed of 3000 rpm, in which 519 nm (a) , 442 nm (b) , and 222 nm silica spheres (c) were confined within the interstices between hexagonal close-packed 891 nm silica spheres. The closed or open circles mark locations of small spheres, while the polygon frames highlight their patterns. Reproduced with permission from Ref. [27].

6.2.4
Colloidal Crystallization at the Water/Air Interface

During recent years, extensive investigations have been conducted into the use of a water/air interface as a platform for molecular self-assembly. In particular, the Langmuir–Blodgett (LB) technique has proved to be a powerful and versatile method of organizing amphiphilic molecules at macroscopic monolayer films at the water/air interface, and to transfer these films to solid substrates in a controlled manner [28]. It has also been shown, but recognized to a lesser degree, that in a biphasic system such as water/oil, colloidal particles behave in rather similar fashion to amphiphilic molecules, in that from a thermodynamic standpoint they prefer to attach to the interface [29]. Based on this analogy, the water/air interface has been extended to support the self-assembly of colloidal particles. For example, when Pieranski conducted the first deliberate microscopic observation of 2-D colloidal crystallization at the water/air interface, it was hypothesized that there was a repulsive interaction between the dipoles of colloidal particles trapped at the interface, and that this was due to the asymmetric charge distribution on the particle surface driving the particles to self-assemble in an ordered array (Figure 6.5) [30]. Later, Park *et al*. developed the technique of *heat-assisted interfacial colloidal crystallization*, the success of which relied

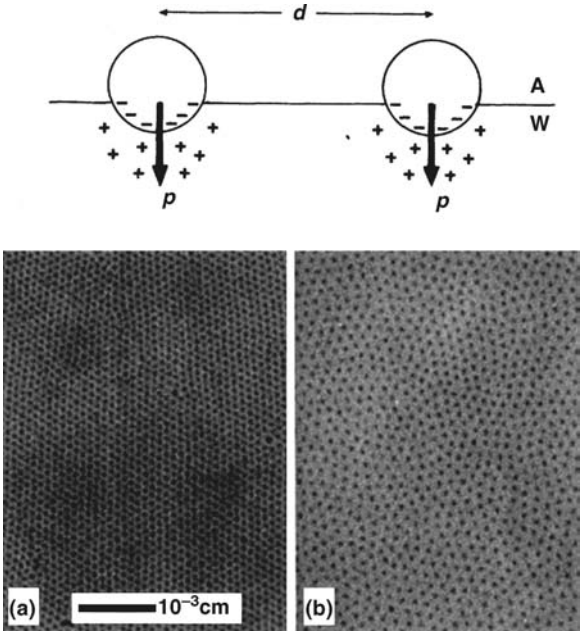


Figure 6.5 Upper panel: Schematic of the model of interaction of colloidal particles at the water (W)/air (A) interface. Lower panel: Photographs of polystyrene spheres (black dots) trapped at water/air interface. (a) Crystalline structure; (b) Disordered structure. Reproduced with permission from Ref. [30].

on the convective flow generated during heating rather than on the interface activity of the colloidal particles [31]. Once 2-D colloidal crystals have been formed at the water/air interface, the LB technique can be used to transfer them onto different substrates [32–35]. Significantly, the LB technique allows a repetition of the transfer of 2-D colloidal crystals onto a substrate into 3-D colloidal crystals with precisely defined layer numbers [35].

In comparison with the water/air interface, a water/oil interface represents a much better platform to trap colloidal particles, due to the relatively low interfacial tension [29]. Thus, water/oil interfaces have been used to grow 2-D colloidal crystals [36, 37], while transfer of the resultant 2-D colloidal crystals to solid substrates remains problematic. In addition to water/air interfaces, air/water/air interfaces have also been used for colloidal crystallization. For example, Velikov and colleagues have studied colloidal crystallization in thinning foam films [38], whilst by using air/water/air interfaces for crystallization Wang and coworkers have successfully obtained free-standing and crack-free colloidal crystal films with sizes in excess of several square millimeters [39]. In another study, rather than use a water/air interface, Zental and coworkers used the interface between melted germanium and air for colloidal crystallization, and obtained crack-free colloidal crystals [40].

6.3
Top-Down Modification of Colloidal Masks

As discussed above, many techniques have been developed to produce 2-D colloidal crystals of high quality. However, in order to increase the structural complexity of the surface patterns obtained via CL, the as-prepared 2-D colloidal crystals can be either etched or deformed, using physical or chemical methods, to tune the size and geometry of the interstices between the solid particles in the crystals. An overview of the strategies developed to modify colloidal crystals is provided in the following subsections.

6.3.1
Controlled Deformation

In general, polymers undergo a second-order phase transition from hard glassy state to a soft rubbery state above a glass transition temperature (T_g), due to the free-volume change between the polymer chains. Therefore, annealing slightly above the T_g can cause the deformation of spherical polymeric beads. It has been shown that, compared to heating in an oven, microwave radiation can provide a much more precise control of such deformation, since its intensity can be easily adjusted [41]. Giersig and coworkers have recently developed a new annealing technique in which a microwave pulse is used to heat polystyrene (PS) microspheres, in a mixture of good and poor solvents for PS. This allowed not only a reduction in the sizes of the interstices of 2-D PS colloidal crystals, but also a deformation of their geometry, from triangular to rodlike, while preserving the interparticle spacing and packing order of

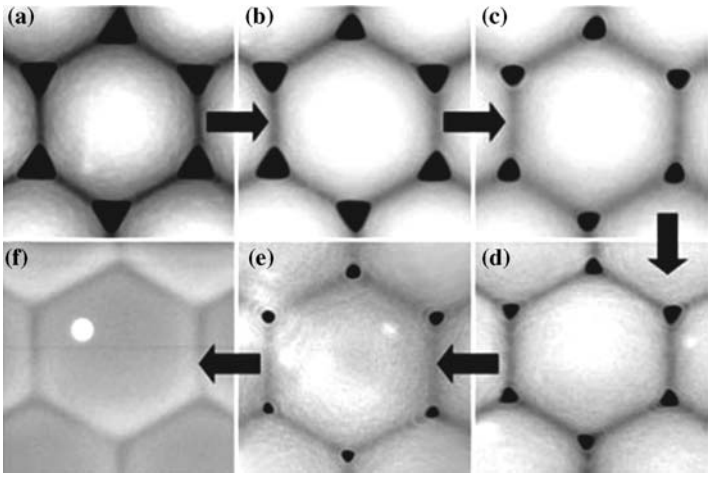


Figure 6.6 Precise control of the degree of annealing is achieved via adjustment of the number of microwave exposures. A 540 nm polystyrene latex mask annealed in 25 ml of water/EtOH/acetone mixture by (a) 1, (b) 2, (c) 4, (d) 6, (e) 7, and (f) 10 microwave pulses. Reproduced with permission from Ref. [42].

the original crystals (Figure 6.6) [42]. Recently, Yang *et al*. described a photolithographic process for the production of hierarchical arrays of nanopores or nanobowls by using colloidal crystals of photoresist particles [43]. In this case, the major difference in T_g between the crosslinked (UV-exposed) and non-crosslinked (UV-screened) particles was the most favorable factor for producing a high contrast in interstitial pore sizes during the baking stage. Although, in the case of inorganic particles, deformation is difficult to achieve by thermal annealing, Polman and colleagues have successfully deformed silica@Au core-shell microspheres into oblate ellipsoids by using high-energy ion irradiation. Such transition occurred due to the fact that the ion-induced deformation of the silica core was counteracted by the mechanical constraint of the gold shell (Figure 6.7) [44]. Vossen and coworkers recently reported that silica particles could undergo an anisotropic deformation under ion bombardment, due to an expansion in the plane perpendicular to the ion beam [45].

6.3.2
Reactive Ion Etching

During the early 1980s, Deckmann and Dunsmuir pioneered investigations into the etching of a colloidal crystal into a textured surface, using a reactive ion beam [46]. Since then, reactive ion etching (RIE) has been widely used to interdependently reduce the particle sizes and thus widen the interstitial space in 2-D colloidal crystal masks; this eventually led to the close-packing structures of the crystals to become non-close-packing (*vide infra*). In 3-D colloidal crystals, RIE is an anisotropic process,

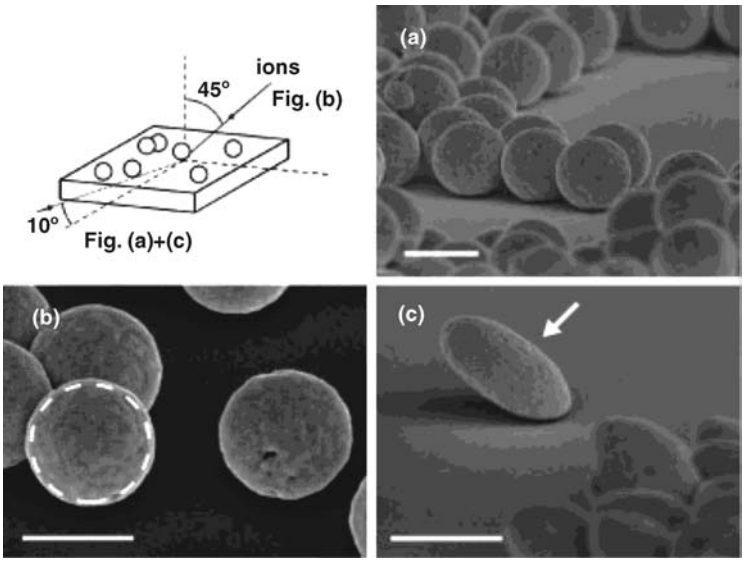


Figure 6.7 Scanning electron microscopy (SEM) images of silica-core/Au-shell colloids on a silicon substrate. (a) Unirradiated; (b, c) Irradiated with 30 MeV Cu ions at 45° and at 77 K, to a fluence of 5×10^{14} ions cm^{-2}, viewing angle parallel (b) or almost perpendicular (c) to the ion-beam direction. The scheme shown at the upper-left shows the ion-beam direction and the SEM viewing angles. Scale bars = 500 nm. Reproduced with permission from Ref. [44].

as the upper layers act as shadow masks for etching the lower layer particles. Such anisotropic RIE can cause spherical particles to become nonspherical particles, while the particle shapes and hierarchical nanostructures obtained will depend heavily on the stacking sequence of the colloidal crystals, the crystal orientation relative to the substrate, the number of colloidal layers, and the RIE conditions employed (Figure 6.8) [47]. Of greatest significance is the fact that the anisotropic RIE can provide a new method for machining the surfaces of the colloidal particles. First, the double layers of PS colloidal crystals were partially filled with silica nanoparticles, such that removal of the top layer PS particles left behind an ordered macroporous silica matrix with regularly arranged openings, beneath which were located the bottom layer particles. The macroporous silica matrices were then used as masks for further RIE of the PS particles beneath. The nanopores, which were arranged in threefold or fourfold symmetry depending on the crystalline orientation of the original colloidal crystals, were then machined on the PS particles [48]. Together, the integration of an anisotropic RIE protocol, the use of binary colloidal crystals composed of PS and silica particles with identical or different sizes as masks, and the use of macroporous matrices as masks represented a powerful method of sculpting spherical particles to multifaceted and nanobored particles [49, 50]. The morphologies of the resultant PS particles were largely dependent on the crystal orientation with respect to the etchant flow, the number of colloidal layers, the size ratio of silica to PS microspheres, the etching angle in the RIE process, the stacking sequence of

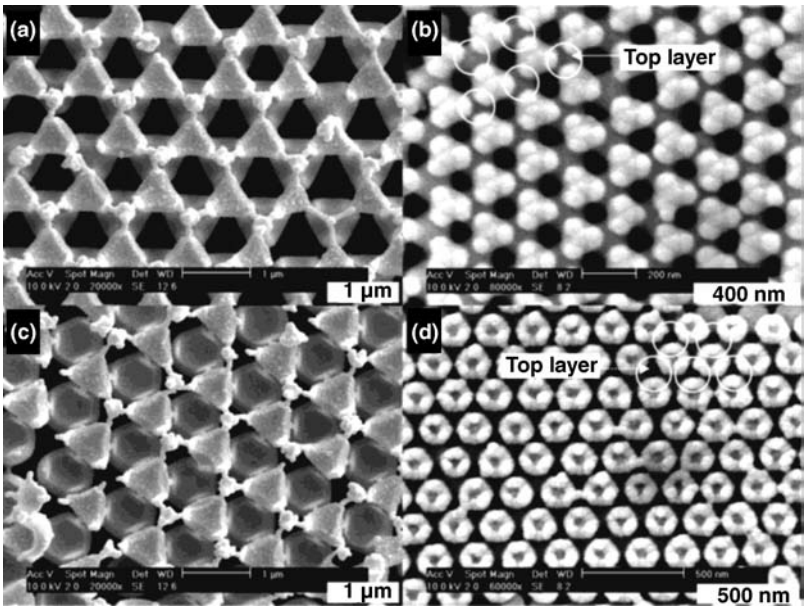


Figure 6.8 Modification of a mask using reactive ion etching (RIE) for the fabrication of binary and ternary particle arrays with nonspherical building blocks. (a, b) Triangle arrays using binary and ternary colloidal spheres with an *hcp* arrangement; (c, d) Polygonal structures produced from colloidal layers with the (111) plane and the (100) plane of the face-centered-cubic structure, respectively. Reproduced with permission from Ref. [47].

the binary colloidal layers, and the tilt angle of the substrate and the orientation angle of the crystal plane with respect to the etchant flow in the RIE process.

6.4 Colloidal Lithography

6.4.1 Colloidal Mask-Assisted Etching

When a 2-D colloidal crystal is formed on a solid substrate, the interstices between the solid particles can used as masks for reactive ions to create patterned bumps or pores on the substrate. Forests of silicon pillars with diameters less than 500 nm and an aspect ratio of up to 10 were fabricated by first, conducting an O_2 RIE to turn close-packed PS particle monolayers into non-close-packed layers, and subsequently conducting a "Bosch" process to etch the supporting silicon wafers [51]. Subsequent scanning electron microscopy (SEM) imaging showed that the etching rate of the resultant structures decreased in line with the increased aspect ratio, which suggested that the etching process was limited by the chemical transport

rate. The reactive plasma can be dispersed by a particle in a point-contact with a substrate, which leads to the induction of so-called "underetching" of the colloidal mask and eventually produces a dome structure of the substrate. Underetching can be avoided by modifying the shape, size, and coverage of the colloidal mask. Sow *et al*. have demonstrated the characteristic features of a RIE silicon substrate using a PS colloidal crystal mask, and produced a double-dome structure by a simultaneous etching of the mask and the regions beneath the particles [52]. When compared to conventionally used polymer masks, such as photoresists removed by organic developers, colloidal masks can be removed easily by sonication, thereby causing very little damage to the nanostructured substrates obtained via RIE. Ordered arrays of polyacrylic acid domes have been fabricated by using 2-D PS colloidal crystals as masks for the O_2 RIE of polymeric films; removal of the PS masks caused no damage to the surface chemistry and the structure of the resultant polymeric domes, thus enabling the conjugation of proteins [53]. Previously, 2-D PS colloidal crystals have also been used as masks for the dry etching of SiO_2 slides so as to create periodic arrays of nanoplates that can be transferred onto polymer films by imprinting [54]. By using colloidal crystals as masks for catalytic etching, Zhu *et al*. have fabricated large-scale periodic arrays of silicon nanowires, the diameters, heights and center-to-center distances of which could be accurately controlled (Figure 6.9) [55]. In a similar study, by using colloidal crystals as masks to create arrays of nanopores on supporting solid substrates via RIE, followed by the consecutive deposition of gold films and removal of the colloidal masks, Ong *et al*. fabricated 2-D ordered arrays of gold nanoparticles nested in the nanopores of the templated substrate [56]. One potential extension of having gold nanoparticles confined in nanopores would be their use as catalysts for the growth of nanowires composed of other materials, such as ZnO.

6.4.2
Colloidal Mask-Assisted Chemical Deposition

By combining microcontact printing with colloidal crystal masking, Xia *et al*. were able to develop a simple method, termed edge-spreading lithography (ESL), that could be used to generate mesoscopic structures on substrates [57]. As the name suggests, ESL utilizes the edges of masks – the perimeters of the footprint of particles on substrates – to define the features of the resultant structures. The ESL procedure begins with the formation of 2-D colloidal crystals of silica beads on the surfaces of gold or silver thin films. Silica beads are used for several reasons: (i) they are inert to most organic solvents; (ii) they are commercially available as monodispersed samples in a range of sizes; (iii) they can be readily assembled into ordered arrays over large areas; (iv) they are mechanically more robust than most polymer beads of equivalent size; and (v) their hydrophilic surfaces support the spreading of the thiols [57]. As shown in Figure 6.10a, typically, a planar polydimethylsiloxane (PDMS) stamp bearing a thin film of the ethanol solution of an alkanethiol was placed on a 2-D silica colloidal crystal. The thiol molecules were then released from the stamp to the silica particle during contact, and subsequently transferred to the substrate along the

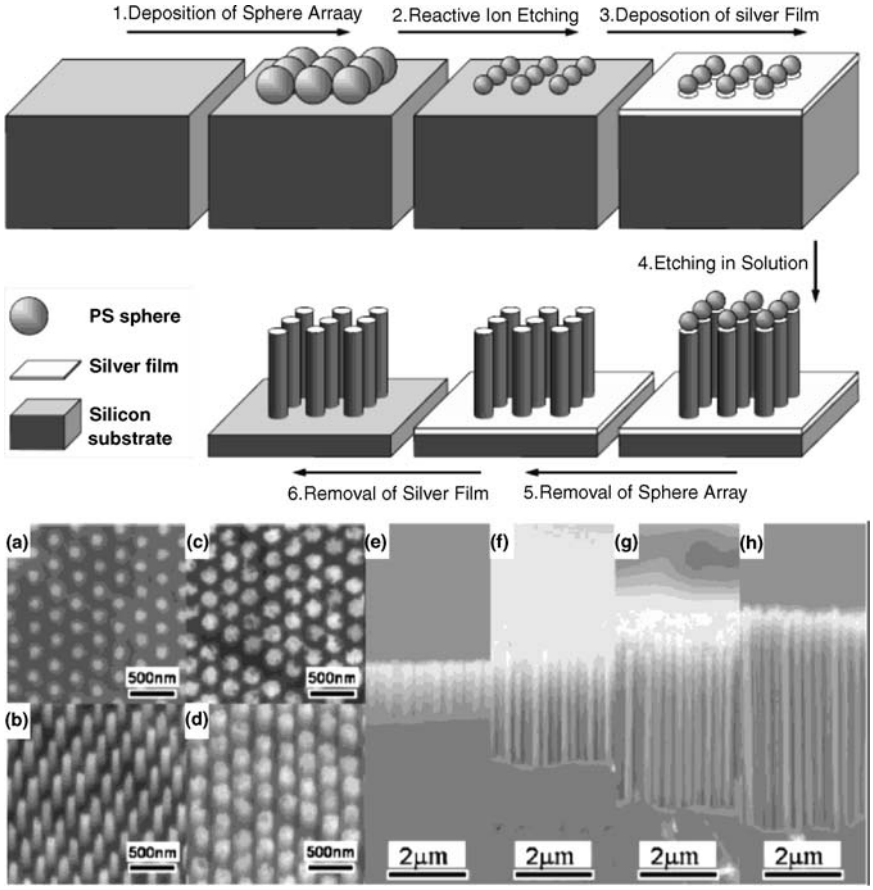


Figure 6.9 Upper panel: Schematic depiction of the fabrication process. Lower panel: SEM images of samples where PS spheres with a nominal diameter of 260 nm have been used. Plane-view and title-view (ca. 15°) images of samples fabricated using PS spheres with a reduced diameter of (a, b) 100 nm and (c, d) 180 nm; (e–h) Cross-sectional SEM images of samples after etching for (e) 4 min, (f) 8 min, (g) 12 min, and (h) 16 min. Reproduced with permission from Ref. [55].

surfaces of the silica particles; this led to the creation of a self-assembled monolayer (SAM) that encircled the footprint of each silica particle. The area of the thiol SAM was able to expand laterally via reactive spreading, as long as the thiols were continuously supplied, such that the width of the thiol SAM rings could be varied between 30 and 340 nm. Following removal of the stamp and bead lift-off, the ring pattern was developed by wet-etching with aqueous Fe^{3+}/thiourea, using the patterned SAM as a resist [57]. The most important point here was that ESL allowed the generation of concentric rings of different alkanethiol SAMs by successive printing with different thiol inks, while removal of the silica particle templates and selective etching yielded concentric gold rings (Figure 6.10b) [58].

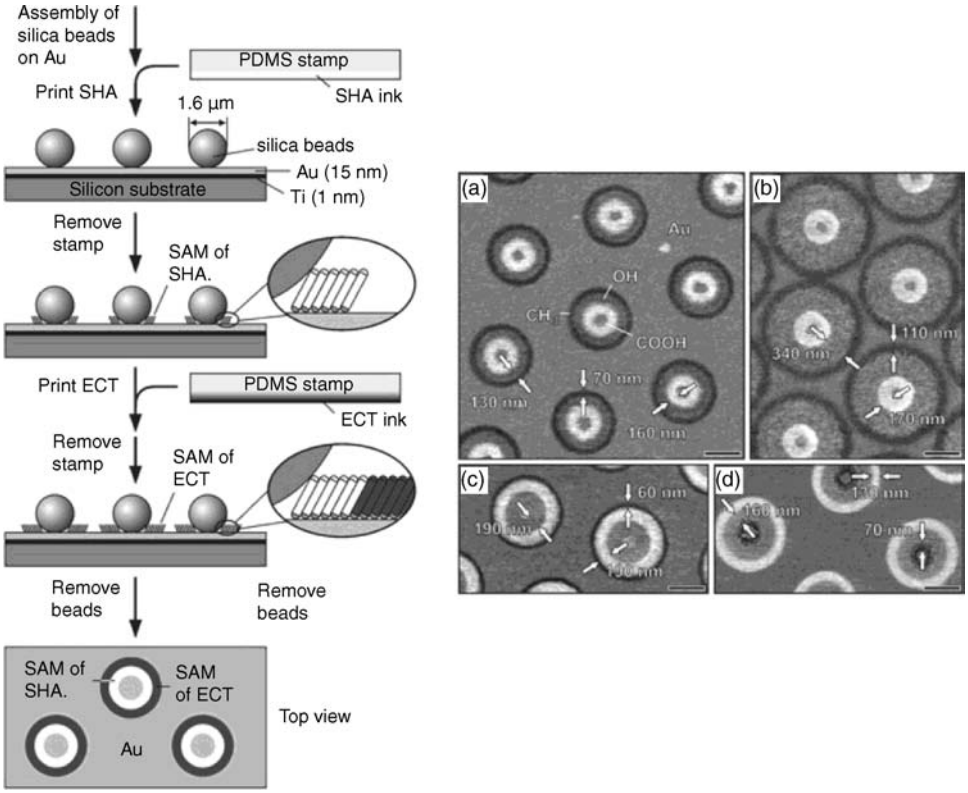


Figure 6.10 Left column: Schematic illustration of the edge-spreading lithography (ESL) procedure used for side-by-side patterning of sulfanylhexadecanoic acid (SHA) and eicosanethiol (ECT) monolayer rings on a gold substrate. The process involves two successive prints that are performed on a 2-D array of silica beads supported on a thin film of gold. In the first step, SHA molecules (white) are guided from a planar stamp to the gold surface, where they assemble into a monolayer, as directed by the circular footprint of each bead and lateral spreading. In the second step, ECT molecules (black) are applied in a similar fashion, thus forming a self-assembled monolayer (SAM) that emerges from the edges of the SHA monolayer. Removal of the beads yields an array of concentric rings of SHA and ECT SAMs on the surface. Right column: Lateral force microscopy (LFM) images of concentric rings of carboxy- (bright), hydroxy- (gray), and methyl-terminated (dark) thiolate monolayers on gold. (a) The rings were fabricated under the following conditions: 1 min for SHA, 1.5 min for 12-hydroxydodecanethiol (HDDT), and 3 min for ECT; (b) An increase in the printing times for HDDT and ECT to 3.5 and 4 min, respectively, resulted in wider rings for these two monolayers; (c, d) The position of each monolayer in the concentric structure could be varied by changing the printing order. The pattern in (c) was generated by printing HDDT for 1.5 min, followed by printing of SHA and ECT for 3 min each. The sample shown in (d) was prepared by printing both ECT and HDDT for 1 min, and SHA for 2 min. All scale bars = 500 nm. Reproduced with permission from Ref. [58].

Shin *et al.* have developed an alternative means of integrating colloidal masking and contact printing, termed contact area lithography (CAL), which can be used for the direct generation of periodic surface chemical patterns at the sub-100 nm scale [59, 60]. In contrast to ESL, CAL relies on the self-assembly of octadecyltrichlorosilane (OTS). Following the formation of a 2-D colloidal crystal of silica on a silicon wafer, the SAM of OTS was grown homogeneously both on the silica particles and on the supporting silicon wafer, via a sol–gel process. The removal of silica particles left behind a periodically arranged array of openings in the OTS SAM, with the same symmetry as that of the 2-D colloidal crystals. The openings were subsequently used as masks, either for the growth of ordered arrays of nanoparticles (e.g., of titania), or for the selective etching of ordered arrays of silica cavities on the silicon wafer. In the case of titania growth, nucleation proved to be rather site-selective due to significant differences in surface energy between the growing and surrounding surfaces [60].

6.4.3
Colloidal Mask-Assisted Physical Deposition: Nanosphere Lithography

In 1981, Fischer and Zingsheim were the first to use 2-D colloidal crystals as masks for contact imaging with visible light [22], whilst a year later Deckman and Dunsmuir demonstrated the feasibility of using 2-D colloidal crystals as masks for both the physical deposition of materials and, in turn, patterning the surfaces of the supporting substrates (Figure 6.11) [61]. Consequently, the latter authors coined the term "natural lithography" to describe this process, since "naturally" assembled single layers of latex particles were used as masks rather than lithographic masks. Later, the capabilities of natural lithography were expanded, with the RIE process in particular

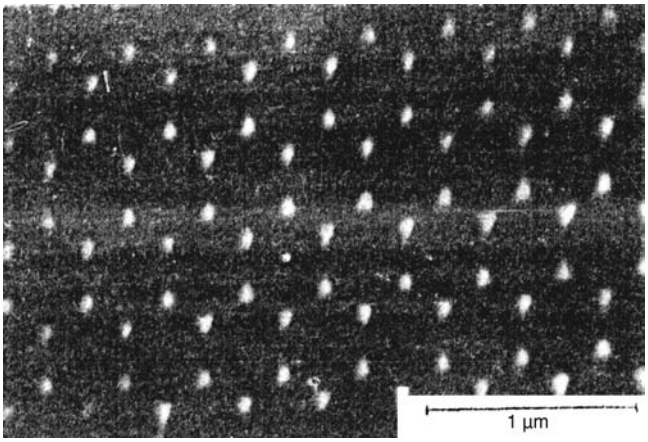


Figure 6.11 Triangular silver posts fabricated using a lift-off process. Silver was evaporated over a densely packed array of 0.4 μm spheres, and the spheres were dissolved in methylene chloride. Reproduced with permission from Ref. [61].

being developed to increase the structural complexity of 2-D colloidal crystal masks [46]. Since then, the group of Van Duyne has expended much effort towards developing patterning techniques that used colloidal crystals as masks for metallic vapor deposition [23, 62–64]. In the context of nanoscience, the term "natural lithography" was changed to "nanosphere lithography" (NSL), and its use explored over a variety of experimental parameters, notably of the incident angle that would lead to a diversification of the patterns obtained. The same group also extended single-layer masking to double-layer masking and, perhaps most importantly, conducted intensive investigations into the plasmon resonance properties of metallic patterns obtained via NSL. In this case, the correlation with feature morphology was of particular interest, the intention being ultimately to develop highly sensitive biosensors based on surface-enhanced Raman spectroscopy (SERS) [65]. Following the seminal studies of Van Duyne and colleagues, colloidal crystals came to be identified as being low-cost, flexible, and easily adoptable masks for the growth of new nanostructures with a diverse structural complexity. The uses of NSL and their variants for surface patterning on planar substrates, and especially on microparticles, are summarized in the following subsections.

6.4.3.1 Surface Patterning on Planar Substrates

In the NSL procedure, a 2-D colloidal crystal is used as a mask for the physical deposition of a material, the latter being freely chosen without any limitations, and often including various metals such as gold and silver. The projection of the interstices between ordered close-packed particles defines the shape of the nanodots deposited on substrates; the dots usually show a quasi-triangular shape, and are arranged in a P_{6mm} array due to the hexagonal packing of the colloidal crystal mask (Figure 6.12a–c). The dot size is about one-fourth of the particle diameter, while the distance between nearest-neighbor dots is about one-half of the sphere diameter. The dot height is controlled by the physical deposition conditions, notably the deposition time. Van Duyne *et al.* have extended colloidal crystal masking from single layers of hexagonally close-packed particles to double layers [23]. Since overlapping of the interstices between the upper and lower layers leads to an hexagonal array of quasi-hexagonal projections on a substrate, the use of a double-layer colloidal crystals as a mask will yields an hexagonal array of quasi-hexagonal nanodots (Figure 6.12d and e).

In a general NSL procedure, the substrate to be patterned is positioned normal to the direction of material deposition. The in-plane shape of the nanodots and spacing of the nearest-neighbor dots derived from NSL are then dictated by a projection of the interstices of single or double layers of colloidal crystals on the substrates. These can be fine-tuned by varying the projection geometry of the interstices on substrates, for example by tilting the masks with respect to the incidence of the vapor beam. This approach has inspired the development of angle-resolved NSL (AR-NSL), as pioneered by the group of Van Duyne [64].

In the AR-NSL process, the incident angle of the propagation vector of the material deposition beam with respect to the normal direction of the colloidal mask (θ) and/or the azimuth angle of the propagation vector with respect to the nearest-neighbor

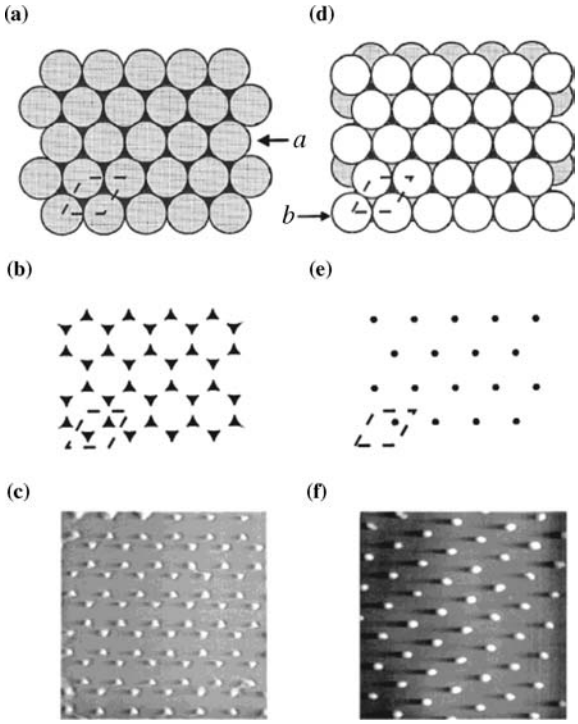


Figure 6.12 Schematic diagrams of single-layer (SL) and double-layer (DL) nanosphere masks and the corresponding periodic particle array (PPA) surfaces. (a) *a* (111) SL mask, dotted line = unit cell, *a* = first layer nanosphere; (b) SL PPA, two particles per unit cell; (c) 1.7 × 1.7 μm constant-height AFM image of a SL PPA with M = Ag, S = mica, $D = 264$ nm, $d_m = 22$ nm, $r_d = 0.2$ nm s^{-1}; (d) *a* (e) $p(1 \times 1)$-*b* DL mask, dotted line = unit cell, *b* = second layer nanosphere; (f) DL PPA, one particle per unit cell; (g) 2.0 × 2.0 μm constant-height AFM image of a DL PPA with M = Ag, S = mica, $D = 264$ nm, $d_m = 22$ nm, $r_d = 0.2$ nm s^{-1}. Reproduced with permission from Ref. [23].

particles in the colloidal masks (φ) – the mask registry with respect to the vector of the material deposition beam – have been used to reduce the size of the nanodots obtained and, at the same time, to elongate their triangular shape (Figure 6.13). By rotating substrates, Giersig and coworkers have recently found that AR-NSL can generate much more complicated metallic nanostructures, and they referred to this process as "shadow NSL" [42, 66, 67]. Due to rotation of the colloidal mask, the shadow NSL process is resolved by the azimuth angle (φ) of the incidence deposition beam rather than the incident angle (θ).

An elegant extension of AR-NSL was to conduct a stepwise physical vapor deposition (PVD) of identical or different materials, but at different angles of incidence. In this case, the group of Van Duyne succeeded in growing surface-patterning features composed of two triangular nanodots that were either overlapped or separated by two deposition steps at $\theta = 0^\circ$ and $\theta > 0^\circ$, respectively [63]. The

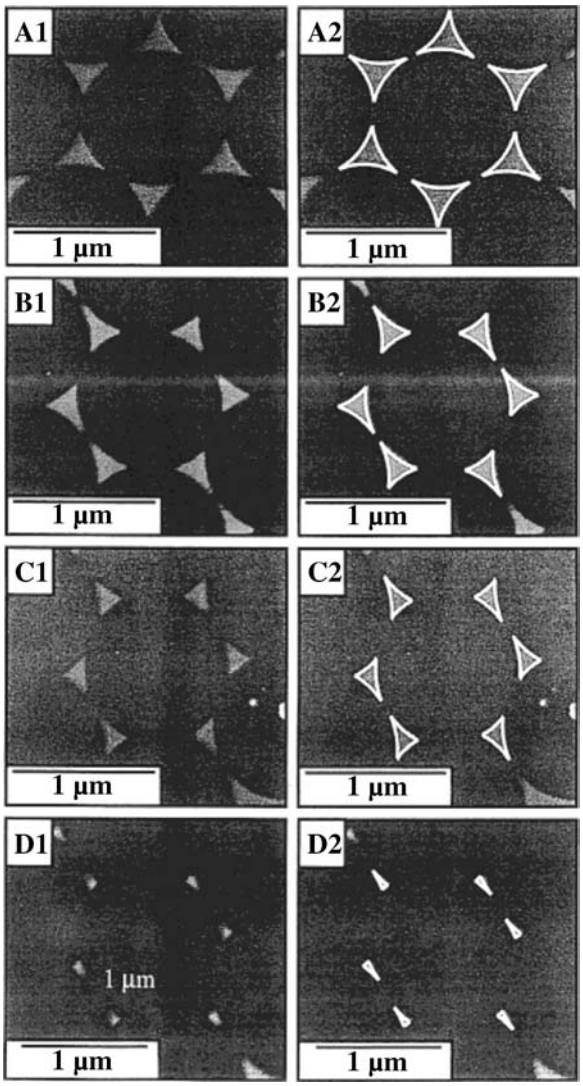


Figure 6.13 Field-emission SEM images of AR NSL-fabricated gold nanodot arrays and images with simulated geometry superimposed, respectively. (A1, A2) $\theta = 108°$, $\phi = 288°$, (B1, B2) $\theta = 208°$, $\phi = 28°$, (C1, C2) $\theta = 268°$, $\phi = 168°$, and (D1, D2) $\theta = 408°$, $\phi = 28°$. All samples were Cr-deposited onto Si(111) substrates. Original magnification of images $= \times$ 40 000. θ is the incidence angle and ϕ the azimuth angle. Reproduced with permission from Ref. [64].

overlapping or spacing between the two nanodots was seen to depend on the incident angle at the second step of deposition. When three deposition steps at three different incident angles – zero, tilted forward and tilted backward – were conducted, chains of triangular dots were obtained. although the registry of colloidal crystal masks – the

azimuth of the incident deposition beam (φ) – was changed only minimally. Giersig *et al*. have also developed a stepwise shadow NSL protocol to deposit different materials at different incident angles when the colloidal masks were rotating, and have succeeded in encapsulating the metallic structures so as to protect them against oxidation [67].

Recently, Zhang and Wang demonstrated the feasibility of consecutively depositing two different metals, such as gold and silver, at two different incident angles, in order to construct ordered binary arrays of gold and silver nanoparticles [68]. This approach was seen to be independent of the sphere sizes of the colloidal masks and the chemical nature of materials deposited, but did demonstrate a profound dependence on the registry of colloidal masks with respect to the incident vapor beam and the incident angle. When the projection of the incident beam onto the substrates was coincident with the vector between the nearest-neighbor particles, triangular gold and silver nanodots were obtained, interspaced by a tiny gap, and each of these was arranged in an array with P_{6mm} symmetry (Figure 6.14). However, when the projection of the incident beam on the substrates was coincident with the vector between the next-nearest-neighbor particles, then triangular gold nanodots, triangular silver nanodots, and rectangular nanodots composed of triangular gold and silver nanodots were obtained, each of which was arranged in an hexagonal array.

Prior to PVD, colloidal crystal masks can undergo RIE to reduce the sizes of the particles and widen the interstitial spaces, thus increasing the dimensions of the triangular nanodots obtained via NSL. Increasing the RIE time can also cause

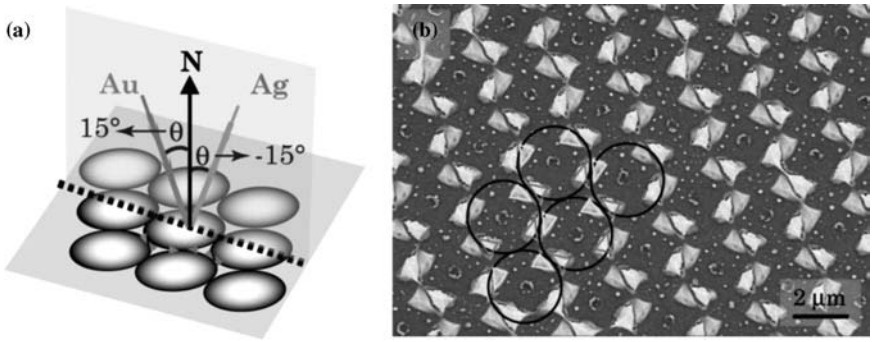


Figure 6.14 (a) Schematic illustration of depositing gold and silver onto a hexagonally close-packed sphere monolayer at incident angles (θ) of 15° and −15°, respectively. The colloidal mask is registered so that the vector between nearest-neighbor spheres is in line with the projection of the incidence beam on the mask, highlighted by a black dotted line. The incidence beams of gold and of silver, and the normal direction of the colloidal template are highlighted by yellow, blue, and black arrows, respectively. The incident angles (θ) of the vapor beam with respect to the normal direction of the colloidal masks are marked as red; (b) SEM image of the resultant heterogeneous binary array. The mask is a monolayer of hexagonally close-packed 830 nm PS spheres. The original location of PS spheres, gold nanoparticles (NPs), and silver NPs are highlighted by black circles, yellow triangles, and blue triangles, respectively. Reproduced with permission from Ref. [68].

close-packed colloidal crystal masks to become non-close-packed, which in turn leads to thin films with hexagonally arranged pores [69, 70]. Wang *et al*. have recently integrated AR-NSL with the use of RIE-modified colloidal crystals as masks, to diversify the structural complexity of the patterning feature derived from NSL from triangular (or deformed) nanodots to nanorods and nanowires [71]. After modification via O_2 RIE, 2-D colloidal crystal masks were registered so that the projection of the metal vapor beam on the colloidal mask was coincident with the vector between the nearest-neighbor particles. When PVD was conducted at the incident angle of 45°, zigzag nanowires were obtained that were well separated and aligned in parallel. However, when the projection of the metal vapor beam on the colloidal mask was adjusted in line with the vector between the next-nearest-neighbor particles, only nanorods were obtained and these were arranged in an hexagonal array. A stepwise rotation of the colloidal crystal masks by 120°, to deposit identical or different materials, led to quasi-3-D grids of nanowires or nanorods with a defined vertical, and especially lateral, heterogeneity (Figure 6.15). The lateral arrangement of different nanowires into a periodic array with a defined alignment is difficult to implement by other means, whether conventional lithographic or self-assembly techniques.

Wang *et al*. have recently extended the RIE process for the modification of double layers of colloidal crystals for AR-NSL [72]. By using O_2 plasma-etched bilayers of hexagonally packed particles as masks for gold deposition, it was possible to fabricate highly ordered binary arrays of gold nanoparticles of various shapes, including a shuttlecock-shaped array composed of small, crescent-shaped nanoparticles, and a large fan-shaped array (see Figure 6.16). The size and shape of both the small and large nanoparticles obtained could be manipulated by altering the plasma-etching period and the incident angle of the Au vapor flow. When compared to the corresponding bulk materials, the melting point of the nanoparticles was much lower, and they were much more sensitive to the surface tension. As the large curvature caused a high surface tension, annealing of the non-round nanoparticles might give rise to a retraction of their apexes, and eventually the creation of a round shape [73]. Wang *et al*. have successfully transformed the shape of Au nanoparticles obtained from crescent- or fan-like array to a round form, with a rather narrow distribution in terms of size and shape (Figure 6.17) [72].

Dmitriev *et al*. have extended colloidal crystal masking from a use for material deposition to one of controlled etching, and have developed an interesting variant of NSL, termed hole-mask colloidal lithography (HCL) [74]. HCL differs from conventional NSL in that the substrate and colloidal crystal mask are interspaced by a sacrificial layer. After PVD, removal of the colloidal mask leads to a thin film mask with nanoholes. This so-called "hole-mask" is subsequently used for vapor deposition and/or etching steps to further define a patterning feature on the substrate. Removal of the sacrificial layer, along with the hole-mask, leaves behind the substrate with a predesigned surface pattern composed of, for instance, discs, ellipsoids, and cores (Figure 6.18). HCL displays several advantages over NSL, notably a large area coverage, a high fabrication speed (the fabrication time does not scale with area), an independent control over the feature size and spacing, and processing simplicity

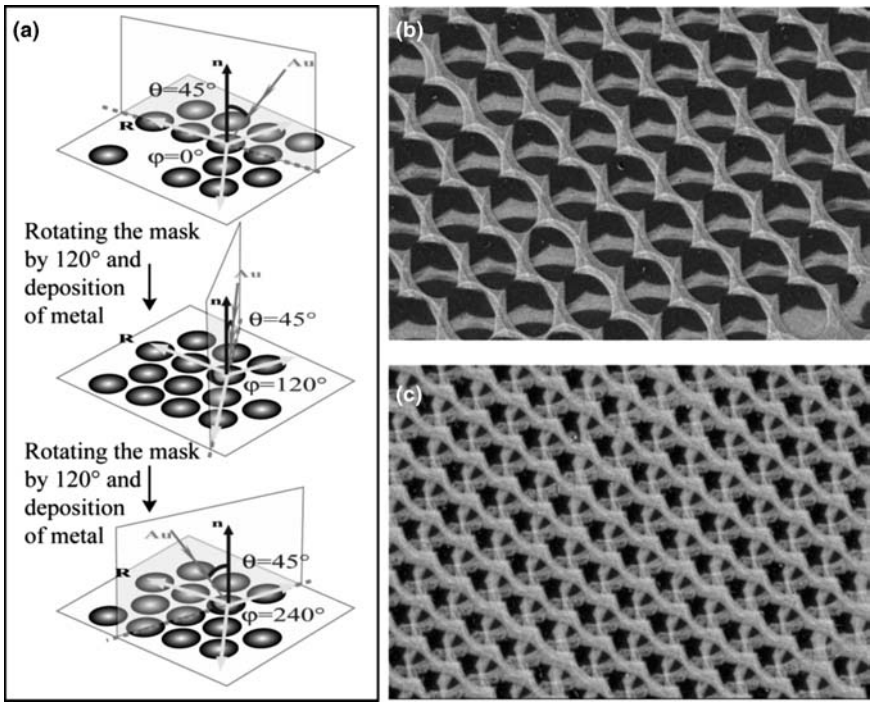


Figure 6.15 (a) Schematic depiction of constructing quasi-3-D grids of multiplex zigzag nanowires by stepwise rotation of the colloidal mask by 120° with respect to the reference vector (*R*) between nearest-neighboring spheres over the course of metallic vapor deposition. The projection of metal vapor on the mask was set coincidence with the reference vector (*R*), namely $\varphi = 0°$. SEM images of quasi-3-D grids of multiplex zigzag nanowires obtained by stepwise depositing gold, silver, and nickel at $\varphi = 0°$, $\varphi = 120°$, and $\varphi = 240°$, using plasma-etched close-packed 830 nm PS sphere monolayers as masks. The structure obtained by two and three deposition steps are shown in panels (b) and (c), respectively. The plasma etching time was 20 min, θ was 45°, and the deposition time 30 min. Reproduced with permission from Ref. [71].

(the nanofabrication process is reduced to conventional material deposition and RIE). Moreover, HCL can be applied to a wide range of materials, including Au, Ag, Pd, Pt, and SiO_2.

6.4.3.2 Surface Patterning on Particles

Various colloidal spheres, whether organic and inorganic, can be produced that are exceedingly monodisperse in terms of their size and shape. Nevertheless, their surfaces remain chemically homogeneous or heterogeneous in an undefined way, despite there being well-established methods for their modification. As this surface chemistry renders the coupling of spheres spatially isotropic, it is difficult to spatially direct the organization of the spheres, and they tend to self-assemble only into simple and energy-favorable *fcc* or *hcp* structures. Controlling the surface properties of

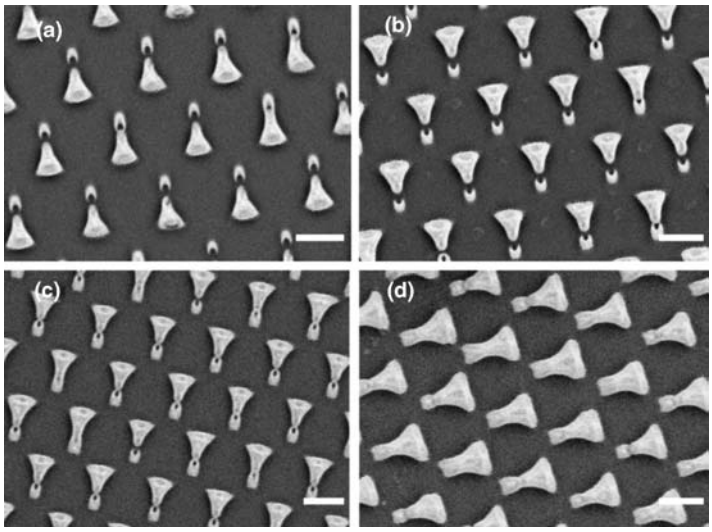


Figure 6.16 SEM images of hexagonally arranged Au nanoshuttlecocks obtained by using bilayers of hexagonal close-packed 925 nm PS spheres, etched by O_2-plasma for 10 (a), 20 (b), 25 (c), and 30 min (d), as masks for Au vapor deposition. The incidence angle of Au vapor flow was set as 15°. Scale bars = 500 nm. Reproduced with permission from Ref. [72].

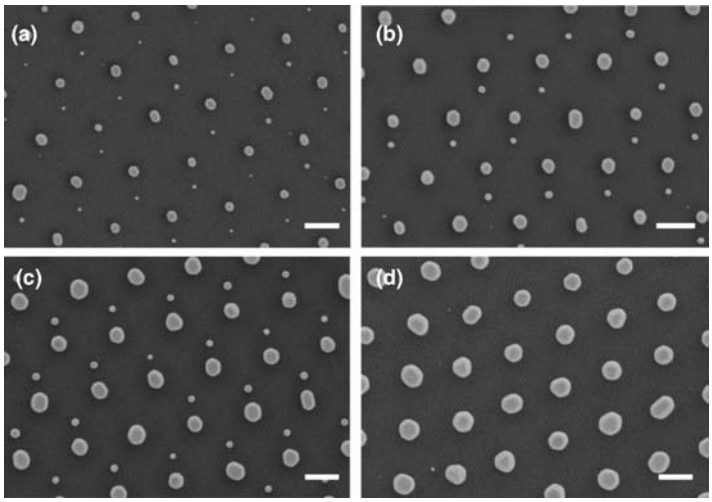


Figure 6.17 SEM images of the hexagonal binary arrays obtained by annealing the Au nanoshuttlecock arrays derived from the hexagonal close-packed 925 nm PS sphere bilayers etched for 10 (a), 20 (b), 25 (c), and 30 min (d). The SEM images of the original shuttlecock arrays are shown in Figure 6.16. Scale bars = 500 nm. Reproduced with permission from Ref. [72].

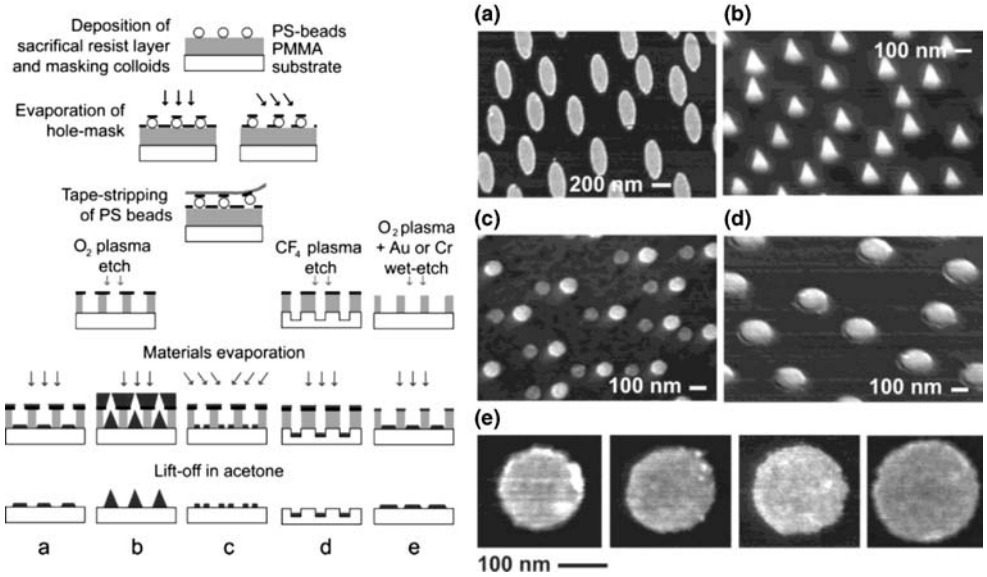


Figure 6.18 Left column: Diagram illustrating the basic process steps and resultant structures produced with HCL nanofabrication. Right column: SEM images of the five different nanostructure types produced by HCL. (a) Array of identically oriented elliptical Au nanostructures; (b) Au nanocone array; (c) Binary arrays of Au–Ag nanodisc pairs (note the slightly different imaging conditions for Au and Ag nanodiscs in each pair; Au is imaged brighter); (d) Embedded nanodiscs; (e) Discs with fine-tuned diameters, where the disc size increases from left to right. Reproduced with permission from Ref. [74].

colloidal particles is one of the oldest and, at the same time, the most vital topics in colloid science and physical chemistry. Patchy particles – that is, particles with more than one patch or patches that are less than 50% of the total particle surface – should present the next generation of particles for assembly [75–77]. However, patterning the surface of colloidal particles with sizes of micrometers or submicrometers represents a formidable challenge, due to the lack of a proper mask.

When 2-D colloidal crystals are used as masks for PVD, it is expected that only the upper surfaces of the colloidal particles (which are exposed directly to the vapor beam) will be coated with the new materials. This leads to the creation of two spatially well-separated halves on the colloidal particles – coated and noncoated – with two distinct surface chemical functionalities [78, 79]. Such particles are usually referred to as *Janus particles*. By embedding a monomer of close-packed colloidal particles in a photoresist layer, Bao *et al*. managed to tune the surface areas of the colloidal particles exposed to the vapor beam during material deposition, by etching the photoresist layer with O_2 plasma; this in turn led to a good control of the domain sizes deposited on the particles [80]. When a monolayer of close-packed colloidal particles is constructed at the water/air interface or the wax/liquid interfaces, a selective modification can be implemented in either of the two phases, and this leads to the production of Janus particles [81, 82]. Shin *et al*. have recently extended the CAL

procedure to decorate silica particles with ordered arrays of titania nanoparticles by selective removal of the upper layer particles [60].

Wang *et al.* pioneered the use of upper single layers of colloidal crystals as masks for the lower layer particles during PVD [83]. By using the upper single layer of a colloidal crystal as masks for gold vapor deposition, various Au patterns were embossed on the upper halves of the particles in the second layer, such as triangles, squares, and bow-ties, the size and shape of which were predominantly manipulated by orientation of the template crystals. Most importantly, the methodology reported by Wang *et al.*, which used colloidal crystals for self-masking, was independent of the curvature and chemical composition of the surfaces (Figure 6.19). This would clearly

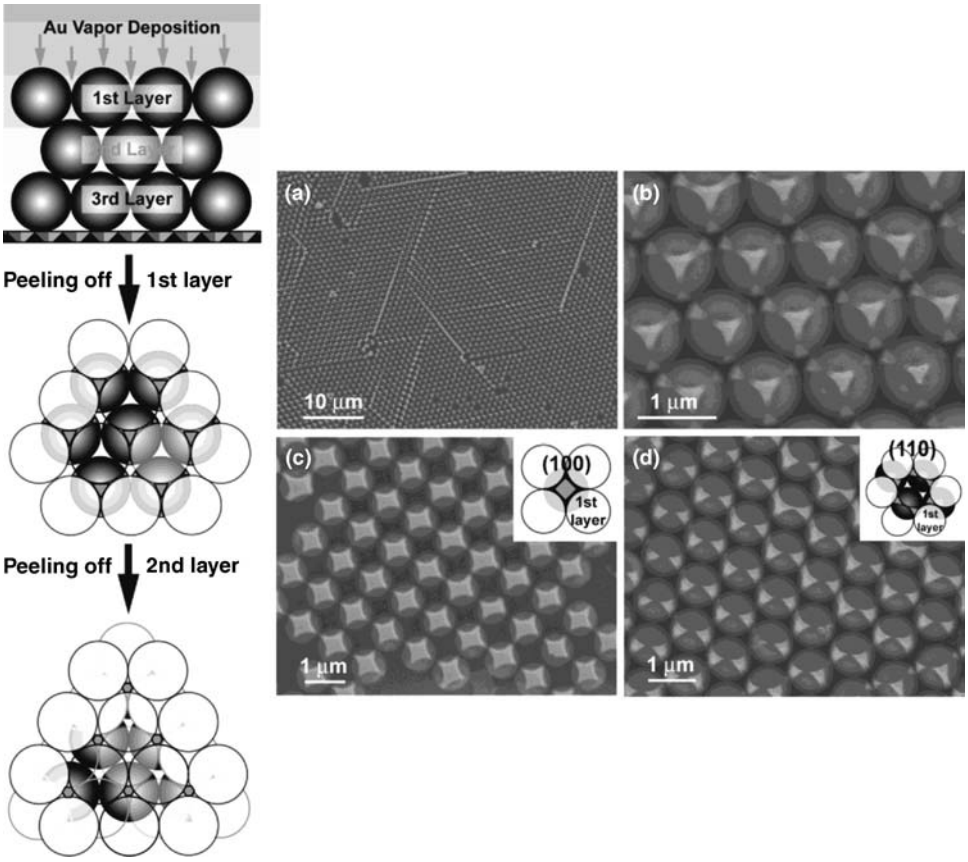


Figure 6.19 Left column: Schematic illustration of the procedure to create colloidal spheres with Au-patterned surfaces by the combination of Au vapor deposition and using the top mono- or bilayers of colloidal crystals with (111) facets parallel to the substrates as masks. Right column: (a) Low-magnification and (b–d) high-magnification SEM images of 925 nm polystyrene (PS) spheres with Au-patterned surfaces, generated by templating the top monolayers of colloidal crystals with (b) (111), (c) (100), and (d) (110) facets parallel to the substrates. Reproduced with permission from Ref. [83].

provide a versatile way in which to pattern highly curved surfaces, a situation that is difficult to achieve by using routine lithographic techniques. By using O_2 plasma to etch the colloidal crystal templates (mainly the top layer) and conducting PVD at the non-zero incident angle, Wang *et al*. also found that the size and shape of the patterns obtained on the second layer particles showed a pronounced dependence on the plasma etching time and the incident angle [84]. Pawar and Kretzschmar have recently extended the use of colloidal crystals for self-masking for glancing angle deposition [85]. During PVD at a glancing incident angle, the shadow effects caused by neighboring particles were used for surface patterning particles with the same particle monolayer. This differed from Wang's strategy, where the upper layers were used as masks for surface patterning the particles in the lower layers. The size and shape of the resultant patterns were determined by the incident angle and monolayer orientation, such that the smallest patch produced via glancing angle deposition was 3.7% of the particle surface.

Wang *et al*. have recently used the upper double layers as masks for patterning particles in the third layers, via PVD [86]. Whilst the smaller interstices in the upper bilayers cause a nonuniform diffusion of the Au vapor, the dimension and features of the Au dots obtained previously were neither uniform nor clear-cut when compared to patterns derived from single-layer masking. In order to achieve an homogeneous diffusion of the vapor through the upper double layers to reach the deeper layers in a colloidal crystal, the crystal was made to undergo RIE with O_2 plasma, which widened the interstitial spaces between the particles. The use of RIE-treated colloidal crystals as a mask has greatly improved the uniformity of the patterns generated on the third layer particles. Notably, widening the interstitial spaces allows more vapor to diffuse into and through an RIE-treated colloidal crystal. Any excess vapor that is scattered or reflected by the spheres beneath the third layer or the substrate would be envisioned to condense into a round dot on the lower half of each sphere in the third layer, opposite to the Au vapor flow. As a consequence, Wang *et al*. succeeded in the stereo-decoration of colloidal particles with two, three, four, or five nanodots. The number of dots per sphere was proved to depend on the crystalline structure of the colloidal crystal masks, the plasma etching time, and the incident angle. The nanodots decorated on particles were arranged in a linear, trigonal, tetrahedral, or right-pyramidal fashion, which provided nanoscale analogues of sp-, sp^2-, and sp^3 -hybridized atomic orbitals of carbon (Figure 6.20). The Au nanodots obtained on microspheres, therefore, could be recruited as the bonding site to dictate the integration of the spheres, thus paving a new approach to colloidal self-assembly – colloidal valent chemistry of spheres [87] – to create hierarchical and complicated "supraparticles" [75].

6.4.3.3 **Extension of Nanosphere Lithography**

One extension of NSL is to use the surface patterns obtained as templates to grow nanostructures of a variety of materials, via bottom-up self-assembly. Mulvaney's group has grown monolayer and multilayer films of semiconductor quantum dots on surface patterns derived from NSL, leading to nanostructured luminescent thin films (Figure 6.21) [88, 89]. Valsesia *et al*. have used ordered arrays of polyacrylic acid

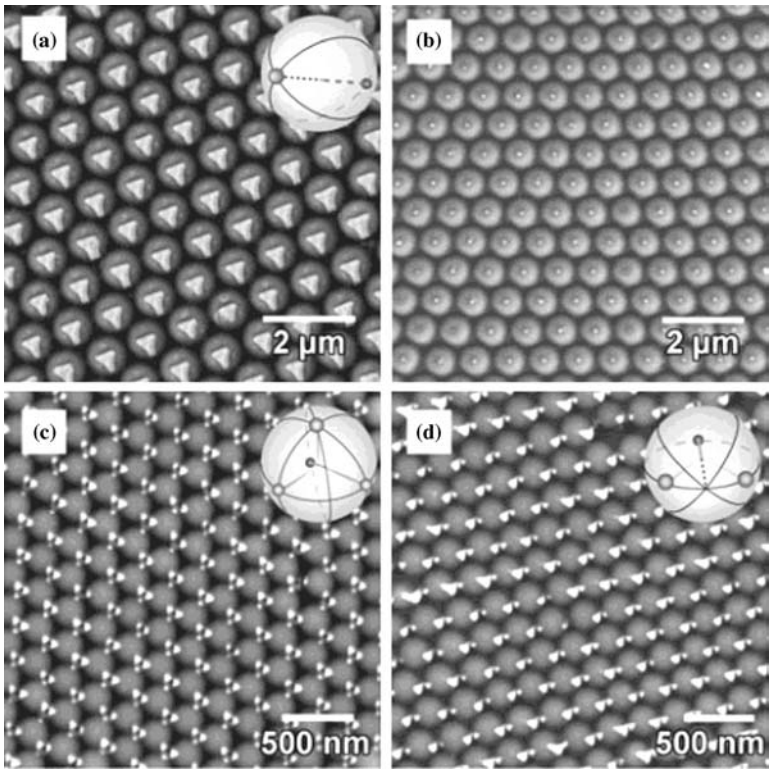


Figure 6.20 (a) SEM image of the gold patterns deposited on the upper halves of 925 nm PS spheres in the third layer, obtained by using their colloidal crystals, and etched by plasma for 10 min, as templates; (b) SEM image of the Au patterns obtained on the lower halves of these spheres; (c, d) SEM images of the gold Au patterns on the upper halves of 270 nm PS spheres by using their colloidal crystals, etched by plasma for 3 min, as templates constructed. The incident angle of the gold vapor was 0° (c) and 10° (d). The insets show schematic illustrations of the spatial configuration of gold nanodots decorated on the microspheres. Reproduced with permission from Ref. [86].

domes derived via NSL to selectively couple with bovine serum albumin [53]. By using NSL-derived surface patterns as templates to grow proteins, Sutherland *et al.* showed that the surface topography could enhance the binding selectivity of fibrinogens to platelets [90].

A second extension is to use NSL-derived surface patterns as etching masks to create surface topography. In this case, Chen *et al.* have fabricated silicon nanopillar arrays with diameters as small as 40 nm and aspect ratios up to 7 [91]. The size and shape of the nanopillars could be controlled by the size and shape of the sputtered aluminum mask, with both parameters being again determined by the feature size of the colloidal mask and the number of the colloid layers. Nanopillars of different shapes can also be fabricated by adjusting the RIE conditions, such as the gas species, bias voltage, and exposure duration for an aluminum mask with a given shape. The

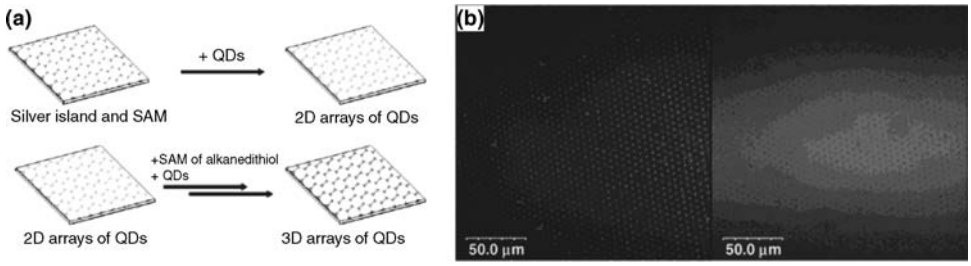


Figure 6.21 (a) Scheme of formation of 3-D quantum dot structures, using a multiple SAM technique based on CL; (b) Photoluminescence and transmission images of QD642 (quantum dots emitting at 642 nm) arrays made with a partially completed bilayer of 5.46 μm PS mask; scan size: 200 × 200 μm. Reproduced with permission from Ref. [89].

as-prepared nanopillar arrays could then be used to imprint a layer of PMMA above its T_g (Figure 6.22) [92]. Similarly, Weekes *et al.* have fabricated ordered arrays of cobalt nanodots for patterned magnetic media [93]. By introducing intermediary layers of SiO_2 between the colloidal crystal masks and substrates, this etching strategy could be applied to a wide range of materials, without too much concern for the surface hydrophilicity of the targeted substrates. By using a similar protocol, large-area ordered arrays of 512 nm pitch holes, and with vertical and smooth sidewalls, have been successfully formed on GaAs substrates [94].

A third extension is to use NSL-derived surface patterns to template or catalyze the growth of other functional materials. Here, Zhou *et al.* have successfully used ordered arrays of gold nanodots derived from NSL as seeds to create highly aligned single-walled carbon nanotubes (CNTs), laid on quartz and sapphire substrates [95]. This method has great potential for the production of CNT arrays, with a simultaneous control over nanotube orientation, position, density, diameter, and even chirality. As a consequence, these CNTs may function as building blocks in future nanoelectronics and ultra-high-speed electronics applications [96]. Wang *et al.* have used gold nanodot arrays as seeds for hexagonally arranged arrays of zinc oxide nanorods, aligned perpendicularly to the substrates [97]. Similarly, Fuhrmann *et al.* have produced ordered arrays of Si nanorods by using the gold nanodots as seeds for molecular beam epitaxy (Figure 6.23) [98], while discretely ordered arrays of organic light-emitting nanodiodes (OLEDs) have been fabricated based on NSL-derived surface patterns [99]. These are not feasible via any other route, as conventional masking techniques may damage the organic heterostructure of the OLED layers.

6.5 Applications of CL

Surface patterns derived via CL, and especially via NSL, are normally composed of metals such as gold and silver, with their primary technical application being highly

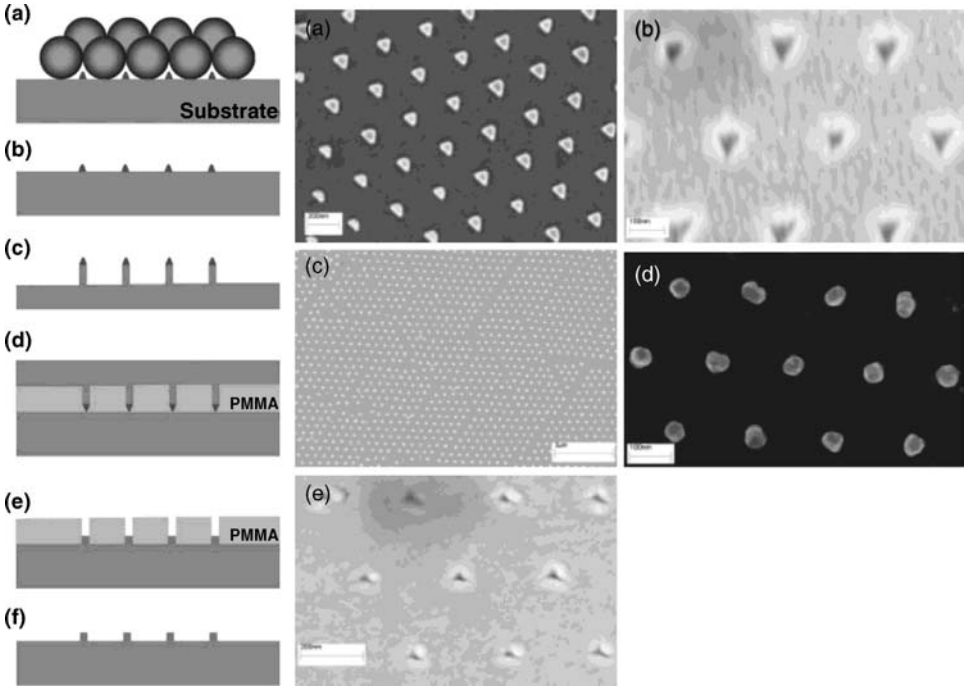


Figure 6.22 Left column: Schematic of the fabrication of large-area periodic nanostructures by a combination of double-layer nanosphere lithography and nanoimprint lithography. (a) The silicon substrate is coated with a double layer of polystyrene spheres and a metal film is deposited on top of the polystyrene beads; (b) After dissolving the polystyrene beads in CH_2Cl_2, periodic metallic arrays are formed on the surface; (c) The silicon substrate is etched using the periodic metal arrays as an etching mask; (d) The silicon nanopillar arrays are used as an imprinting stamp, which is pressed against a PMMA film on a silicon wafer above the polymer's glass transition temperature; (e) The stamp is removed and the desired material deposited; (f) After lift-off, periodic arrays of the desired material are obtained. Right column: (a) SEM image of a nanoimprint stamp fabricated using a 350 nm polystyrene template; (b) The imprinted patterns on PMMA. The base of the triangular hole is about 55 nm; (c) Large-area image of periodic metal dots formed by nanoimprint lithography; (d) SEM image of nanodots formed by nanoimprint lithography. The diameter of the nanodots is about 50 nm; (e) SEM image of the imprinted patterns using a stamp treated with chromium etchant. The lateral dimension of the triangular hole is about 30 nm. Reproduced with permission from Ref. [92].

sensitive biosensors that rely on the localized-surface plasmon resonance (LSPR) of metallic nanostructures [65]. During intensive investigations of the LSPR of metallic nanostructures composed of gold rings [100] and disks [101] and obtained via NSL, it was found that the LSPR could be tuned by varying either the diameter of the disks (at a constant disk height) or the ring thickness. The subsequent shape-dependent red shift originated from the electromagnetic coupling between the inner and outer ring surfaces, and this led to energy shifts and the splitting of degenerate modes [102]. NSL has been also used to create nanocaps and nanocups, the LSPR

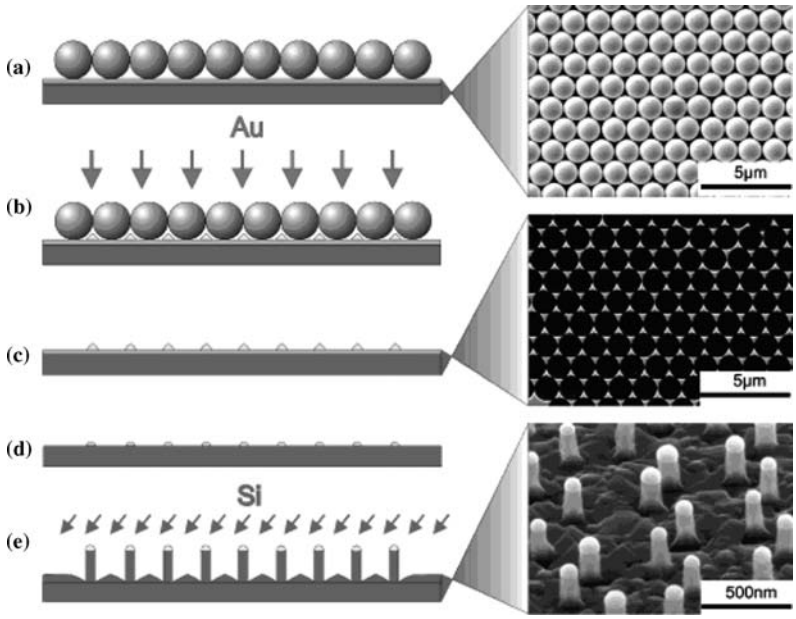


Figure 6.23 Steps of Si nanowire fabrication by NSL. (a) Deposition of a mask of polystyrene particles; (b) deposition of gold by thermal evaporation; (c) removal of the spheres; (d) thermal annealing and cleaning step to remove the oxide layer; (e) Si deposition and growth of nanowires by MBE. The corresponding SEM images at the right show the wafers at the different fabrication steps. Reproduced with permission from Ref. [98].

behavior of which has been studied [103]. For example, by using NSL, Lee *et al.* created gold crescent-moon structures which had a sub-10 nm sharp edge and exhibited a very strong SERS [104]. Moreover, the field emitted on the circular sharp edge of the nanocrescent moon (the "hot spot") could be enhanced more than 1000-fold when illuminated at 785 nm with a near-infrared diode laser. Interestingly, when 1 μM of rhodamine 6G was adsorbed onto the single gold nanocrescent moon it could be detected by a recognizable difference in the SERS spectrum. Such high sensitivity was considered due to the sharp edge of the nanostructure that had been created from the colloidal template.

Besides the exploitation of CL, and the patterns thus obtained via LSPR-assisted sensing, the magnetic properties of CL-derived nanostructures have continued to attract attention. In general, nanoscale magnetic materials often exhibit superparamagnetic behavior, and an ordered nanostructure of magnetic materials is required when conducting investigations into the mesoscopic effects induced when magnetic materials are confined within nanoscale domains [105]. As the magnetic properties depend heavily on the domain size and inter-domain distances, Weekes *et al.* set out to create ordered arrays of isolated magnetic nanodots via NSL [93]. In this case, the coercivity and switching width of the isolated nanodot arrays were enhanced when compared to those of continuous magnetic films. In

addition, well-organized arrays of magnetic nanorings over a large area have been prepared via NSL, and have demonstrated a stable vortex state due to the absence of a destabilizing vortex core. Such findings should hold promise for applications in vertical magnetic random access memories [106, 107]. Albrecht *et al.* have also shown that Co/Pd multilayers on a colloid surface exhibited a pronounced magnetic anisotropy (Figure 6.24) [108].

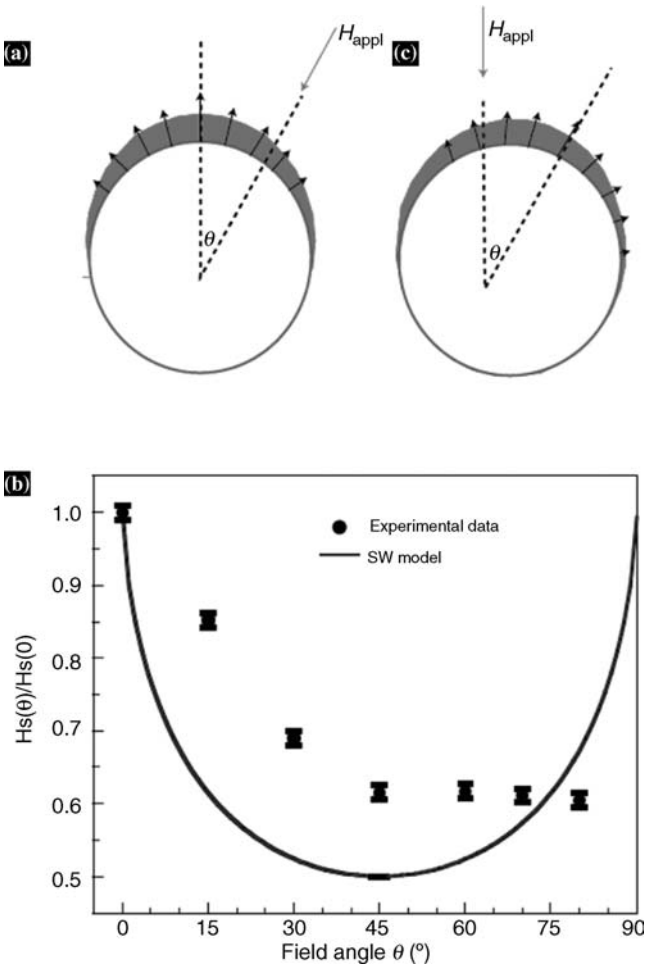


Figure 6.24 (a) Schematic of a magnetic film deposited on a nanosphere, showing the anisotropy distribution indicated by arrows; (b) Switching field as a function of applied field angle for an array with 50 nm particles (black dots). The angular dependence based on the Stoner–Wohlfarth (SW) model is shown as a solid line for comparison. The error bars are the standard deviation of the measurement; (c) The average anisotropy axis can be tuned to the required angle by changing the deposition direction. Reproduced with permission from Ref. [108].

6.6 Summary

The recent development of CL, and especially the integration of etching the colloidal mask, altering the incident angle, and the stepwise and regular changes in mask registry, have led to the creation of a powerful nanochemical patterning tool which is inexpensive (in terms of both capital and operation), has a high throughput, and can easily be adopted on various planar and curved surfaces, including microparticles. Unlike conventional mask-assisted lithographic processes, where the mask design and production tend to pose a challenge when scaling down the feature size and diversifying the feature shape, CL embodies a simple approach to masking, namely the self-assembly of monodisperse microspheres on a targeted substrate. As a consequence, the feature size may easily be shrunk below 100 nm, simply by reducing the diameter of the microspheres used, and according to a simple correlation between the interstice size and the sphere diameter. Likewise, the feature shape can be easily modified by changing the crystalline structure of a colloidal crystal mask, the time of anisotropic etching of the mask, the incident angle of the vapor beam, and the mask registry (the azimuth angle of the vapor beam). Currently, CL permits the fabrication of complicated 2-D and 3-D nanostructured features, such as multiplex nanostructures, with a clear-cut lateral and vertical heterogeneity. These several new nanostructures have proven difficult to implement by conventional lithographic techniques, and cannot be implemented in some cases. Hence, CL provides both a nanochemical and complementary tool for conventional and fully top-down lithographic techniques and, as a result, continues to show immense promise in the field of surface patterning.

Despite the great progress in colloidal crystallization, CL remains at a very early stage of development, and represents a formidable challenge for the creation of defect-free single crystals with defined crystalline faces. The presence of defects dramatically reduces the patterning precision of CL. For instance, the random orientation of polycrystalline crystalline domains in a colloidal mask is disastrous when collimating the mask registry, and in this respect the template-assisted epitaxy for colloidal crystallization shows great promise, as it allows the growth of colloidal crystals with defined packing structures and orientations. As a patterned substrate is a prerequisite for colloidal epitaxy, its applicability to patterning is limited. The ability to transfer a colloidal crystal, derived from such colloidal epitaxy, onto different substrates without causing any deterioration in crystal quality represents an important task for CL. The fabrication of large-area monolayers of periodically close-packed microspheres less than 100 nm in size is also a clear challenge that will involve reducing feature sizes to less than 10 nm, via the process of CL. Unfortunately, in a CL patterning process the feature size and interspace size between features cannot be separately manipulated, since both are directly proportional to the sphere size in a colloidal mask, and this greatly limits the patterning capabilities of CL.

Acknowledgments

D.W. was supported financially by the Max Planck Society, and in part by a DFG grant (WA 1704/4-1) and an EU-FP6 grant (BONSAI, LSHB-CT-2006-037639). G.Z. thanks the NSFC (No. 50703015) and MOST (No. 2007CB936402) for financial support. The authors also thank Dr. W. Li for his role in the preparation of the chapter, and Prof. G. Schmid for his organization of the present themed monograph with Wiley.

References

1 Wang, D. and Möhwald, H. (2004) Template-directed colloidal self-assembly – the route to 'top-down' nanochemical engineering. *J. Mater. Chem.*, **14**, 459–468.

2 Xia, Y., Gates, B., Yin, Y., and Lu, Y. (2000) Monodisperse colloidal particles: old materials with new applications. *Adv. Mater.*, **12**, 693–713.

3 Leunissen, M.E., Christova, C.G., Hynninen, A.-P., Royall, C.P., Campbell, A.I., Imhof, A., Dijkstra, M., van Roij, R., and van Blaaderen, A. (2005) Ionic colloidal crystals of oppositely charged particles. *Nature*, **437**, 235–240.

4 Mayoral, R., Requena, J., Moya, J.S., López, C., Cintas, A., Miguez, H., Meseguer, F., Vazquez, L., Holgado, M., and Blanco, A. (1997) 3D long-range ordering in and SiO_2 submicrometer-sphere sintered superstructure. *Adv. Mater.*, **9**, 257–260.

5 Miguez, H., Meseguer, F., López, C., Mifsud, A., Moya, J.S., and Vazquez, L. (1997) Evidence of FCC crystallization of SiO_2 nanospheres. *Langmuir*, **13**, 6009–6011.

6 Migurez, H., López, C., Meseguer, F., Blanco, A., Vazquez, L., Mayoral, R., Ocafia, M., Fornes, V., and Mifsud, A. (1997) Photonic crystal properties of packed submicrometric SiO_2. *Appl. Phys. Lett.*, **71**, 1148–1150.

7 Denkov, N.D., Velev, O.D., Kralchevsky, P.A., Invanov, I.B., Yoshimura, H., and Nagayama, K. (1992) Mechanism of formation of 2D crystals from latex-particles on substrates. *Langmuir*, **8**, 3183–3190.

8 Micheletto, R., Fukuda, H., and Ohtsu, M. (1995) A simple method for the production of a 2D ordered array of small latex-particles. *Langmuir*, **11**, 3333–3336.

9 Dimitrov, A.S. and Nagayama, K. (1996) Continuous convective assembling of fine particles into 2D arrays on solid surfaces. *Langmuir*, **12**, 1303–1311.

10 Jiang, P., Bertone, J.F., Hwang, K.S., and Colvin, V.L. (1999) Single-crystal colloidal multilayers of controlled thickness. *Chem. Mater.*, **11**, 2132–2140.

11 Gu, Z.–Z., Fujishima, A., and Sato, O. (2002) Fabrication of high-quality opal films with controllable thickness. *Chem. Mater.*, **14**, 760–765.

12 Zhou, Z. and Zhao, X.S. (2004) Flow-controlled vertical deposition method for the fabrication of photonic crystals. *Langmuir*, **20**, 1524–1526.

13 Wong, S., Kitaev, V., and Ozin, G.A. (2003) Colloidal crystals films: advances in universality and perfection. *J. Am. Chem. Soc.*, **125**, 15589–15598.

14 Chen, X., Chen, Z., Fu, N., Lu, G., and Yang, B. (2003) Versatile nanopatterned surfaces generated via 3D colloidal crystals. *Adv. Mater.*, **15**, 1413–1417.

15 Chung, Y.W., Leu, I.C., Lee, J.H., and Hon, M.H. (2006) Influence of humidity on the fabrication of high-quality colloidal crystals via a capillary-enhanced process. *Langmuir*, **22**, 6454–6460.

16 Kim, M.H., Im, S.H., and Park, O.O. (2005) Rapid fabrication of two- and three-dimensional colloidal crystal films via confined convective assembly. *Adv. Funct. Mater.*, **15**, 1329–1335.

17 Cheng, Z., Russel, W.B., and Chaikin, P.M. (1999) Controlled growth of hard-sphere colloidal crystals. *Nature*, **401**, 893–895.

18 Im, S.H., Kim, M.H., and Park, O.O. (2003) Thickness control of colloidal crystals with a substrate dipped at a tilted angle into a colloidal suspension. *Chem. Mater.*, **15**, 1797–1802.

19 Kitaev, V. and Ozin, G.A. (2003) Self-assembled surface patterns of binary colloidal crystals. *Adv. Mater.*, **15**, 75–78.

20 Velikov, K.P., Christova, C.G., Dullens, R.P.A., and van Blaaderen, A. (2002) Layer-by-layer growth of binary colloidal crystals. *Science*, **296**, 106–109.

21 Kim, M.H., Im, S.H., and Park, O.O. (2005) Fabrication and structural analysis of binary colloidal crystals with 2D superlattices. *Adv. Mater.*, **17**, 2501–2505.

22 Fischer, U.C. and Zingsheim, H.P. (1981) Submicroscopic pattern replication with visible light. *J. Vac. Sci. Technol.*, **19**, 881–885.

23 Hulteen, J.C. and van Duyne, R.P. (1995) Nanosphere lithography-a materials general fabrication process for periodic particle array surfaces. *J. Vac. Sci. Technol. A.*, **13**, 1553–1558.

24 Rehg, T.J. and Higgins, B.G. (1992) Spin coating of colloidal suspensions. *AIChE*, **38**, 489–501.

25 Jiang, P. and McFarland, M.J. (2004) Large-scale fabrication of wafer-size colloidal crystals, macroporous polymers and nanocomposites by spin-coating. *J. Am. Chem. Soc.*, **126**, 13778–13786.

26 Jiang, P. and McFarland, M.J. (2005) Wafer-scale periodic nanoholes arrays templated from 2D non-close-packed colloidal crystals. *J. Am. Chem. Soc.*, **127**, 3710–3711.

27 Wang, D. and Möhwald, H. (2004) Rapid fabrication of binary colloidal crystals by stepwise spin-coating. *Adv. Mater.*, **16**, 244–247.

28 Ulman, A. (1991) *An Introduction to Ultrathin Organic Films: FROM LANGMUIR-Blodgett to Self-Assembly*, Academic Press, Boston.

29 Binks, B.P. (2002) Particles as surfactants-similarities and differences. *Curr. Opin. Colloid Interface Sci.*, **7**, 21–41.

30 Pieranski, P. (1980) Two dimensional interfacial colloidal crystals. *Phys. Rev. Lett.*, **45**, 569–572.

31 Im, S.H., Lim, Y.T., Suh, D.J., and Park, O.O. (2002) 3D self-assembly of colloids at a water-air interface: A novel technique for the fabrication of photonic bandgap crystals. *Adv. Mater.*, **14**, 1367–1369.

32 Kondo, M., Shinozaki, K., Bergström, I., and Mizutani, N. (1995) Preparation of colloidal monolayers of alkoxylated silica particles at the air-liquid interface. *Langmuir*, **11**, 394–397.

33 Fulda, K.-U. and Tieke, B. (1994) Langmuir films of monodisperse 0.5 μm spherical polymer particles with a hydrophobic core and a hydrophilic shell. *Adv. Mater.*, **6**, 288–290.

34 van Duffel, B., Ras, R.H.A., De Schryver, F.C., and Schoonheydt, R.A. (2001) Langmuir-Blodgett deposition and optical diffraction of two dimensional opal. *J. Mater. Chem.*, **11**, 3333–3336.

35 Reculusa, S. and Ravaine, S. (2003) Synthesis of colloidal crystals of controllable thickness through the Langmuir-Blodgett technique. *Chem. Mater.*, **15**, 598–605.

36 Goldenberg, L.M., Wagner, J., Stumpe, J., Paulke, B.-R., and Grnitz, E. (2002) Simple method for the preparation of colloidal particle monolayers at the water/alkane interface. *Langmuir*, **18**, 5627–5629.

37 Reynaert, S., Moldenaers, P., and Vermant, J. (2006) Control over colloidal aggregation in monolayers of latex particles at the oil-water interface. *Langmuir*, **22**, 4936–4945.

38 Velikov, K.P., Durst, F., and Velev, O.D. (1998) Direct observation of the dynamics of latex particles confined inside thinning water-air films. *Langmuir*, **14**, 1148–1155.

39 Gu, Z.-Z., Wang, D., and Möhwald, H. (2007) Self-assembly of microspheres at the air/water/air interface into free-standing colloidal crystal films. *Soft Matter*, **3**, 68–70.

40 Griesebock, B., Egen, M., and Zental, R. (2002) Large photonic films by crystallization on fluid substrates. *Chem. Mater.*, **14**, 4023–4025.

41 Hanarp, P., Kall, M., and Sutherland, D.S. (2003) Optical properties of short range ordered arrays of nanometer gold disks prepared by colloidal lithography. *J. Phys. Chem. B*, **107**, 5768–5772.

42 Kosiorek, A., Kandulski, W., Glaczynska, H., and Giersig, M. (2005) Fabrication of nanoscale rings, dots, and rods by combining shadow nanosphere lithography and annealed polystyrene nanosphere masks. *Small*, **1**, 439–444.

43 Moon, J.H., Kim, W.-S., Ha, J.-W., Jang, S.G., Yang, S.-M., and Park, J.K. (2005) Colloidal lithography with crosslinkable particles: Fabrication of hierarchical nanopore arrays. *Chem. Commun.*, 4107–4109.

44 Peninkhof, J.J., Graf, C., van Dillen, T., Vredenberg, A.M., van Blaaderen, A., and Polman, A. (2005) Angle-dependent extinction of anisotropic silica/Au core/shell colloids made via ion irradiation. *Adv. Mater.*, **17**, 1484–1488.

45 Vossen, D.L.J., Fific, D., Penninkhof, J., Van Dillen, T., Polman, A., and Van Blaaderen, A. (2005) Combined optical tweezers/ion beam technique to tune colloidal masks for nanolithography. *Nano Lett.*, **5**, 1175–1179.

46 Deckmann, H.W. and Dunsmuir, J.H. (1983) Applications of surface textures produced with natural lithography. *J. Vac. Sci. Technol. B*, **1**, 1109–1112.

47 Choi, D.-G., Yu, H.K., Jang, S.G., and Yang, S.-M. (2004) Colloidal lithographic nanopatterning via reactive ion etching. *J. Am. Chem. Soc.*, **126**, 7019–7025.

48 Yang, S.-M., Jang, S.G., Choi, D.-G., Kim, S., and Yu, H.K. (2006) Nanomachining by colloidal lithography. *Small*, **2**, 458–475.

49 Choi, D.-G., Jang, S.G., Kim, S., Lee, E., Han, C.-S., and Yang, S.-M. (2006) Multifaceted and nanobored particle arrays sculpted using colloidal lithography. *Adv. Funct. Mater.*, **16**, 33–40.

50 Zheng, Y., Wang, Y., Wang, S., and Huan, C.H.A. (2006) Fabrication of nonspherical colloidal particles via reactive ion etching of surface-patterned colloidal crystals. *Colloids Surf. A*, **277**, 27–36.

51 Cheung, C.L., Nikolić, R.J., Reinhardt, C.E., and Wang, T.F. (2006) Fabrication of nanopillars by nanosphere lithography. *Nanotechnology*, **17**, 1339–1343.

52 Tan, B.J.-Y., Sow, C.-H., Lim, K.-Y., Cheong, F.-C., Chong, G.-L., Wee, A.T.-S., and Ong, C.-K. (2004) Fabrication of a two-dimensional periodic non-close-packed array of polystyrene particles. *J. Phys. Chem. B*, **108**, 18575–18579.

53 Valsesia, A., Colpo, P., Silvan, M.M., Meziani, T., Ceccone, G., and Rossi, F. (2004) Fabrication of nanostructured polymeric surfaces for biosensing devices. *Nano Lett.*, **4**, 1047–1050.

54 Wang, B., Zhao, W., Chen, A., and Chua, S.-J. (2006) Formation of nanoimprinting mould through use of nanosphere lithography. *J. Crystal Growth*, **288**, 200–204.

55 Huang, Z., Fang, H., and Zhu, J. (2007) Fabrication of silicon nanowire arrays with controlled diameter, length, and density. *Adv. Mater.*, **19**, 744–748.

56 Tan, B.J.Y., Sow, C.H., Koh, T.S., Chin, K.C., Wee, A.T.S., and Ong, C.K. (2005) Fabrication of size-tunable gold nanoparticles array with nanosphere lithography, reactive ion etching, and thermal annealing. *J. Phys. Chem. B*, **109**, 11100–11109.

57 McLellan, J.M., Geissler, M., and Xia, Y. (2004) Edge spreading lithography and its application to the fabrication of mesoscopic gold and silver rings. *J. Am. Chem. Soc.*, **126**, 10830–10831.

58 Geissler, M., McLellan, J.M., Chen, J., and Xia, Y. (2005) Side-by-side patterning of multiple alkanethiolate monolayers on gold by edge-spreading lithography. *Angew. Chem., Int. Ed.*, **44**, 3596–3600.

59 Bae, C., Shin, H., Moon, J., and Sung, M.M. (2006) Contact area lithography (CAL): A new approach to direct formation of nanometric chemical patterns. *Chem. Mater.*, **18**, 1085–1088.

60 Bae, C., Moon, J., Shin, H., Kim, J., and Sung, M.M. (2007) Fabrication of monodisperse asymmetric colloidal clusters by using contact area lithography (CAL). *J. Am. Chem. Soc.*, **129**, 14232–14239.

61 Deckmann, H.W. and Dunsmuir, J.H. (1982) Natural lithography. *Appl. Phys. Lett.*, **41**, 377–379.

62 Hulteen, J.C., Treichel, D.A., Smith, M.T., Duval, M.L. Jensen, T.R., and Van Duyene, R.P. (1999) Nanosphere lithography: size-tunable silver nanoparticle and surface cluster arrays. *J. Phys. Chem. B*, **103**, 3854–3863.

63 Haynes, C.L. and Van Duyne, R.P. (2001) Nanosphere lithography: A versatile nanofabrication tool for studies of size-dependent nanoparticle optics. *J. Phys. Chem. B*, **105**, 5599–5611.

64 Haynes, C.L., McFarland, A.D., Smith, M.T., Hulteen, J.C., and Van Duyne, R.P. (2002) Angle-resolved nanosphere lithography: Manipulation of nanoparticle size, shape, and interparticle spacing. *J. Phys. Chem. B*, **106**, 1898–1902.

65 Willets, A. and Van Duyne, R.P. (2007) Localized surface plasmon resonance spectroscopy and tension. *Annu. Rev. Phys. Chem.*, **58**, 267–297.

66 Giersig, M. and Hilgendorff, M. (2005) Magnetic nanoparticle superstructures. *Eur. J. Inorg. Chem.*, 3571–3583.

67 Kosiorek, A., Kandulski, W., Chudzinski, P., Kempa, K., and Giersig, M. (2004) Shadow nanosphere lithography: Simulation and experiment. *Nano Lett.*, **4**, 1359–1363.

68 Zhang, G. and Wang, D. (2008) Fabrication of heterogeneous binary arrays of nanoparticles via colloidal lithography. *J. Am. Chem. Soc.*, **130**, 5616–5617.

69 Weekes, S.M. and Ogrin, F.Y. (2005) Torque studies of large-area Co arrays fabricated by etched nanosphere lithography. *J. Appl. Phys.*, **97** (10), J503.

70 Choi, D.-G., Kim, S., Jang, S.G., Yang, S.-M., Jeong, J.-R., and Shin, S.-C. (2004) Nanopatterned magnetic metal via colloidal lithography with reactive ion etching. *Chem. Mater.*, **16**, 4208–4211.

71 Zhang, G., Wang, D., and Möhwald, H. (2007) Fabrication of multiplex quasi-three-dimensional grids of one-dimensional nanostructures via stepwise colloidal lithography. *Nano Lett.*, **7**, 3410–3413.

72 Zhang, G., Wang, D., and Möhwald, H. (2007) Ordered binary arrays of Au nanoparticles derived from colloidal lithography. *Nano Lett.*, **7**, 127–132.

73 Habenicht, A., Olapinski, M., Burmeister, F., Leiderer, P., and Boneberg, J. (2005) Jumping nanodroplets. *Science*, **309**, 2043–2045.

74 Fredriksson, H., Alaverdyan, Y., Dmitriev, A., Langhammer, C., Sutherland, D.S., Zäch, M., and Kasemo, B. (2007) Hole-mask colloidal lithography. *Adv. Mater.*, **19**, 4297–4302.

75 Edwards, E.W., Wang, D., and Möhwald, H. (2007) Hierarchical organization of colloidal particles: from colloidal crystallization to supraparticle chemistry. *Macromol. Chem. Phys.*, **208**, 439–445.

76 Zhang, H., Edwards, E.W., Wang, D., and Möhwald, H. (2006) Directing the self-assembly of nanocrystals beyond colloidal crystallization. *Phys. Chem. Chem. Phys.*, **8**, 3288–3299.

77 Glotzer, S.C. and Solomon, M.J. (2007) Anisotropy of building blocks and their assembly into complex structures. *Nat. Mater.*, **6**, 557–562.

78 Fujimoto, K., Nakahama, K., Shidara, M., and Kawaguchi, H. (1999) Preparation of unsymmetrical microspheres at the interfaces. *Langmuir*, **15**, 4630–4635.

79 Lu, Y., Xiong, H., Jiang, X., Xia, Y., Prentiss, M., and Whitesides, G.M. (2003) Asymmetric dimers can be formed by dewetting half-shells of gold deposited on the surfaces of spherical oxide colloids. *J. Am. Chem. Soc.*, **125**, 12724–12725.

80 Bao, Z., Chen, L., Weldon, M., Chandross, E., Cherniavskaya, O., Dai, Y., and Tok, J. (2002) Toward controllable self-assembly of microstructures: Selective functionalization and fabrication of patterned spheres. *Chem. Mater.*, **14**, 24–26.

81 Perro, A., Reculusa, S., Ravaine, S., Bourgeat-Lami, E., and Duguet, E. (2005) Design and synthesis of Janus micro- and nanoparticles. *J. Mater. Chem.*, **15**, 3745–3760.

82 Hong, L., Cacciuto, A., Luijten, E., and Granick, S. (2006) Clusters of charged Janus spheres. *Nano Lett.*, **6**, 2510–2514.

83 Zhang, G., Wang, D., and Möhwald, H. (2005) Patterning microsphere surfaces by templating colloidal crystals. *Nano Lett.*, **5**, 143–146.

84 Zhang, G., Wang, D., and Möhwald, H. (2006) Nanoembossment of Au patterns on microspheres. *Chem. Mater.*, **18**, 3985–3992.

85 Pawar, A.B. and Kretzschmar, I. (2008) Patchy particles by glancing angle deposition. *Langmuir*, **24**, 355–358.

86 Zhang, G., Wang, D., and Möhwald, H. (2005) Decoration of microspheres with gold nanodots—giving colloidal spheres valences. *Angew. Chem., Int. Ed.*, **44**, 7767–7770.

87 Nelson, D.R. (2002) Toward a tetravalent chemistry of colloids. *Nano Lett.*, **2**, 1125–1129.

88 Pacifico, J., Gómez, D., and Mulvaney, P. (2005) A simple route to tunable two-dimensional arrays of quantum dots. *Adv. Mater.*, **17**, 415–418.

89 Pacifico, J., Jasieniak, J., Gómez, D.E., and Mulvaney, P. (2006) Tunable D-3 arrays of quantum dots: Synthesis and luminescence properties. *Small*, **2**, 199–203.

90 Sutherland, D.S., Broberg, M., Nygren, H., and Kasemo, B. (2001) Influence of nanoscale surface topography and chemistry on the functional behaviour of an adsorbed model macromolecule. *Macromol. Biosci.*, **1**, 270–273.

91 Kuo, C.-W., Shiu, J.-Y., and Chen, P. (2003) Size- and shape-controlled fabrication of large-area periodic nanopillar arrays. *Chem. Mater.*, **15**, 2917–2920.

92 Kuo, C.-W., Shiu, J.-Y., Cho, Y.-H., and Chen, P. (2003) Fabrication of large-area periodic nanopillar arrays for nanoimprint lithography using polymer colloid masks. *Adv. Mater.*, **15**, 1065–1068.

93 Weekes, S.M., Ogrin, F.Y., and Murray, W.A. (2004) Fabrication of large-area ferromagnetic arrays using etched nanosphere lithography. *Langmuir*, **20**, 11208–11212.

94 Han, S., Hao, Z., Wang, J., and Luo, Y. (2005) Controllable two-dimensional photonic crystal patterns fabricated by nanosphere lithography. *J. Vac. Sci. Technol. B*, **23**, 1585–1588.

95 Ryu, K., Badmaev, A., Gomez, L., Ishikawa, F., Lei, B., and Zhou, C. (2007) Synthesis of aligned single-walled nanotubes using catalysts defined by nanosphere lithography. *J. Am. Chem. Soc.*, **129**, 10104–10105.

96 Park, K.H., Lee, S., Koh, K.H., Lacerda, R., Teo, K.B.K., and Milne, W.I. (2005) Advanced nanosphere lithography for the areal-density variation of periodic arrays of vertically aligned carbon nanofibers. *J. Appl. Phys.*, **97**, 024311.

97 Wang, X., Summers, C.J., and Wang, Z.L. (2004) Large-scale hexagonal-patterned growth of aligned ZnO nanorods for nano-optoelectronics and nanosensor arrays. *Nano Lett.*, **4**, 423–426.

98 Fuhrmann, B., Leipner, H.S., Höche, H.-R., Schubert, L., Werner, P., and Gösele, U. (2005) Ordered arrays of silicon nanowires produced by nanosphere lithography and molecular beam epitaxy. *Nano Lett.*, **5**, 2524–2537.

99 Veinot, J.G.C., Yan, H., Smith, S.M., Cui, J., Huang, Q., and Marks, T.J. (2002) Fabrication and properties of organic light-emitting "nanodiode" arrays. *Nano Lett.*, **2**, 333–335.

100 Aizpurua, J., Hanarp, P., Sutherland, D.S., Kall, M., Bryant, G.W., and Garcia de Abajo, F.J. (2003) Optical properties of gold nanorings. *Phys. Rev. Lett.*, **90**, 057401.

101 Hanarp, P., Kall, M., and Sutherland, D.S. (2003) Optical properties of short range ordered arrays of nanometer gold disks prepared by colloidal lithography. *J. Phys. Chem. B*, **107**, 5768–5772.

102 Lamprecht, B., Schider, G., Lechner, R.T., Ditlbacher, H., Krenn, J.R., Leitner, A., and Aussenegg, F.R. (2000) Metal nanoparticle gratings: Influence of dipolar particle interaction on the plasmon resonance. *Phys. Rev. Lett.*, **84**, 4721–4724.

103 Liu, J., Maaroof, A.I., Wieczorek, L., and Cortie, M.B. (2005) Fabrication of hollow metal "nanocaps" and their red-shifted optical absorption spectra. *Adv. Mater.*, **17**, 1276–1281.

104 Lu, Y., Liu, G.L., Kim, J., Mejia, Y.X., and Lee, L.P. (2005) Nanophotonic crescent moon structures with sharp edge for ultrasensitive biomolecular detection by local electromagnetic field enhancement effect. *Nano Lett.*, **5**, 119–124.

105 Moser, A., Takano, K., Margulis, D.T., Albrecht, M., Sonobe, Y., Ikeda, Y., Sun, S., and Fullerton, E.E. (2002) Magnetic recording: advancing into the future. *J. Phys. D*, **35**, R157–167.

106 Zhu, J., Zheng, Y., and Prinz, G.A. (2000) Ultrahigh density vertical magnetoresistive random access memory. *J. Appl. Phys.*, **87**, 6668–6673.

107 Zhu, F., Fan, D., Zhu, X., Zhu, J.-G., Cammarata, R.C., and Chien, C.-L. (2004) Ultrahigh-density arrays of ferromagnetic nanorings on macroscopic areas. *Adv. Mater.*, **16**, 2155–2159.

108 Albrecht, M., Hu, G., Guhr, I.L., Ulbrich, T.C., Boneberg, J., Leiderer, P., and Schatz, G. (2005) Magnetic multilayers on nanospheres. *Nat. Mater.*, **4**, 203–206.

7
Diblock Copolymer Micelle Nanolithography: Characteristics and Applications

Theobald Lohmueller and Joachim P. Spatz

7.1
Introduction

Materials at the nanoscale show remarkable physical and chemical properties as a consequence of their small size, and this makes them useful for a wide range of conceivable new applications and technologies [1]. In 1959, Richard P. Feynman announced the issue for the future of ". . .manipulating and controlling things on a small scale" [2]. Since that time, nanotechnology has indeed evolved to become an important topic with enormous scientific and industrial interest [3], and this has resulted in increasing progress in terms of the development of novel micro- and optoelectronic components, with a subsequent high impact on everyday life. The most famous, yet impressive, example of this rapid evolution was expressed by Moore's law, which stated that the number of components per integrated circuit on a microprocessor would double approximately every two years [4]. Although dating back to the 1970s and aimed at predicting the increase of processing power and data storage capacity in computer technology, this forecast remains valid today, and greatly emphasizes the current vitality of this field of research. Today, nanotechnology enfolds a broad interdisciplinary area in the applied sciences, and is used for the investigation of various fundamental questions in physics, chemistry, and biology [1,3,5].

Two strategies for nanofabrication can be identified, namely "top-down" and "bottom-up":

- Top-down methods comprise photo- [6, 7], X-ray [8, 9], as well as electron (e-beam) [10] and focused ion beam (FIB) [11, 12] lithography, where nanostructured surfaces are generated either by exposing a sensitive material to UV-light or X-ray radiation, or by "writing" a nanopattern with a focused beam of electrons or ions [13]. For patterning, a sample is covered with a photosensitive resist and illuminated through a mask, in close proximity to the substrate. The final structure is developed by subsequent treatment of the substrate with a selective solvent or etching agent; this enables surface patterning with high resolution, depending on the wavelength of the irradiation. Consequently,

Nanotechnology, Volume 8: Nanostructured Surfaces. Edited by Lifeng Chi

ISBN: 978-3-527-31739-4

photolithography is currently the most widely used technology in the semiconductor industry where, by using state-of-the-art systems with deep-ultraviolet (DUV) excimer lasers (e.g., ArF: 193 nm) it is possible to pattern structures smaller than 50 nm [14, 15]. No mask is needed in the case of e-beam and FIB lithography, since the pattern is generated directly by a focused beam of electrons and ions, rather than by irradiating the whole sample at once. Although, when using these methods it is possible to achieve resolutions down to only a few nanometers, because they represent a serial process they have the main disadvantages of low processing rates and the need for expensive equipment.

- Bottom-up approaches are the competing concept for nanofabrication, with the idea of the self-organization of small components into larger materials and devices, without any external intervention [16–18]. As a concept for the materials sciences, self-assembly based methods are particularly attractive as they represent an inexpensive and widely applicable fabrication technology, with possible resolution down to a single nanometer. Several strategies have been developed and shown to be capable for patterning of materials at the nanoscale; these include self-assembled monolayers (SAMs) [19, 20], block copolymer lithography [21–23], and colloidal lithography [24, 25], all of which provide cheap and fast processing technologies that produce nanometer-scale resolution. Whilst the applicability of the bottom-up methods is limited by a greater demand on the complexity of the nanopattern, strategies to overcome these drawbacks include placing the molecules directly into an aperiodic topology, as is the case for microcontact printing (μCP) [18, 26, 27] and dip-pen lithography (DPN) [28, 29]. Here, the desired molecules are used as an "ink" which can either be layered (in DPN) or stamped (in μCP) on top of a solid-state substrate. The SAM itself can then be used as a lithographic resist for further modification. Examples of this approach include chemical lithography and scanning probe lithography-based technologies, such as nanografting [30, 31] and near-field scanning optical lithography (NSOM) [32].

The aim of this chapter is to introduce block copolymer micelle nanolithography (BCML) as a versatile bottom-up approach, to generate extended arrays of metallic nanoparticles, and to provide some examples of its related applications [33, 34]. Block copolymers are compounds of separate polymer chains (or blocks respectively), where each block is built up from individual types of monomer. In the case of diblock copolymers, for instance, one molecule contains two polymer chains (or blocks), which are linked by a covalent bond. For entropic reasons, the two blocks do not mix in solution, and thus show a strong tendency to segregate into microphase-separated morphologies, depending on the molecular weight, the segment sizes, and the strength of the molecular interactions between the respective blocks [35, 36]. How structural parameters such as the lateral particle spacing and particle size can be controlled on the substrate, with single-nanometer precision, will be examined in the following subsections. Two intriguing examples will also be provided of how these nanoparticle patterns can be used to mimic materials found in nature, such as the

antireflective corneal lens of moths, and artificial extracellular interfaces in the study of cell adhesion.

7.2
Block Copolymer Micelle Nanolithography

7.2.1
Introduction

The underlying principle of BCML relates to the spontaneous formation of amphiphilic block-copolymers to from microphase-separated morphologies [35], and the transfer of these units into nanometer-scale patterns on top of rigid substrates such as silicon wafers or glass coverslips. When diblock copolymers of polystyrene (PS) and polyvinylpyridine (PVP) chains are dissolved in toluene at low concentration, they are present in the form of single chains. However, above their critical micelle concentration (CMC), these molecules begin to segregate while the number of individual free chains in solution remains constant [37–39]. As toluene is a more selective solvent towards PS, the PS block will form the outer micellar shell, surrounding the less-soluble PVP block which builds up the core [40]. This core–shell configuration can be considered as a nanoscopic reactor that allows the selective dissolution of metal precursor salts into the micelle [41]. The distribution of precursor salt per micelle varies only within narrow limits [42], and the incorporation of metal salts into the micellar core has certain effects on the micellar stability of the micelles in solution, compared to the neutral case. Due to increasing interactions between the different copolymer blocks, the CMC shifts towards lower concentrations [43, 44]. The amount of metal salt inside each micelle is adjustable, depending on the ratio of the neutralized number of vinylpyridine versus the total number of vinylpyridine units:

$$m_{metal} = \frac{m_{PS-P_2VP} M_{metal} [VP]_n L}{M_{PS-P_2VP}}$$

where L depicts the loading ratio $L = n(\mathrm{HAuCl_4})/n(\text{PS-PVP})$ and $1 \geq L \geq 0$. A nanopattern is formed by either spin-coating or dipping a rigid substrate into the polymer solution. Dip-coating is more favorable as it enables a uniform decoration of plain as well as curved substrates over a total area of up to several square centimeters, within a short period of time, and with high accuracy. During substrate retraction, the micelles assemble into a quasi-hexagonal ordered monolayer on top of the surface, the driving force for which process is the evaporation of solvent at the immersion edge. Subsequent plasma treatment can then be applied to remove the whole polymer matrix and to induce the formation of pure metal particles on the surface of the substrate. A schematic overview of the process, together with relevant scanning electron microscopy (SEM) images of the different nanoparticle arrays, are shown in Figure 7.1.

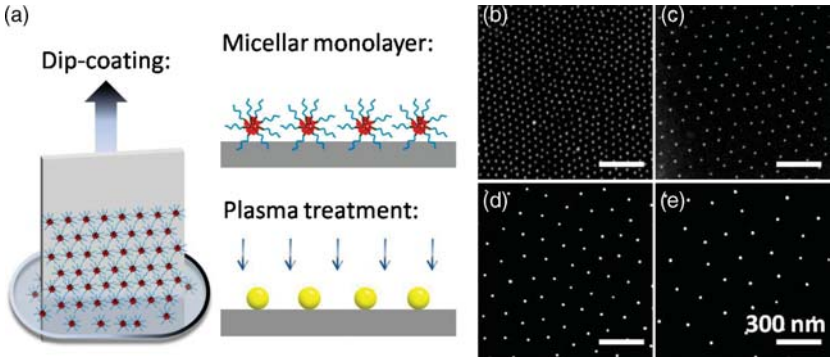


Figure 7.1 Schematic description of the dip-coating process. (a) Diblock copolymer micelles self-assemble into a hexagonal-ordered film when a rigid substrate is dragged out of the toluene solution. Gold nanoparticles are generated by subsequent plasma treatment of the substrate. The distance between individual particles on top of the substrate is a result of the diameter of the micelles, the concentration of the micelle solution, and the dipping velocity. Examples of gold nanoparticle pattern with different spacing: (b) 50 nm; (c) 100 nm; (d) 150 nm; (e) 200 nm.

The particles formed from one polymer solution during the process are all of similar size and shape, and aligned hexagonally on the surface (Figure 7.1b-e). The only major requirement arising from the fabrication process is that the substrate must be resistant against the solvent, and stable during the plasma process. The technique thus offers a great applicability, and has been successfully applied to pattern different materials such as glass, silicon, diamond, sapphire, $SrTiO_3$, and mica, with various different particle compositions. A literature overview dealing with the concept of BCML is provided in Table 7.1.

Table 7.1 A literature overview of block copolymer micelle nanolithography.

Parameter	Component/technique	Reference(s)
Metallic NP	• Au, Pt, Pd	[33, 45, 107]
	• TiO_2	[108]
	• Fe_xPt_y	[109]
Particle size	• Intramicellar electroless deposition	[107]
Particle spacing	• Concentration and velocity dependence	[34, 61, 62]
	• Spacing gradients	[63]
Lateral order		[110–112]
Micro- nanopatterning	• E-beam lithography	[71, 72]
	• Photoresist e-beam; Photolithography	[70, 74]
	• FIB	[73]
	• Micro-contact printing	[113]
Transfer lithography	• PDMS, Polystyrene, PEGDA 700	[106]

The great flexibility of BCML makes it an ideal tool for nanopatterning, with the advantage of high-throughput sample processing but only a minimal need for highly technical (and expensive) equipment [45]. A more detailed description as to how nanoparticle patterns can be characterized, and how experimental parameters such as the lateral order, the interparticle distance, and nanoparticle spacing can be controlled in a large frame, are outlined in the following subsections. A brief overview is also provided on the synthesis of micro-nanopatterned surfaces, using a combination of BCML and conventional top-down lithography.

7.2.2 Characterization on Nanoparticle Arrays

7.2.2.1 Scanning Electron Microscopy (SEM)

Scanning electron microscopy (SEM) represents a powerful tool for sample imaging with nanoscale spatial resolution [46]. During such measurements, the surface of a conductive sample is imaged by scanning the surface topology with a focused electron beam. Nonconductive substrates such as glass or polymeric materials must be covered with a metal or carbon layer prior measurements. Although state-of-the-art SEM systems show a resolution power of less than 1 nm, the theoretical limit is impaired by aberrations of the electron lenses and the interaction volume of the electron beam with the substrate material. Atomic-level resolution can be achieved by using transmission electron microscopy (TEM), in which case the detector is located beneath the sample, such that those electrons transmitted *through* the sample are analyzed. Hence, very thin sample materials are required for these measurements. Consequently, although TEM imaging is currently recognized as the most accurate method for characterizing the size and shape of individual particles, even with atomic resolution, is not practical consider nanoparticle arrays fabricated on large area samples such as glass coverslips or whole silica wafers. Several types of signal are generated by interactions between the electron beam and the probe material, and this provides both topographic and chemical information concerning the sample. The most important of these signals are secondary electrons (SEs), back-scattered electrons (BSEs) and low-energy X-rays. Whereas, secondary and back-scattered electrons provide information concerning surface morphology and material contrast, the chemical composition of the specimen can be analyzed via its characteristic X-ray emission, using energy-dispersive X-ray (EDX) spectroscopy. SEM imaging is particularly advantageous for the characterization of extended nanoparticle arrays, since it is possible to carry out rapid imaging of large areas at several positions on the sample, so as to reveal accurate information concerning the separation distances between individual particles and the particle diameter in-plane. Unfortunately, SEM images provide very little information relating to the nanoparticle height.

7.2.2.2 Atomic Force Microscopy (AFM)

Atomic force microscopy (AFM) is a high-resolution imaging technique used to measure the attractive or repulsive forces between a sharp tip brought into proximity to the surface of a sample [47]. The tips used for AFM are typically fabricated from

silicon or silicon nitride, and have a diameter of between 10 and 50 nm. The available spring constants range from 0.01 to 100 N m^{-1}, with resonant frequencies between 5 and 350 kHz.

When taking measurements, the microscope is used to mechanically scan a certain area of the sample, and is operated in either *contact mode* or *tapping mode*:

- While operating in *contact mode*, the AFM tip is actually in contact with the surface, so that any topographic features encountered will cause the tip to undergo a vertical deflection; these deflections are translated into a feedback signal that carries information about the surface. The tip deflection is monitored by following a laser spot that is reflected from the top of the cantilever, using a photodiode.
- In *tapping mode*, the cantilever oscillates close to its resonance frequency, so that it is barely touching the substrate surface. With increasing tip-to-sample proximity the oscillation amplitude and phase, as well as the resonance frequency of the cantilever are damped, and the damping movements then provide information regarding the surface morphology. Tapping mode is commonly used to measure soft samples, where friction and lateral forces should be minimized.

In general, AFM displays several advantages over SEM, the most notable being that there are virtually no limitations regarding the substrate composition. The samples do not need to be conductive, and the measurements can be performed under atmospheric conditions. The image of the sample also represents a true three-dimensional profile of the surface, with a resolution that is comparable to that of SEM. The main drawbacks of AFM are a high sensitivity to environmental noise, and a low scanning size of approximately 100 μm^2. One other problem is that a convolution of the tip size with the surface topology can create image artifacts in-plane (xy). Thus, in order to provide a complete characterization of the height and spatial orientation of a nanopatterned substrate, a combination of both AFM and SEM should ideally be utilized [48].

7.2.2.3 Spacing and Order of Nanoparticle Arrays

The lateral geometry of a nanoparticle array can be considered to be in a state between crystalline or completely random [49]. When perfectly ordered, crystalline systems are defined by the long range order of their spatial pattern. In case of a 2-D hexagonal lattice, each particle is surrounded by an infinite number of equally distributed neighbors with a constant spacing period. In contrast, a random, amorphous system does not show any long-range order, and all of the pattern features are uniformly distributed with no specific positional information. The perfect pattern geometry is usually affected by defects and deformations, notably dislocations and disclinations as a consequence of the self-assembly process. Disclinations occur if a particle is surrounded for example by five or seven instead of six nearest-neighbors, causing a dislocation of the hexagonal orientation. A model system which is used to identify the quality of a nanoparticle array from an SEM image, compares the structure with the order transition during the melting of a 2-D crystal [49, 50]. A theoretical approach to describe such 2-D melting is known as the Kosterlitz–Thouless–Halperin–Nelson–Young (KTHNY) theory [51–54], where a perfect crystalline phase is disrupted with increasing temperature until the

state of complete disorder is reached. Following this model, the translational order of any symmetric lattice during the phase transitions can be described by an order parameter $\psi_{\vec{G}}(\vec{r})$ of the form [55]:

$$\psi_{\vec{G}}(\vec{r}) = \exp(i\vec{G}\vec{r})$$

where $\vec{G}$ is the reciprocal lattice vector of the hexagonal array. For both, disclinations and dislocations, the crystalline order is disrupted and $\psi_{\vec{G}}(\vec{r})$ falls to zero. However, there is a qualitative difference when considering the local-range and long-range order of a crystalline structure. Whereas dislocations cause a translational displacement that does not affect the local hexatic orientation, both translational and long-range orientational orders are impaired by disclinations. Thus, in order to quantify the bond-to-bond orientation of a whole hexagonal lattice, a complement to $\psi_{\vec{G}}(\vec{r})$, the global bond-orientational order parameter $\psi_6(\vec{r}_i)$ must be introduced:

$$\psi_6(\vec{r}_i) = \left| \frac{1}{N} \sum_j \sum_k \exp(6i\theta_{jk}) \right|$$

where θ_{ik} is the orientation angle of the connecting sides (or neighboring particles) j and k relative to $(\vec{r}_i)$. The influence of dislocations and disclinations can be distinguished by their influence on $\psi_{\vec{G}}(\vec{r})$ and $\psi_6(\vec{r}_i)$. For an array of nanoparticles, ψ_6 reaches a value between one and zero:

$$0 \leq \psi_6 \leq 1$$

where ψ_6 equals 1 for a perfect hexagonal lattice, but falls to zero with increasing disorder. By calculating the sixfold global order parameter, the average nanoparticle spacing and the corresponding standard error can be derived from all individual particles from a single SEM or AFM image.

7.2.3 Tuning the Pattern Properties

7.2.3.1 Controlling the Nanoparticle Spacing

The precise adjustment of the nanoparticle spacing represents one of the most important issues to realize a broad applicability for BCML. Several experimental parameters may influence the particle separation on the surface. A rather small effect is observed by the amount of metal salt added to the solution. The diameter of the micelles is increased depending on the metal salt loading, and this results in a lowered packing density of the micellar monolayer on the substrate. As a consequence, the hydrodynamic radius of the micelles is also increased, and this is reflected by a lower packing density following transfer to the substrate [43]. The spacing can be adjusted within a few nanometers, but only at the expense of the particle size, which is generally not desirable.

The most obvious way to achieve reproducible and robust control over a distance between single particles is to alter the length of the diblock-copolymer chain, since a longer polymer will result in a bigger micelle [34, 56, 57]. A broad range of lateral

Table 7.2 Examples of spacing values for different diblock copolymer solutions.

Diblock copolymer	*c* (mg ml^{-1})	*L*	MnPS – (g mol^{-1})	MnPVP – (g mol^{-1})	Mw/Mn	Spacing (nm)	Order (△)
PS(190)-*b*-P_2VP(190)	5	0.2	19 900	21 000	1.09	28 ± 5	0.54
PS(500)-*b*-P_2VP(270)	5	0.5	52 400	28 100	1.05	58 ± 7	0.50
PS(990)-*b*-P_2VP(385)	5	0.5	103 000	40 500	1.07	73 ± 8	0.39
PS(1350)-*b*-P_2VP(400)	5	0.5	140 800	41 500	1.11	85 ± 9	0.42
PS(1824)-*b*-P_2VP(523)	3	0.5	190 000	55 000	1.10	91 ± 1	0.63
PS(990)-*b*-P_2VP(385)	3	0.5	103 000	40 500	1.07	110 ± 1	0.61
PS(1824)-*b*-P_2VP(523)	2	0.5	190 000	55 000	1.10	139 ± 2	0.71
PS(5348)-*b*-P_4VP(713)	1.5	0.2	557 000	75 000	1.07	176 ± 5	0.40
PS(5348)-*b*-P_4VP(713)	1	0.2	557 000	75 000	1.07	222 ± 7	0.50

distances can be realized by simply varying the molecular weight, as shown in Table 7.2 (the dipping velocity in each case was 12 mm min^{-1}).

For a certain micelle size, it has been shown experimentally that the separation distance between the nanoparticles is also influenced by the concentration of the polymer solution (cf. Table 7.2). Although for higher concentrations the micelles are packed more closely, interestingly the same results are observed for different velocities during the dip-coating process. This is a particularly important finding, since varying either the polymer concentration or the retraction velocity will allow the creation of a broad range of different nanoparticle densities from a single polymer solution, and thus a complete decoupling of the metal salt loading and particle size. By accelerating the velocity during retraction of the substrate, it might even be possible to generate a continuous nanoparticle gradient over the range of several tens of nanometers. The gradient slope can be adjusted between a few micrometers and several millimeters, depending on the acceleration. Consequently, nanoparticle gradients with separation distances of between 80 and 250 nm can be achieved by using only three different polymer solutions, as shown in Figure 7.2.

As depicted in the graphs of Figure 7.2, the interparticle spacing on the samples is smaller for higher retraction velocities. To explain this observation, it is necessary to consider the influence of the retraction velocity of the deposition process during dip-coating. As reported by Darhuber *et al.* [58], the height of the adsorbed film deposited during the dip-coating of a substrate perpendicular to the fluid interface depends on the retraction speed at which the sample is withdrawn from the solution. The dependency between velocity and film thickness can be expressed as [58, 59]:

$$h_\infty = 0.946\sqrt{\frac{\sigma}{\varrho g}}Ca^{2/3}$$

where *Ca* denotes the capillary number $Ca = \mu U/\sigma$, and where μ, σ, and ϱ represent the solution viscosity, the surface tension, and the density of the polymer solution respectively; *g* denotes gravitational acceleration. This expression is only valid for

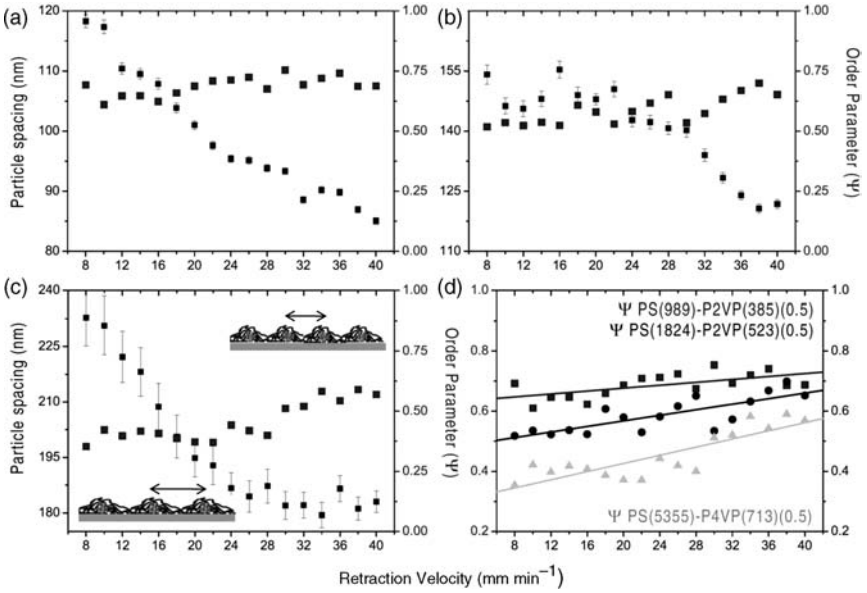


Figure 7.2 Particle spacing realized from three different diblock copolymer solutions as a function of the retraction speed. The retraction velocity is adjusted between 8 and 40 mm min^{-1} in all cases. (a) 3 mg ml^{-1} PS(990)-*b*-P2VP(385) (0.5); (b) 2 mg ml^{-1} PS(1824)-*b*-P2VP(523) (0.5); (c) 1 mg ml^{-1} PS(5348)-*b*-P4VP(713) (0.5). For high retraction velocities above 30 mm min^{-1}, an almost constant value of the interparticle distance was observed for the PS (5348)-*b*-P4VP(713) (0.5) polymer; (d) Linear fit of the corresponding order parameters for all three polymers as a function of the dipping velocity. A trend towards a higher order parameter was observed for smaller interparticle spacings.

very low capillary numbers ($Ca << 1$), which can be estimated in the present system (due to the very low capillary number for pure toluene [60, 62]. According to this equation, the film thickness h_∞ is dependent on the viscosity of the polymer solution and the dipping velocity. With increasing polymer concentration, the viscosity of the solution will therefore also increase. The optimum thickness h_∞ of the micellar film for a constant dipping velocity can thus be expressed by:

$$h_\infty \propto \frac{\mu^{2/3}}{(\varrho g)^{1/2}\sigma^{1/6}}$$

At the same time, assuming a constant number of micelles in solution for a constant polymer concentration, h_∞ is proportional to the dipping velocity U according to:

$$h_\infty \propto U^{2/3}$$

As the film thickness is proportional to the dipping velocity, the number of micelles per area is also proportional to h_∞. As the micelles are all of the same size, the maximum number on the surface will reach a saturation value for high dipping

speeds, where the interparticle spacing is almost constant and the nanopattern displays the highest packing density at this point. This is shown in Figure 7.2c, where the interparticle spacing of the PS(5348) polymer is unchanged between 32 and 40 mm min^{-1}. It is important to note that the order parameter is affected depending on the separation distance and the corresponding retraction velocity, and this is especially the case for polymers with a high molecular weight. As the viscosity of the solution increases with higher polymer concentrations, both parameters will influence the dipping process in the same way.

To recapitulate, the spacing period can be controlled by three parameters: (i) the molecular weight of the diblock copolymer; (ii) the concentration of the polymer solution [61]; and (iii) the dipping velocity [61–64]. The latter two points are advantageous for the fabrication of particle arrays, as same type of diblock copolymer can be used for a certain loading parameter, and the particle size is therefore independent of the particle density. However, by neglecting the influence of the order parameter, a wide range of different interparticle spacings can be covered with a single diblock-copolymer by choosing appropriate conditions.

7.2.3.2 **Controlling the Particle Size**

The particle size is restricted to the amount of metal salt that can be loaded inside each micelle, and is therefore limited to a small range (typically 1–15 nm), depending on the loading ratio and the number of PVP units per micelle. In order to control the size over a wider range, the gold nanoparticles may be used as a seed for an additional growing step by "hydroxylamine seeding" [65, 66]. Here, the metal particles act as catalytic nuclei for the electroless deposition of metal ions from solution by hydroxylamine (NH_2OH). As the kinetics for the reduction of adsorbed metal ions exceeds the rate of reduction in solution, the nucleation of new particles is prevented and all of the ions will take part in the production of larger colloids [67, 68]. Unfortunately, this technique cannot be adapted to nanoparticulate substrates as the particles would lift off and the pattern geometry would be destroyed during the growth procedure.

The pattern geometry is preserved during the growth procedure by either: (i) embedding the nanoparticles into a SAM [69, 107]; or (ii) by using the polymer matrix as a stabilizing template [107]. A schematic overview of both strategies is shown in Figure 7.3.

In this case, the glass coverslips were patterned over the whole area below the dotted line. In the first approach, the bare glass area between the gold particles was functionalized by a monolayer of hexadecyltrimethoxysilane (HTMS) to form a stabilizing environment for the gold particles. The formation of a SAM in this system was caused by the selective silanization of the surface by forming a covalent Si−O−Si bond. The average particle height on the unmodified glass coverslip was revealed (via AFM measurements) to be 6.5 ± 0.4 nm, while the height of the particles embedded in the HTMS monolayer was 4.0 ± 0.4 nm. The difference of 2.5 ± 0.8 nm corresponded to the thickness of the monolayer, and was responsible for the lateral stabilization of the particles on the surface. The substrates were then immersed into an aqueous seeding solution of hydroxylamine and gold acid. Prior to the electroless

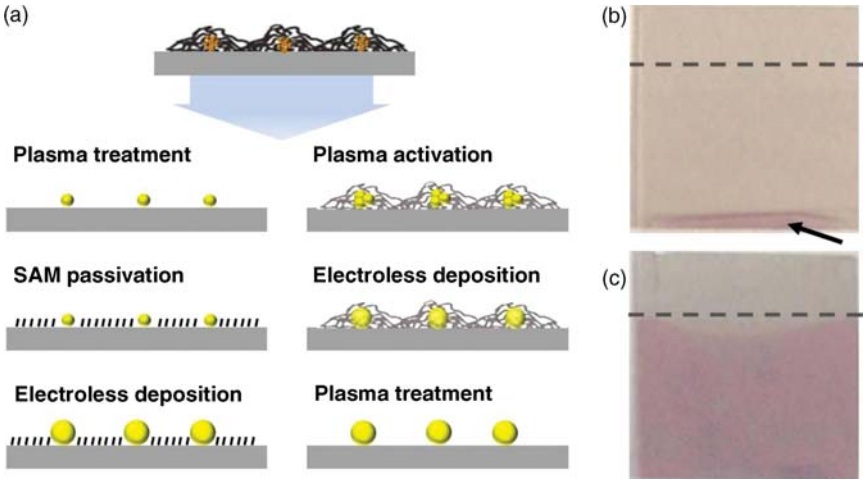


Figure 7.3 Schematic of the particle growth strategies. (a) Formation of a micellar monolayer by dip-coating. The particles are either stabilized by embedding them into a hexadecyltrimethoxysilane (HTMS) layer, or the micelles themselves act as a template at the surface to keep the particles in position; (b, c) Photographs of glass coverslips (size 20×20 mm) before (b) and after (c) particle enlargement. Both samples were decorated up to the dipping edge with nanoparticles (indicated by the dotted line). The successful reaction is made apparent by the homogeneous red color of the structured part of the substrate. For small particles, the appearance of the nanostructure is only visible at the dripping edge (arrow).

deposition process, the initial particle size of about 6 nm was too small to observe plasmon absorbance by the naked eye. The presence of gold particles, however, soon became evident by presence of a red stripe at the bottom of the dipping edge; this was an area of total disorder where multilayer formation and the agglomeration of small particles had led to a dense layer of gold clusters being formed after the plasma treatment. The course of the reaction was detectable after only a few seconds of immersion, as the patterned part of the substrate turned red due to increasing particle plasmon absorbance (Figure 7.3c), while the homogeneous red color of the sample indicated a uniform growth of the particles. Varying the metal type of the seeding solution and the interparticle distance enabled the precise preparation of highly ordered single and bimetallic core–shell nanostructures. Unfortunately, it was necessary to modify the sample by applying a stabilizing monolayer, though this was not always wanted – nor even possible – depending on the substrate chemistry. Notably, as the micellar technique was seen to be adaptive to a broad range of different substrate materials, its versatility was clearly limited.

In a second approach, the polymer shell of the micelles itself was used as a stabilizing matrix surrounding the metal core. In order to enable hydroxylamine reduction, the precursor salt inside each micelle must be reduced to generate a seed of the pure metal, and this was achieved by a short hydrogen plasma activation of the micellar films. The formation of elemental particles in the micelles occurred within

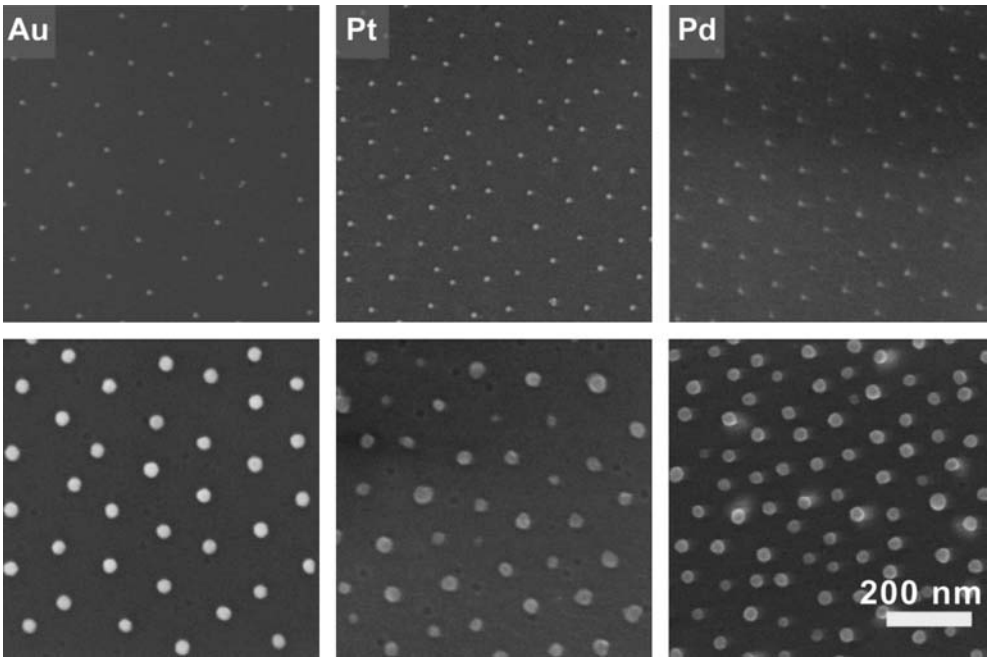


Figure 7.4 Au, Pt, Pd particles grown by intramicellar electroless deposition. Scanning electron microscopy (SEM) images of Au, Pt and Pd particles on glass coverslips before (top; 7 nm initial size) and after (bottom; 25 nm) particle growth. The initial particle size in all cases was ca. 6 nm. All particles were enlarged by intramicellar electroless deposition up to a size of ca. 25 nm, using different immersion times and conditions: Au (0.1% $HAuCl_4$/0.2 mM NH_3OHCl, 60 s); Pt (1% H_2PtCl_6/2 M NH_3OHCl, 20 h); Pd (0.1% Pd $(Ac)_2$/200 mM NH_3OHCl, 18 h). Adapted from Ref. [107].

the first few minutes, during which activation step some parts of the polymer matrix were etched, although the reaction time was not long enough to remove the polymer film completely [46]. The activated samples were dipped into the particular seeding solution, and the reaction then stopped by rinsing the substrates in ultrapure water. Finally, the samples were exposed to hydrogen plasma to remove the complete polymer residuals. Figure 7.4 shows the SEM images of the Au, Pt, and Pd particles arrays before (upper row) and after (lower row) the metal deposition.

The initial particle size in all cases was 6 nm, and all particles were enlarged by intra-micellar electroless deposition up to a size of 25 nm, using different immersion times and conditions. The geometry of the nanopattern was not affected by the seeding procedure. It should be noted that the experimental parameters for the electroless deposition of Au, Pt or Pd are very different. For example, the deposition time for platinum and palladium particles is much longer than for gold, and the seeding solution must be more concentrated. As a result of working with higher concentrations and longer reaction times, the size distribution of the enlarged platinum particles was less homogeneous than that of gold particles.

When comparing both strategies, the intra-micellar approach had the advantage that the enlargement step could be applied directly to the activated micellar film, which translated as a broader applicability. Particles with a diameter greater than 50 nm not only began to lift off from the surface, but also had a much more disperse size distribution. The latter effect may have been due to the polymer shell surrounding the particles and, in turn, affecting the particle size. Although the stabilization of particles by a SAM required additional modification of the substrate materials, it also allowed the controlled preparation of metal core–shell clusters.

7.2.3.3 **Micro-Nanopatterned Interfaces**

Besides varying the particle spacing, micro-nanopatterned morphologies represent an additional means of controlling the nanoparticle density. A combination of BCML with conventional "top down" technology enables the generation of aperiodic micro-/nanopatterned surfaces, and the directed location of nanometer-sized features. A schematic overview of these different strategies is shown in Figure 7.5.

Micro-nanopatterned interfaces may be fabricated by either a direct modification of the adsorbed micelles by irradiation with UV-light [70], with e-beam lithography [71, 72], or low-dose FIB milling [73]. In the simplest case, the monomicellar film itself can be used as a negative resist for e-beam lithography on conductive as well as nonconductive substrates. This allows the deposition of single submicron particle patches, or even single gold nanoparticles in an aperiodic pattern on conductive and nonconductive substrates. The micelles are pinned on the substrates as a result of the

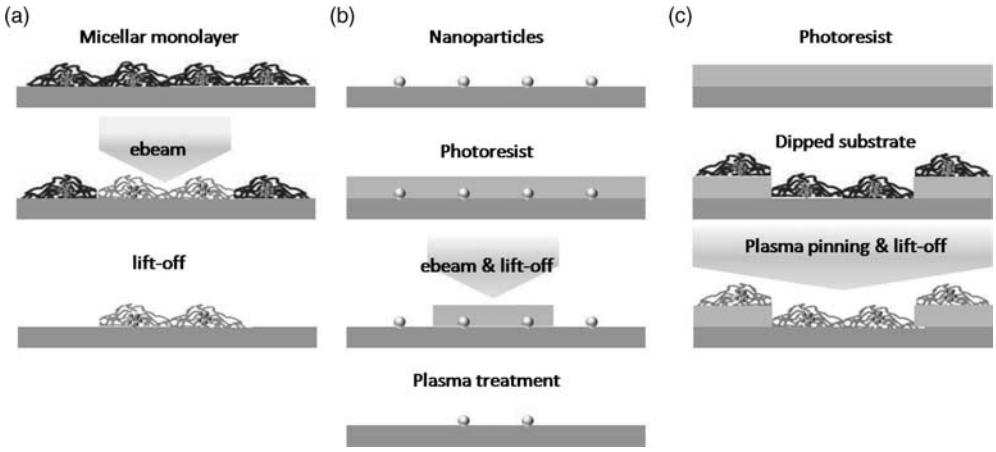


Figure 7.5 Schematic overview of different micro-nanopatterning strategies. (a) The micellar monolayer itself may be used as a resist for electron-beam lithography; (b) Alternatively, the gold nanoparticle array may be covered with a photosensitive resist; (c) By preparing nanoscopic cavities using top-down lithography, it is possible to trap individual micelles inside the prepatterned structure by dip-coating. The micro- nanopatterned substrate is generated by plasma-pinning the micelles to the surface, removal of the photoresist, and a final plasma treatment.

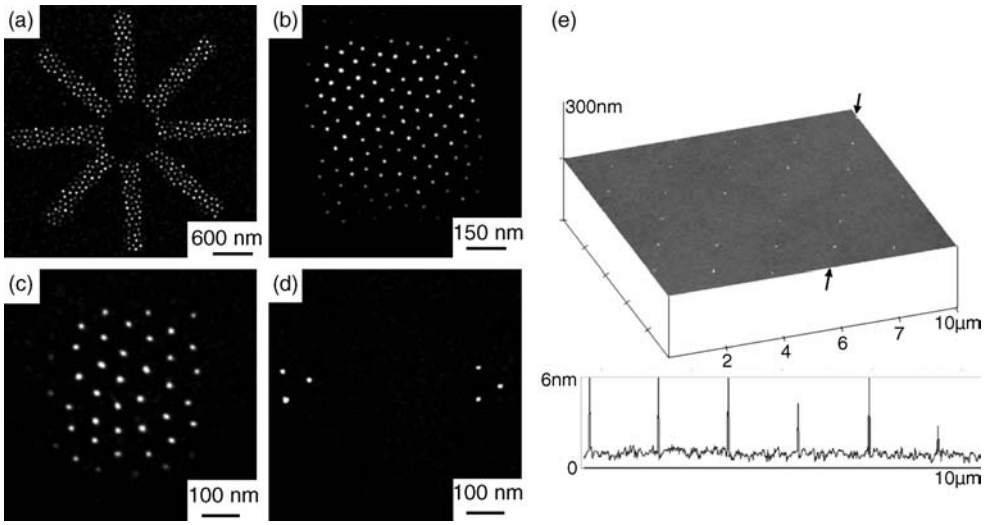


Figure 7.6 (a–d) Micro-nanopatterned particles arrays fabricated by e-beam lithography. (a) Gold nanoparticles arranged in a star pattern; (b, c) Squares patch side length of (b) ca. 600 nm and (c) ca. 200 nm side length; (d) Three single gold particles separated by ca. 400 nm; (e) By preparing separated photoresist cavities and subsequent deposition of single micelles, single particles can be precisely separated over 1 μm apart in a square pattern. Panels (a–d) adapted from Ref. [71]; Panel (e) adapted from Ref. [75].

interaction of the polymer film and the electron beam; subsequent immersion of the substrate into an organic solvent reveals the exposure pattern. The polymer shell of the tethered micelles can then be removed by a subsequent plasma process that is analogous to the normal protocol.

Whilst e-beam lithography has the advantage of high feature resolution, it suffers from the disadvantage of being a serial and time-consuming process. Nonetheless, it can be accelerated by using a combination of standard lithography and subsequent nanoparticle removal (Figure 7.5b) [74]. In this case, although the entire substrate with its gold particles is coated with a photoresist, only certain areas will be illuminated and the final pattern can be generated by removing any unmodified resist and the appropriate underlying particles. Moreover, this approach can be transferred to photolithography, which has the advantage that a large area can be patterned in one step. The major advantage with this combined method is that photolithography is capable of generating aperiodic microstructures, while the high-resolution nanopattern is generated by the micelles via a rapid and simple self-assembly process. Hence, the combination of these methods represents an intriguing approach to nanopatterning that can be conducted considerably faster and simpler than by using a conventional "top-down" approach. In a variation of this strategy, micelles can be deposited either into prestructured cavities (Figure 7.5c) or directly, using μCP [113]. The use of this approach has made possible the generation of linear and circular configurations of

gold nanoparticles, and the deposition of single nanoparticles separated by several micrometers [35, 75].

7.3
Applications of Nanopatterned Materials

7.3.1
Moth Eye Antireflective Surfaces

The faceted eye of dawn-active moths is equipped with a periodic array of subwavelength-structured protuberances, which behaves as a gradation of the refractive index between the air/cornea interface [76]. In the experiments described here, gold nanoparticles were used as a shadow mask for the subsequent reactive ion etching (RIE) of glass coverslips and fused silica substrates, with the aim of generating a nanometer-sized surface texture similar to the biological example.

The reduction of Fresnel reflections at optical interfaces is a topic of enormous interest for a wide range of applications [77]. The performance of the projection optics of both photographic and microscopy units is greatly affected by the reflection and transmission of light at optical interfaces. In the case of semiconductors, the reflection loss of light in the visible and near-infrared spectral regions may reach 40%, due to the high refractive indices of these materials [78]. Today, state-of-the-art antireflection (AR) coatings are most frequently based on multilayer interference structures with alternating high and low refractive indices [79, 80]. Unfortunately, however, such layer systems tend to perform suboptimally in many aspects, with thin-film coatings suffering from both adhesion problems and radiation damage if the optical device is used over a broad thermal range, or in high-power laser applications. Typical light sources for DUV illumination include excimer lasers such as KrF (248 nm) and ArF (193 nm). The number of available materials with a suitable refractive index to realize broadband antireflection coatings in this spectral region is greatly limited. Whilst thin-film coatings are commonly used to reduce the reflection of optical components in the visible range, similar technologies in the DUV spectral region are difficult to implement, and extremely expensive [81]. One alternative to these multilayer films would be to use subwavelength or antireflective structured surfaces [82] which, in nature, are found on the eyes of nocturnal insects. The compound eye of an insect consists of an arrangement of identical units, the *ommatidia*, each of which represents an independent eye with its own cornea and lens to focus light on the subjacent photoreceptor cells. In the case of nocturnal moths, the surface of each cornea is equipped with a hexagonal array of cuticular protuberances. This structure was first discovered by Bernhard [83], who proposed that the function of these "nipple arrays" might be to suppress reflections from the faceted eye surface in order to avoid fatal consequences for the moth if the reflection were to be detected by a bird or any other predator. The optical properties of a "moth eye" surface can, in principle, be understood as a gradation of the refractive index between air and the corneal material [84, 85]. Some SEM images of the surface

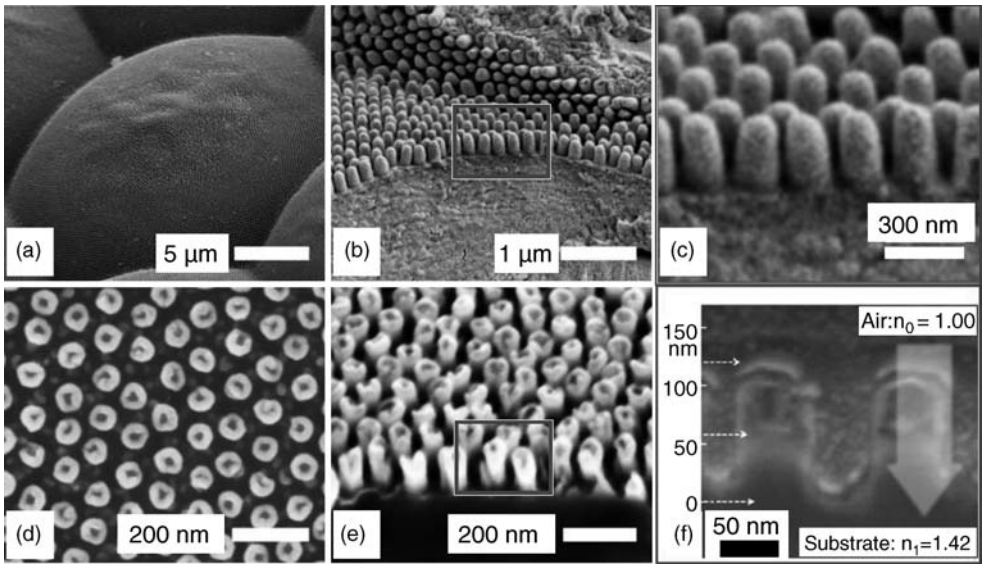


Figure 7.7 (a–c) Scanning electron microscopy (SEM) images of the surface of a genuine moth eye. The insect compound eye consists of microarrays of several thousand single lenslets. The lens of a single ommatidia is equipped with a fine array of protuberances with a structural period smaller than the wavelength of the incoming light. This special profile leads to a continuous increase of material density at the air/cornea interfaces, which results in a gradation of the refractive index; (d–f) SEM images of the "moth eye" structure on fused silica. Note: panel (c) is an enlargement of the box in panel (b); (d) Top-view SEM images of the structure displaying the quasi-hexagonal arrangement; (e) Side-view image of the pillar array measured with a tilt angle of 45°; (f) Focused ion beam (FIB) cross-section through the antireflective structure. The air/material transition is schematically implemented in the micrograph. The pillars have a diameter of 60 ± 4 nm and a lateral spacing of 114 ± 3 nm (center to center), respectively. The height of the structure was measured as 120 ± 5 nm, which corresponds to the effective thickness of the antireflective layer. A cone-type hole is etched into each pillar tip to approximately half of the pillar height. Adapted from Ref. [86].

patterns of a moth eye are shown in Figure 7.7 (details of the eyes of different butterfly species are available in Ref. [76]).

As the distance between the pillars is sufficiently small, the structure cannot be resolved by the incident light. Hence, transition between the air/material interface will appear as a continuous boundary, with the effect of a decreased reflection and improved transmittance of all light with a wavelength larger than the spacing period. Several characteristics of antireflective structured surface offer distinct advantages compared to layers of thin dielectric films. For example, thin-film coatings suffer from problems of mechanical stability, such as layer ablation and tensile stress, while appropriate coating materials with suitable refractive indices barely exist. Moreover, whereas common single- and multi-layer configurations are applicable only within a small wavelength range and normal incidence of light, moth eye structured materials show a reduced and angle-independent reflectance over a broad spectral bandwidth [84].

Artificial "moth eye" antireflective structures can be created by the RIE of prepatterned fused silica samples [86], such that the gold nanoparticles function as a protective resist due to their higher stability against the plasma treatment compared to the underlying material. Remarkably, the tips of the pillars on top of the fused silica sample are hollow, and pores are formed at spots where the gold particles had been placed originally. During the plasma process the reactive ions of the plasma are focused to the contact area of the metallic nanoparticles with the underlying fused silica substrate. This causes a strong depletion of the plasma-generated reactive ion concentration around the metal islands, which in turn causes the particles to act as an etching mask for the processing of hollow, cone-like pillars oriented perpendicular to the substrate. During the etching process, the particles sink into the material. As the RIE process represents an unselective physical ion bombardment of the sample, the gold particles are continuously reduced in size until completely used up, at which point the whole surface is considered to be uniformly etched.

The optical properties of the fabricated samples were observed via wavelength-dependent transmission measurements, and compared to an unstructured reference substrate. As indicated by the SEM images, the topology of the fused silica sample was similar to the corneal surface of a real moth. However, the moth eye lens showed a superior optical performance compared to many non-natural materials, as the overall reflection was reduced whilst the transmission of light in the visible range was increased. The optical properties of plane-fused silica samples were investigated by wavelength- and angle-dependent transmission and reflection measurements, and the data compared to theoretical values for unstructured reference samples (Figure 7.8).

An increase in total transmission was observed over a spectral range from 300 to 800 nm, with a maximum value of transmittance of 99.0%, while the reflectivity of the same sample was damped to 0.7%. As the improved transmission was in accordance

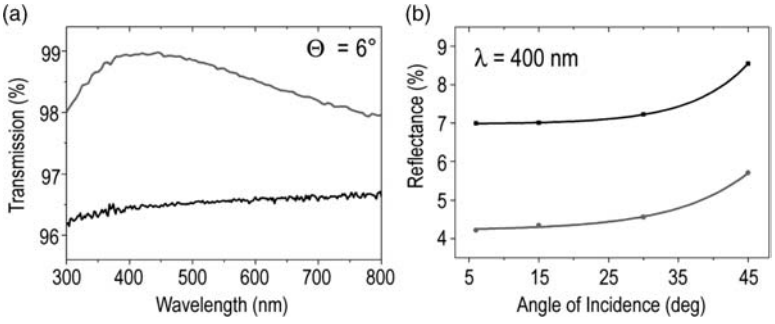


Figure 7.8 Broadband antireflective properties with varying incidence angles. (a) Wavelength-dependent transmission and reflectivity of a "moth eye" fused silica sample (upper line) compared to a reference (lower line); (b) Reflectance of a "moth eye"-structured (lower line) and a reference (upper line) sample as a function of the incident angle. The increase in transmission corresponds to an equivalent decrease in reflection. Adapted from Ref. [86].

with a reduced reflectance, however, it seemed apparent that the light-scattering defects or absorption losses were negligible.

As noted above, AR-structured surfaces enjoy certain advantages over layer-coating configurations, since the reflection is reduced for the omnidirectional incidence of light. This effect was revealed by ellipsometry measurements in which the non-polarized spectral reflection $(R_s + R_p)/2$ was investigated for different angles of incidence. When the reflectance data were compared with calculations of the reflectance of light from nonstructured fused quartz interfaces, the reflectivity of the subwavelength-structured interface was reduced to about 3% over the whole spectral region for incidence angles up to 45° (Figure 7.8b).

In order to demonstrate the excellent applicability of the method to nonplanar optical components, the convex side of a fused silica lens was processed and characterized by sub-300 nm transmission measurements. The planar-convex lens had a diameter of 22.4 mm and a focal distance of 100 mm, which corresponded to a radius of curvature of 50 mm. The reduced reflectivity of the structured part of the lens surface is shown in Figure 7.9a, where the dipping edge is indicated by a white arrow that corresponds to the border line between the antireflective-structured and nonstructured regions. More intense light reflectivity was seen above the dipping edge, whereas the antireflective part of the lens appeared less bright. When transmission in the DUV range was measured between 185 and 300 nm (Figure 7.9b), the performance was improved over the entire DUV spectral region, by 5% for 193 nm and 3% for 248 nm at the excimer laser wavelengths of ArF and KrF, respectively. This increase in transmission, of about 5%, was considered to relate to a virtual elimination of reflection at the modified optical interface.

Besides their remarkable optical properties, these structures offer additional advantages compared to thin-film coatings, in terms of mechanical stability and

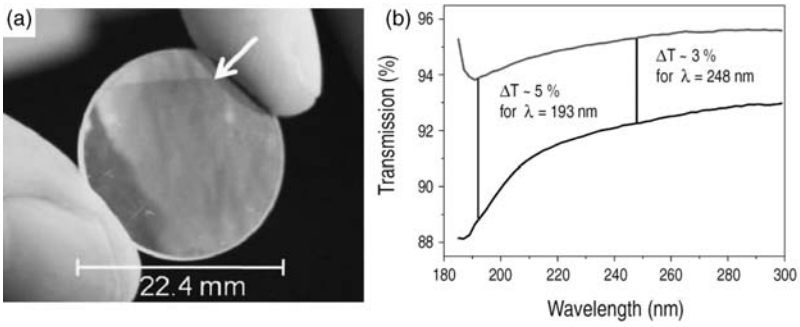


Figure 7.9 The optical properties of a "moth eye"-structured lens. (a) Photograph of the processed lens, demonstrating the antireflective effect. The borderline between the structured (below) and unstructured area is indicated by the white arrow; (b) Transmission spectra of the same lens before (lower line) and after (upper line) processing. An increase in transmission was observed over the whole DUV range, from 185 nm to 300 nm. The improved transmission values at the excimer laser wavelengths 193 nm (ArF) and 248 nm (KrF) are shown as examples. Adapted from Ref. [86].

durability. Moth-eye structured devices can also be used over a broad thermal range, as they are essentially free from adhesion problems and tensile stress between the substrate and the AR layer.

7.3.2 Cells on Nanostructured Interfaces

During recent years, increasing efforts have been made to acquire a deeper understanding of how cells interact with their environment, and in this respect nanopatterned substrates have been identified as powerful tools to engineer cellular environments to study cell-c interaction and adhesion [88]. These experiments, the main aims of which have been to mimic biological interfaces with defined chemical and physical properties, represent some of the most striking examples of the use of gold nanoparticle substrates [87]. For such nanoparticles, the adjustable separation distance of between 20 and 250 nm is within the size range of the nanoscopic subunits of the extracellular matrix (ECM), such as collagen [89]. These findings have underlined the initial proposal that biological processes occur at the nanoscale [90], while the artificial platforms designed specifically for biological applications have permitted identification of the details of cellular functions such as adhesion, migration [63, 91–93], proliferation [94], and differentiation [95], and also confirmed that these functions are regulated on the molecular level. Indeed, the site-controlled immobilization of bioactive molecules on nanoparticle patterns has opened the door to a variety of biologically active templates [96, 97], and such well-defined systems have provided insights into the density effects and spatial organization of cell membrane receptors. Arrays of micro- and nanopatterned adhesion molecules have also been used to investigate how tiny structural differences of only a few nanometers can influence the fate of a cell, whether it lives or dies [98, 99]. Clearly, these experiments have together provided important information regarding the mechanisms involved in cell–cell and cell–ECM interactions.

The adhesion of cells to the ECM is mediated by a class of transmembrane receptors of the integrin family. One important recognition sequence for the $\alpha_v\beta_3$ integrin is RGD, a peptide sequence that consists of arginine, glycine, and aspartic acid and is found in many ECM components [100].

In these experiments, gold nanoparticles were used as anchor points to tether cyclic RGD molecules c(RGDfK)-thiol to the gold nanoparticles on top of a glass coverslip (the set-up is shown schematically in Figure 7.10a). In order to avoid any unspecific binding of proteins to the substrate, and unspecific interactions of the cell with the glass substrate, the area between the nanoparticles must be passivated by either covalent [101] or electrostatic [102] functionalization with poly(ethylene glycol) (PEG). When MC3T3 osteoblasts were seeded onto the nanopattern, the cells showed a sensitive response depending on the separation distance of the RGD ligands on the surface (Figure 7.11); a particle size of 6–8 nm matched the size of a single integrin membrane receptor. When the cells were spread normally on a 58 nm patterned substrate, very few became attached and the spreading area was reduced on a sample where the RGD sequence was tethered more than 73 nm apart. Weak adhesion forces

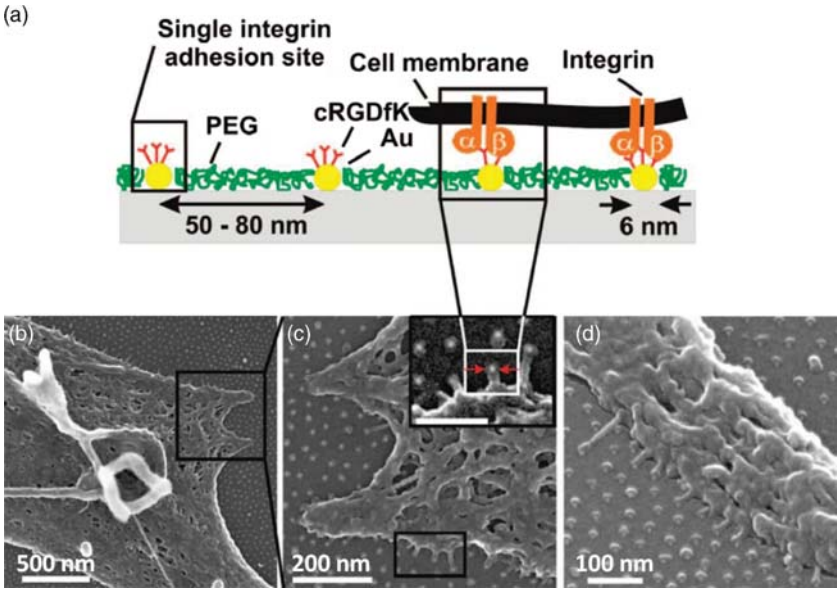


Figure 7.10 (a) Schematic of the biofunctionalized substrate. c(-RGDfK-) molecules are tethered to gold particles, which serve as adhesive anchor points for single integrin membrane receptors of living cells. Note that the particle diameter matches the structural dimensions of the integrin receptors itself. The bare glass surface between the gold nanoparticles is functionalized with a PEG monolayer to avoid protein adsorption and unspecific cell adhesion; (b–d) Scanning electron microscopy images of critical point-dried MC3T3 osteoblasts plated for 21 h on nanopatterned glass substrates. The particle spacing is set to ca. 60 nm. Close-up imaging of the cell rims reveals the interactions of cell protrusions with individual gold nanoparticles, underlining the sensitivity of the cells to communicate with the artificial substrate Adapted from Ref. [63].

were also measured on a nanopattern with a spacing larger than 58 nm [103]. Subsequent high-resolution SEM imaging of the fixed cells showed them to have formed nanoscopic protrusions that were attached to single nanoparticles. The results of these experiments confirmed that cell spreading and focal adhesion formation was impaired above a distance of 58 nm between two neighboring RGD motifs, despite the cells being sensitive enough to attach to individual particle islands. This, in turn, emphasized the fact that clustering of the $\alpha_v\beta_3$ integrin was necessary for the formation of focal adhesions [104, 105]. However, if a gradient nanopattern was presented to the cells, they migrated actively to areas with a lower separation distance between the adhesion motifs [63].

Remarkably, these cell-spreading experiments were conducted with several cell types, each of which showed a similar behavior. Similar experiments were carried out using micro-nanopatterned interfaces, and each produced similar results. Although the overall RGD density on a 58 nm pattern within micrometer-sized squares was lower than in the case of an homogeneous array with a 73 nm separation distance, focal adhesions were formed preferably on the finer-spaced substrates.

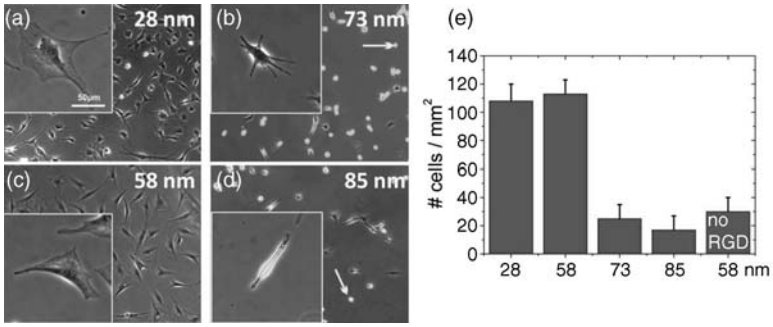


Figure 7.11 Phase-contrast microscopy images of MC3T3 osteoblasts seeded on nanopatterned surfaces, with interparticle distances of: (a) ~28 nm; (b) ~58 nm; (c) ~73 nm; and (d) ~85 nm. While cells spread very well on the 28 and 58 nm pattern, impaired cell spreading is observed on substrates with a interparticle distance >73 nm. Motionless cells appear round (arrows), while migrating cells show long extensions, sensing their environment (b, d). The number of cells attached to substrates with a spacing greater than 73 nm falls to almost the same level found on reference samples with no RGD Adapted from Ref. [98].

This confirmed that cell adhesion was indeed a consequence of the structural parameters on a nanoscopic length scale, rather than an overall ligand-density effect.

Recent developments have shown that the nanopattern can also be transferred to soft materials such as polydimethylsiloxane (PDMS), polystyrene, and hydrogels by the process of "transfer nanolithography" [106]. In addition, it is possible to create curved and complex, 3-D environments that are more closely related to the ECM of a cell *in vivo*. An example of this, using nanostructured hydrogel microtubes, is shown in Figure 7.12. In this case, the hydrogel-tubes were prepared by embedding nanopatterned glass fibers with diameters between 60 and 500 µm into a matrix of polyethylene glycol diacrylate (PEGDA). In order to ensure a complete transfer, the gold particles were functionalized with propene thiol linker molecules. After crosslinking the PEGDA, the glass fiber was dissolved with hydrofluoric acid, which resulted in a channel structure decorated internally with gold nanoparticles, as revealed by cryo-SEM imaging (Figure 7.12b and c). Nanopatterned soft materials offer the additional advantage of controlling the surface stiffness and viscoelastic properties of the substrate. This ability is especially advantageous when investigating the spatiomechanical properties of cell–ECM interactions.

7.4 Conclusions

Block copolymer micelle nanolithography is a rapid and highly reproducible technique for large-scale nanopatterning on the basis of pure self-assembly. In this

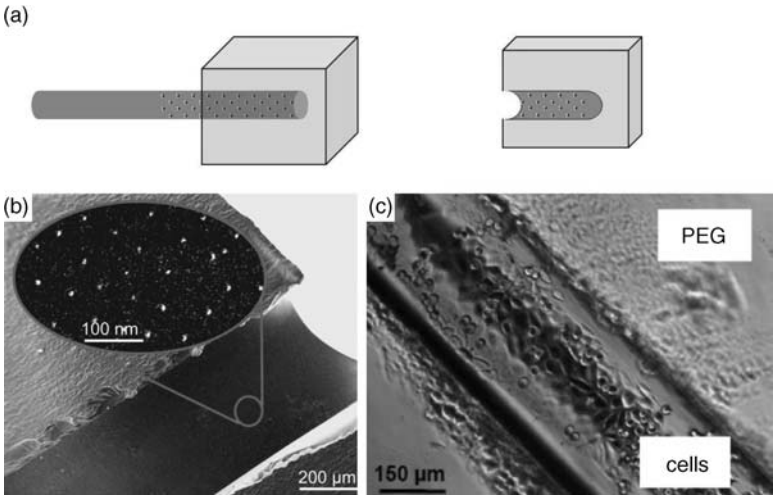


Figure 7.12 Formation of nanostructured hydrogel microtubes. (a) Nanoparticles are transferred from glass microfibers into PEGDA 700 block by using a transfer linker. Removal of the glass fiber resulted in a nanopatterned hydrogel tube; (b) Cryo-SEM image of a PEGDA 700 hydrogel channel decorated with gold nanoparticles; (c) HeLa cells cultured in PEGDA 700 hydrogel channels. The particles at the inside of the tube are functionalized with c(-RGDfK-). Adapted from Ref. [106].

chapter, aspects for characterizing nanoparticle arrays have been discussed, and the experimental conditions presented for controlling the separation distance and size of noble metal particles on top of solid substrates such as glass or silicon wafers. For this, Au, Pt, and Pd nanoparticles were grown homogeneously up to 50 nm and more on the basis of electroless deposition. The micellar approach can also be combined with a conventional top-down technology to prepare micro-nanopatterned interfaces.

The nanoparticle interfaces serve as a platform for biomimetic applications. As the structural period between the particles is of wavelength range, and less than that of visible light, pillar arrays fabricated via RIE demonstrate remarkable antireflective properties. When applied to optical functional materials, this approach represents a rapid, inexpensive and reproducible means of creating highly light-transmissive, antireflective optical devices for use as display panels and projection optics, and also for heat-generating microscopic and excimer laser applications.

Gold particles arrays may also serve as anchor points for the selective binding of extracellular proteins, with such biomimetic environments presenting a striking experimental platform for the investigation of cell adhesion. Application of the high-resolution spatial positioning of signaling molecules to inorganic or polymeric supports will allow the creation of a unique artificial environment for examining the important aspects of spatioregulated cell behavior such as cell attachment, spreading, and migration under molecular control.

References

1 Royal Society and Royal Academy of Engineering, London (2004) *Nanoscience and Nanotechnology: Opportunities and Uncertainties*, The Royal Society and The Royal Academy of Engineering, London.
2 Feynman, R.P. (1960) *Eng. Sci.*, **23**, 22.
3 Whitesides, G.M. (2005) *Small*, **1**, 172.
4 Moore, G.E. (1965) *Electronics*, **38**, 4 pages.
5 Whitesides, G.M. (2003) *Nat. Biotechnol.*, **21**, 1161.
6 Moreau, W.M. (1988) *Semiconductor Lithography: Principles and Materials*, Plenum, New York.
7 Brambley, D., Martin, D., and Prewett, P.D. (1994) *Adv. Mater. Opt. Electron.*, **4**, 55.
8 Feldman, M. and Sun, J. (1992) *J. Vac. Sci. Technol. B*, **10**, 3173.
9 Silverman, J.P. (1997) *J. Vac. Sci. Technol. B*, **15**, 2117.
10 McCord, M.A. (1997) *J. Vac. Sci. Technol. B*, **15**, 2125.
11 Matsui, S., Kojima, Y., Ochiai, Y., and Honda, T. (1991) *J. Vac. Sci. Technol. B.*, **9**, 2622.
12 Melngailis, J. (1987) *J. Vac. Sci. Technol. B.*, **5**, 469.
13 Ito, I. (2004) *Dekker Encyclopedia of Nanoscience and Nanotechnology*, Marcel Dekker Ltd, New York, Vol. 1, pp. 2413–2422.
14 Smith, B.W. (2009) *Proc. SPIE*, 7274.
15 Holmes, S.J., Mitchell, P.H., and Hakey, M.C. (1997) *IBM J. Res. Dev.*, **41** (1 & 2), 7–20.
16 Philip, D. and Stoddart, J.F. (1996) *Angew. Chem., Int. Ed.*, **35**, 1155.
17 Whitesides, G.M. and Grzybowski, B. (2002) *Science*, **295**, 2418.
18 Gates, B.D., Xu, Q., Love, J.C., Wolfe, D.B., and Whitesides, G.M. (2004) *Annu. Rev. Mater. Res.*, **34**, 339.
19 Love, C.J., Estroff, L.A., Kriebel, J.K., Nuzzo, R.G., and Whitesides, G.M. (2004) *Chem. Rev.*, **105**, 1103.
20 Ulman, A. (1996) *Chem. Rev.*, **96**, 1533.
21 Park, M., Harrison, C., Chaikin, P.M., Register, R.A., and Adamson, D.H. (1997) *Science*, **276**, 1401.
22 Black, C.T. *et al.* (2007) *IBM J. Res. Dev.*, **51**, 605.
23 Hamley, I.W. (2003) *Nanotechnology*, **14**, 39.
24 Xia, Y., Gates, B., Yin, Y., and Lu, Y. (2000) *Adv. Mater.*, **12**, 693.
25 Yang, S.M., Jang, S.G., Choi, D.G., Kim, S., and Yu, H.K. (2006) *Small*, **2**, 458.
26 Xia, Y., Rogers, J.A., Paul, K.E., and Whitesides, G.M. (1999) *Chem. Rev.*, **99**, 1823.
27 Xia, Y. and Whitesides, G.M. (1998) *Angew. Chem.*, **110**, 569.
28 Piner, R.D., Zhu, J., Xu, F., Hong, S., and Mirkin, C.A. (1999) *Science*, **283**, 661.
29 Salaita, K. *et al.* (2006) *Angew. Chem., Int. Ed.*, **45**, 7220–7223.
30 Wadu-Mesthrige, K., Xu, S., Amro, N.A., and Liu, G. (1999) *Langmuir*, **15**, 8580.
31 Xu, S., Miller, S., Laibinis, P.E., and Liu, G. (1999) *Langmuir*, **15**, 7244.
32 Herndon, M.K., Collins, R.T., Hollingsworth, R.E., Larson, P.R., and Johnson, M.B. (1999) *Appl. Phys. Lett.*, **74**, 141.
33 Spatz, J.P. *et al.* (2000) *Langmuir*, **16**, 407.
34 Glass, R., Moeller, M., and Spatz, J.P. (2003) *Nanotechnology*, **14**, 1153.
35 Leibler, P. (1980) *Macromolecules*, **13**, 1602.
36 Bates, F.S. and Fredrickson, G.H. (1990) *Annu. Rev. Phys. Chem.*, **41**, 525–557.
37 Israelachvili, J. (1992) *Intramolecular and Surface Forces*, 2nd edn, Academic Press, London.
38 Israelachvili, J. (1994) *Langmuir*, **10**, 3774.
39 Gao, Z. and Eisenberg, A. (1993) *Macromolecules*, **26**, 7353.
40 Izzo, D. and Marques, C.M. (1993) *Macromolecules*, **26**, 7189.
41 Spatz, J.P., Roescher, A., Sheiko, S., Krausch, G., and Moeller, M. (1995) *Adv. Mater.*, **7**, 731.
42 Spatz, J.P., Moessmer, S., and Moeller, M. (1996) *Chem. Eur. J.*, **2**, 1552.
43 Spatz, J.P., Sheiko, S., and Moeller, M. (1996) *Macromolecules*, **29**, 3220.
44 Moessmer, S. *et al.* (2000) *Macromolecules*, **33**, 4791.

45 Kaestle, G. *et al.* (2003) *Adv. Funct. Mater.*, **13**, 853.

46 Heimendahl, M.V. (1980) *Electron Microscopy of Materials: An Introduction*, Academic Press, San Diego, USA.

47 Bhushan, B. (2004) *Springer Handbook of Nanotechnology* (ed. B. Bhushan), Springer-Verlag, Heidelberg, Germany.

48 Grabar, K.C. *et al.* (1997) *Anal. Chem.*, **69**, 471.

49 Angelescu, D.E., Harrison, C.K., Trawick, M.L., Register, R.A., and Chaikin, P.M. (2005) *Phys. Rev. Lett.*, **95**, 025702.

50 Murray, C.A. and Van Winkle, D.A. (1987) *Phys. Rev. Lett.*, **58**, 1200.

51 Kosterlitz, J.M. and Thouless, D.J. (1972) *J. Phys. C*, **6**, 1181.

52 Halperin, B.I. and Nelson, D.R. (1978) *Phys. Rev. Lett.*, **41**, 121.

53 Nelson, D.R. and Halperin, B.I. (1979) *Phys. Rev. B*, **19**, 2457.

54 Young, A.P. (1979) *Phys. Rev. B*, **19**, 1855.

55 Murray, C.A. and Grier, D.G. (1996) *Annu. Rev. Phys. Chem.*, **47**, 421.

56 Spatz, J.P., Roescher, A., and Moeller, M. (1996) *Adv. Mater.*, **8**, 337.

57 Förster, S. and Plantenberg, T. (2002) *Angew. Chem., Int. Ed.*, **114**, 712.

58 Darhuber, A.A., Troian, S.M., Davis, J.M., Miller, S.M., and Wagner, S. (2000) *J. Appl. Phys.*, **88**, 5119.

59 Landau, L.D. and Levich, B. (1942) *Acta Physicochem.*, **17**, 42.

60 Wilson, S.D.R. (1982) *J. Eng. Math.*, **16**, 209.

61 Krishnamoorthy, S., Raphaël, P., Brugger, J., Heinzelmann, H., and Hinderling, C. (2006) *Adv. Funct. Mater.*, **16**, 1469–1475.

62 Bansmann, J. *et al.* (2007) *Langmuir*, **23**, 10150.

63 Arnold, M. *et al.* (2008) *Nano Lett.*, **8**, 2063.

64 Moeller, M. *et al.* (2004) *Polym. Mater. Sci. Eng.*, **90**, 255–256.

65 Brown, K.R. and Natan, M.J. (1998) *Langmuir*, **14**, 726.

66 Sheffer, M. *et al.* (2001) *Langmuir*, **17**, 1713.

67 Stremsdoerfer, G., Martin, J.R., and Clechet, P. (1992) *Electrochem. Soc. Proc.*, **92–93**, 305.

68 Schlesinger, M. (2000) *Modern Electroplating* (ed. M. Schlesinger), John Wiley & Sons, Inc., New York, Weinheim.

69 Resch, R. *et al.* (2001) *Langmuir*, **17**, 5666.

70 Gorzolnik, B., Mela, O., and Moeller, M. (2006) *Nanotechnology*, **17**, 5027.

71 Glass, R. *et al.* (2004) *New J. Phys.*, **6** (101), 17 pages.

72 Glass, R. *et al.* (2003) *Adv. Funct. Mater.*, **13**, 569.

73 Mela, P. *et al.* (2007) *Small*, **3**, 1368–1373.

74 Aydin, D. *et al.* (2009) *Small*, **5**, 1014.

75 Spatz, J.P. *et al.* (2002) *Adv. Mater.*, **14**, 1827.

76 Stavenga, D.G., Foletti, S., Palasantzas, G., and Arikawa, K. (2006) *Proc. Biol. Sci.*, **273**, 661.

77 Kikuta, H., Toyota, H., and Yu, W.i. (2003) *Opt. Rev.*, **10**, 63.

78 Singh, J. (2003) *Electronic and Optoelectronic Properties of Semiconductor Structures*, Cambridge University Press, Cambridge, UK.

79 Sandrock, M. *et al.* (2004) *Appl. Phys. Lett.*, **84**, 3621.

80 Xi, J.Q. *et al.* (2007) *Nat. Photonics*, **1**, 176.

81 Ullmann, J. *et al.* (2000) *Proc. SPIE*, **3902**, 514.

82 Brunner, R. *et al.* (2008) *Proc. SPIE*, **7057**, 707705-1–705705-10.

83 Bernhard, C.G. (1967) *Endeavour*, **26**, 79.

84 Clapham, P.B. and Hutley, M.C. (1973) *Nature*, **244**, 281.

85 Wilson, S.J. and Hutley, M.C. (1982) *Opt. Acta*, **7**, 993.

86 Lohmueller, T., Helgert, M., Sundermann, M., Brunner, R., and Spatz, J.P. (2008) *Nano Lett.*, **8**, 1429.

87 Spatz, J.P. and Geiger, B. (2007) *Methods Cell Biol.*, **83**, 89.

88 Stevens, M.M. and George, J.H. (2005) *Science*, **310**, 1135.

89 Meller, D., Peters, K., and Meller, K. (1997) *Cell Tissue Res.*, **288**, 111–118.

90 Niemeyer, C. and Mirkin, C.A. (2004) *Nano Biotechnology: Concepts, Applications and Perspectives* (eds C. Niemeyer and C.A. Mirkin), Wiley-VCH, Weinheim.

91 Koo, L.Y., Irvine, D.J., Mayes, A.M., Lauffenburger, D.A., and Griffith, L.G. (2002) *J Cell Sci.*, **115**, 1423.

92 Jiang, X.Y., Bruzewicz, D.A., Wong, A.P., Piel, M., and Whitesides, G.M. (2005)

Proc. Natl Acad. Sci. USA, **102** (4), 975–978.
93 Cavalcanti-Adam, E.A. *et al.* (2006) *Eur. J. Cell Biol.*, **85**, 219.
94 Yima, E.K.F. *et al.* (2005) *Biomaterials*, **26**, 5405.
95 Yim, E., Pang, S., and Leong, K. (2007) *Exp. Cell Res.*, **313**, 1820.
96 Wolfram, T., Belz, F., Schön, T., and Spatz, J.P. (2007) *Biointerphases*, **2**, 44.
97 de Mel, A., Jell, G., Stevens, M.M., and Seifalian, A.M. (2008) *Biomacromolecules*, **9**, 2969–2979.
98 Arnold, M. *et al.* (2004) *ChemPhysChem*, **4**, 872.
99 Chen, C.S., Mrksich, M., Huang, S., Whitesides, G.M., and Ingber, D.E. (1997) *Science*, **276**, 1425.
100 Geiger, B., Bershadsky, A., Pankov, R., and Yamada, K. (2001) *Nat. Rev. Mol. Cell Biol.*, **2**, 793–805.
101 Bluemmel, J. *et al.* (2007) *Biomaterials*, **22**, 4739.
102 Kenausis, G.L. *et al.* (2000) *J. Phys. Chem. B*, **104**, 3298–3309.
103 Walter, N., Selhuber, C., Kessler, H., and Spatz, J.P. (2006) *Nano Lett.*, **6**, 4380.
104 Cavalcanti-Adam, E.A. *et al.* (2007) *Biophys. J.*, **92**, 2964.
105 Geiger, B., Spatz, J.P., and Bershadsky, A.D. (2009) *Nat. Rev. Mol. Cell Biol.*, **10**, 21.
106 Graeter, S.V. *et al.* (2007) *Nano Lett.*, **7**, 1413.
107 Lohmueller, T., Bock, E., and Spatz, J.P. (2008) *Adv. Mater.*, **20**, 2297.
108 Spatz, J.P. *et al.* (1998) *Adv. Mater.*, **10**, 473.
109 Ethirajan, A. *et al.* (2007) *Adv. Mater.*, **19**, 406.
110 Yun, S.H., Yoo, S.I., Jung, J.C., Zin, W.C., and Sohn, B.H. (2006) *Chem. Mater.*, **18**, 5646.
111 Park, S., Kim, B., Yavuzcetin, O., Tuominen, M.T., and Russell, T.P. (2008) *ACS Nano.*, **2**, 1363–1370.
112 Park, S. *et al.* (2009) *Science*, **20**, 1030.
113 Chen, J., Mela, P., Moeller, M., and Lensen, M.C. (2009) *ACS Nano.*, **3**, 1451–1456.

8
The Evolution of Langmuir–Blodgett Patterning

Xiaodong Chen and Lifeng Chi

8.1
Introduction

During recent years, surface patterning with nano- or microscopic structures has attracted increasing scientific and technological interest in the research areas of materials science, chemistry, biology, and physics. For instance, patterned surfaces can be used to control the crystal nucleation and to manipulate crystallographic orientation [1], or to guide the self-assembly of polymers [2]. In addition to their uses in templates, patterned surfaces are essential to the development of a number of existing and emerging technologies, such as ultrahigh-density information storage [3]. The ability to fabricate patterned surfaces on the micrometer or nanometer scale also guarantees a continuation in the miniaturization of functional devices, such as the patterned assembly of integrated semiconductor devices [4] and the patterned luminescence of organic light-emitting diodes (LEDs) [5] in microelectronics. Likewise, direct liquid flow on a selectively patterned surface is important for the development of microfluidic systems, and for the miniaturization of flow devices [6]. Active efforts are also under way, for example, to develop micropatterned cell and/or protein arrays for biosensors [7, 8], for microliter chromatography [9], for biological recognition processes [10], and for DNA separation [11, 12]. An easy access to, and cost-effective large-area nanopatterning of, biocompatible films is also important for gene and drug delivery systems, and for tissue engineering [13, 14]. The ability to fabricate patterned surfaces on the microscale or nanoscale also allows for the manipulation of surface wettability [15].

In almost all applications of patterned surfaces, nanostructure fabrication represents the first – perhaps also the most significant – challenges to their realization. Until now, many strategies have been developed for fabricating patterned surfaces, including: (i) photolithography; (ii) electron beam (e-beam) lithography; (iii) scanning probe-based lithography, including dip-pen nanolithography (DPN) [16, 17]; (iv) nanoimprinting lithography (NIL) [18]; and (v) soft lithography [19]. These methods are normally classified as "top-down" approaches, and have demonstrated a high spatial resolution. In contrast, the concepts of self-assembly and self-organization provide an alternative and simple means of realizing small features over large areas via so-called "bottom-up" approaches. These rely on the interactions of building

Nanotechnology, Volume 8: Nanostructured Surfaces. Edited by Lifeng Chi

ISBN: 978-3-527-31739-4

blocks (such as molecules or nanoparticles) that assemble spontaneously into nano/microstructures. A variety of strategies based on self-assembly, including block copolymer-based lithography, have been demonstrated and subsequently used to fabricate patterned structures [20–22]. Among many of these self-assembly techniques, the Langmuir–Blodgett (LB) technique consists of a series of efficient and parallel processes by which to build up patterned structures on solid surfaces that are chemically or physically differentiated on the micro to submicron scale [23]. A unique property of the LB technique is its ability to provide control over nanoscale assembly by tuning macroscopic properties such as the surface pressure, the molecular composition of monolayer, transfer velocity, and the temperature, subphase, and substrate.

The aim of this chapter is to provide a comprehensive description of the development of the LB technique, as used to fabricate and pattern nanostructures on solid substrates. After a description of the technique's history, controlled LB pattern formation based on small organic molecules and macromolecules is discussed, followed by details of the patterning of nanoparticles and nanowires using the LB method. The applications of nanostructures fabricated via the LB technique are summarized.

8.2 The LB Technique in Retrospect: From Homogeneous Film to Lateral Features

The history of the LB technique can be traced back to experiments conducted by Benjamin Franklin in 1773, when he dropped a teaspoon of oil onto the water surface of a pond [24]. Over a century later, Lord Rayleigh [25] and Agnes Pockels [26] quantified the oil film on the water surface, providing details of its thickness ($\sim$0.16 nm) and molecular area coverage ($\sim$0.2 nm^2). Based on these data, Irving Langmuir noted during the early twentieth century that the monolayers of fatty acids could be compressed into a solid-like ordered state on the surface of water, and this led to the development of the Langmuir trough [27]. Subsequently, Langmuir and his student, Katharine Blodgett, showed that the fatty acid monolayers on the water surfaces could be transferred onto a solid support by passing a solid substrate vertically through the air/water interface [28]. Today, this general process is referred to as the LB technique.

Since the 1960s, many stimulating studies have been carried out by Hans Kuhn and others, using LB films, that have led to applications in the fields of electronics, optics, and biology [29, 30]. Yet, within the past two decades the field has undergone a revolution, due mainly to the development of novel experimental techniques or to the enhancement of traditional techniques, including synchrotron X-ray diffraction [31], fluorescence microscopy [32], and Brewster angle microscopy (BAM) [33, 34], each of which can be used to observe monolayers directly on the water surface. Together, these techniques can be used to demonstrate the phase transition behavior and morphological features in Langmuir monolayers, that cannot be observed directly in classical isotherm measurements [35, 36]. Furthermore, the

development of atomic force microscopy (AFM) has provided a means of directly imaging the morphology of the monolayer, when transferred onto a solid substrate, with resolution down to the molecular scale [37, 38]. These findings broke the traditional concept that the LB technique could be used only to create homogeneous and defect-free ultrathin films. However, with the introduction of modern experimental methods and their understanding, the LB method has developed into a high-throughput, low-cost, easily integrated method for the controlled assembly and patterning of building blocks that forms the basis of this chapter.

8.3
LB Patterning of Organic Molecules

8.3.1
Direct Transfer of Featured Structures Onto Solid Substrates

Traditionally, the LB technique has proved to be a highly versatile tool for the fabrication of homogeneous organic films on solid substrates, by transferring the closely packed monolayer onto the substrate itself. In contrast, through a rational molecular design, the organic molecules can form either nanostructures or microstructures on the water surface, and these can be transferred onto a solid substrate using the LB technique. As an example, whereas partially fluorinated long-chain fatty acids can form sharply monodisperse circular nanostructures on water surfaces during their spreading [39], normal fatty acids do not behave in this way. The mismatch between hydrocarbon segments and fluorinated segments is responsible for the formation of clusters. In one circular nanostructure, the hydrocarbon segments are packed due to van der Waals attractive interactions, although the packing is restricted by thick and stiff rods of perfluoroalkane helix chains. Semifluorinated phosphonic acids can also form stable nanoscale clusters on substrates, with similar behavior [40–42]. Furthermore, Krafft *et al.* found that semifluorinated alkanes (FnHm), diblock molecules with one hydrocarbon segment and one perfluorinated segment, can form nanopatterned structures on water surfaces [43]. The semifluorinated alkanes do not behave as typical amphiphiles, which normally contain one hydrophobic chain with one hydrophilic headgroup. Krafft *et al.* confirmed that the hydrocarbon segments of the semifluorinated alkane molecules were directed towards the substrate, while the fluorinated segments pointed outwards, towards the air. Moreover, the size of these nanostructures could be controlled by the density mismatch between the fluorinated and hydrogenated segments, which was originally due to an adjustment of the intermolecular interactions. Depending on the molecular structure of the FnHm diblocks, the nanostructures were either circular or elongated; however, increasing the FnHm length favored the formation of elongated nanostructures, albeit at the expense of the circular forms (see Figure 8.1) [41, 42].

Liu *et al.* found that the properties of nanostructures could easily be tuned through a rational molecular design and deposition condition [44, 45]. Initially, it was found that the achiral molecules could form chiral superstructures, depending on the

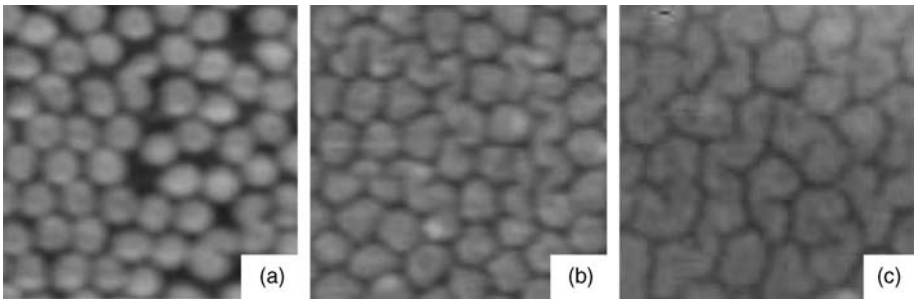


Figure 8.1 Atomic force microscopy images (250 × 250 nm) of transferred monolayers of (a) F8H16, (b) F8H18, and (c) F8H20 transferred onto silicon wafers at 5 mN m^{-1}. Reproduced with permission from Ref. [42].

surface pressure [44]. For example, an achiral amphiphilic derivative of barbituric acid (BA) could form two-dimensional (2-D) spiral structures at a low surface pressure (7 mN m^{-1}), but not at a higher surface pressure (20–30 mN m^{-1}) [44]. The spiral structures showed a clear Cotton effect for the circular dichroism (CD) measurements when they were transferred onto solid substrates. It was suggested that the large aromatic rings of the head groups, together with hydrogen bonding between the BA molecules, might be responsible for a preferential tilting of neighboring molecules in the packed films, and that spiral structures were produced due to the directionality of the hydrogen bonding interactions. Moreover, the morphology of the superstructure was found to play an important role in the properties of the structures [45]. For example, an anthracene derivative would form nanocoils at the water surface with a lower surface pressure (9 mN m^{-1}), but form straight nanoribbons at a higher surface pressure (20 mN m^{-1}). In addition, the straight nanoribbons showed an effect called "photoswitching," whereby their conductivity would change when they were exposed to light. The property of photoswitching arises because the molecules are arranged with their benzene rings stacked almost directly on top of one another, which causes the molecular orbitals of the π-electrons to overlap, leading to a more efficient charge transport. As the coiled nanoribbons had a less efficient stacking, they did not demonstrate photoswitching (Figure 8.2).

The shape of the superstructures (Figure 8.3a) can also be extensively controlled by the subphase conditions. For example, a chiral amphiphilic molecule, C12-(L)Cys-(L) Cys-C18, which consists of two short cysteine peptides as hydrophilic heads and two hydrophobic alkyl chains as tails, can form chiral domains at the air/water interface, owing to intermolecular hydrogen-bonding and hydrophobic interactions [46]. However, when the subphase contained 10^{-8} M $CdCl_2$, the superstructure was changed to a spiral structure, and the size of the structure greatly reduced. Alternatively, when the subphase contained 10^{-6} M CdTe nanoparticles, the superstructure was changed to linear nanostructures (Figure 8.3b) [47]. The reasons for this include: (i) that the addition of electrolytes might reduce the molecular interaction; or (ii) that the presence of thiol groups within the hydrophilic heads of the C12-(L)Cys-(L)Cys-C18 molecule allowed the complexation of metal or semiconductor nanocrystals

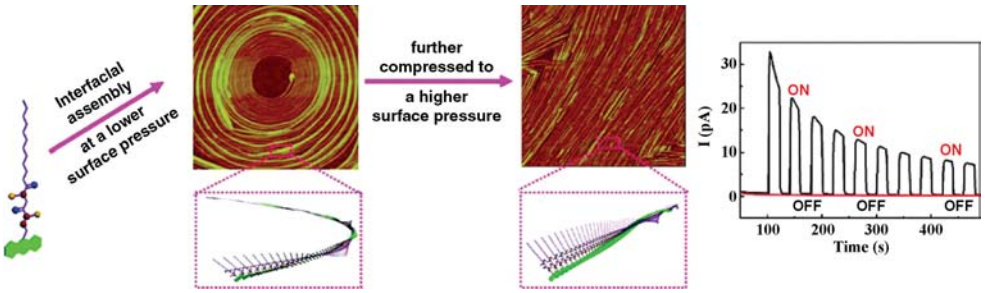


Figure 8.2 Atomic force microscopy image of monolayer on mica deposited at 9 mN m^{-1} (left) and 20 mN m^{-1} (right). The graph at the right shows the photoswitching characteristics of the two-end devices based on the films of BA by white light for the films deposited at 9 mN m^{-1} (red) and 20 mN m^{-1} (green); the scheme for the formation of nanocoils and straight nanoribbons. Reproduced with permission from Ref. [45].

(NCs), which might also change the molecular interaction. In this case, by simply changing the subphase, it would be possible to construct controlled lateral structures from the sub-micrometer scale down to the nanometer scale, simply by adjusting the subtle balance of the molecular interactions (with one or more chemical components). A second example is that the lanthanide ion could induce the stripe formation of phospholipid monolayers through the dynamic binding of subphase lanthanide ions to the phosphocholine headgroups at the air/water interface [48]. Competitive dipole–dipole and electrostatic interactions between the lanthanide-bound and free phospholipid molecules might then have produced long-range ordered arrays of phospholipid stripes, with periodicities of 1.7–1.8 μm.

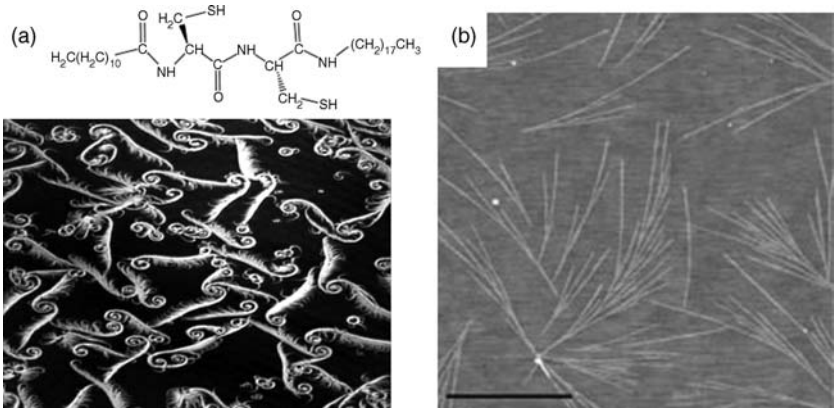


Figure 8.3 (a) Chemical structure of a C12-(L)Cys-(L)Cys-C18 molecule and Brewster angle microscopy (BAM) image of chiral domains formed by C12-(L)Cys-(L)Cys-C18 at the air/water interface; (b) Atomic force microscopy image of LB domains of a chiral compound, C12-(L)Cys-(L)Cys-C18, transferred onto a silicon substrate with the subphase containing 10^{-6} M of a CdTe nanoparticle. Scale bar = 2 μm. Panel (a) reprinted with permission from Ref. [46]; panel (b) reprinted with permission from Ref. [47].

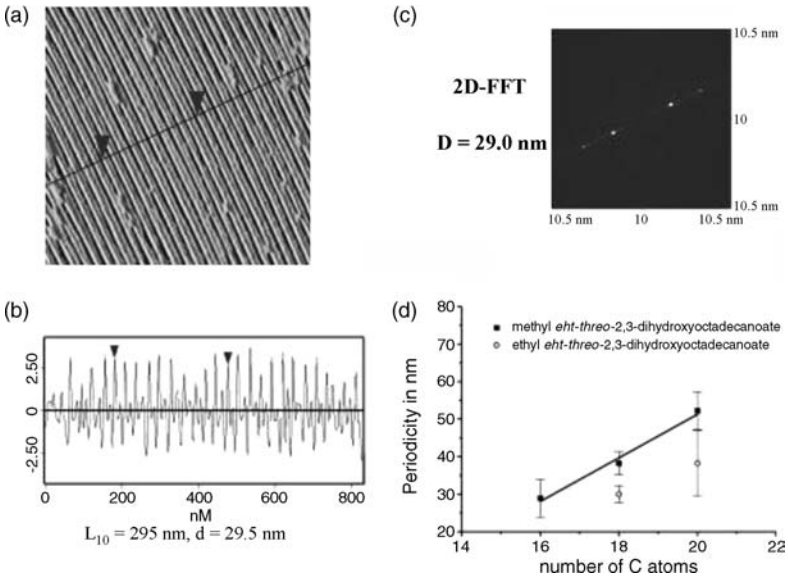


Figure 8.4 (a) Supermolecular periodic structures in a monolayer of ethyl-ent-*threo*-2,3-dihydroxyoctadecanoate. The periodicity was measured as 29 nm, evaluated from (b) the line section of the image or (c) 2-D-FET of the image; (d) The variation in periodicity by changes in chain length or head group. Reproduced with permission from Ref. [49].

The methyl and ethyl esters of ent-*threo*-2,3-dihydroxy fatty acids can also form periodic nanometer-sized structures, as observed with AFM after being transferred onto solid substrates (Figure 8.4) [49]. The ordered structures were not formed by the regular packing of single molecules, but rather by molecular assemblies. Here, it was found that the periodicity could be adjusted by varying the alkyl chain length or head group. For example, in the case of ethyl ent-*threo*-2,3-dihydroxyoctadecanoate, the periodicity was 30 nm, but this was changed to 38 nm for the case of ethyl ent-threo-2,3-dihydroxyicosanoate. Likewise, Kelley *et al.* showed that some amphiphilic β-hairpin peptides could form ordered periodic nanostructures on mica [50].

The molecular compositions of the monolayer can also be used to control the feature on the water surface. The formation of various types of pattern can be achieved through self-organization processes, such as the growth of condensed-phase domains in an expanded phase at the phase-transition region during micro-phase separation in mixed monolayers. For instance, by using a micro-phase separation in binary mixed Langmuir monolayers of cadmium salts of *n*-alkyl fatty acids and a perfluoropolyether surfactant, which separate into microscopic domains of condensed phase and a surrounding matrix of expanded phase, respectively, it was shown that the pattern shape would depend on the alkyl chain length of *n*-alkyl fatty acids and the temperature of the water surface [51, 52].

Besides the small organic molecules, amphiphilic diblock copolymers represent an important class of materials for pattern formation, by selecting suitable molecular

architecture and deposition conditions. The density of the polymer at the surface can also be controlled by the choice of adsorbing block size and deposition conditions, while the properties imparted to the surface can be modified by the choice of free block. For instance, an amphiphilic polyelectrolyte diblock, polystyrene-*b*-poly(4-vinylpyridine) diblock ionomer, can form stable surface structures at the air/water interface. Subsequent transmission electron microscopy (TEM) measurements provided direct evidence of the self-assembly of the diblock copolymers into regular circular surface nanostructures [53] that consisted of a central core of polystyrene chains, from which radiated the ionic poly(vinylpyridium) chains. The distance between nanostructures can be controlled by adjusting the surface pressure. Polyelectrolyte diblocks have also been used to control the spacing between micelles by selection of polymer size, charge, and asymmetry [54]. In addition, nonionic diblock polymers of comparably sized hydrophobic and hydrophilic units, such as polystyrene-poly(*n*-butylmethacrylate), polystyrene-polydimethylsiloxane, and polystyrene-*b*-poly(ethylene oxide) (PS-PEO), can form uniform arrays on solid substrate by LB deposition (Figure 8.5) [55–57].

Similarly, the change in monolayer composition may alter the pattern formation. For instance, when a semifluorinated alkane was blended with a PS-PEO diblock copolymer, a surface nanoscale pattern was obtained which resembled a honeycomb with a hump at the center, with a periodicity of ~40 nm [58]. The same qualitative morphological features were found in all mixed films, independent of the polymer grafting density, while the ordering was increased with the increasing polymer grafting density. These structures arose from the organization of the semifluorinated alkane molecules segregated to the surface of the polymer layer. Other examples included the blends of polystyrene-*b*-poly(ferrocenylsilane) (PS-PF) and polystyrene-*b*-poly(2-vinylpyridine) (PS-P2VP) monolayers, the morphologies of which were distinct from those formed when either of the copolymers was spread alone [59]. Pure PS-P2VP was seen to form a highly ordered hexagonal lattice of spherical

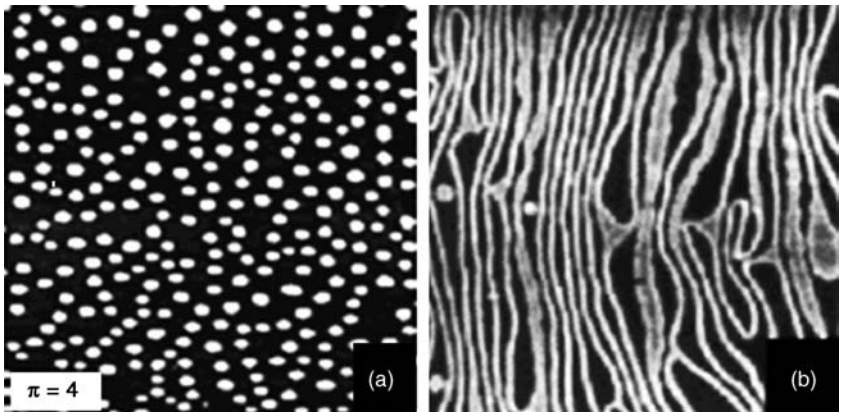


Figure 8.5 The "dots and spaghetti" morphology of a PS-PEO monolayer on a silicon substrate, depending on the transferred conditions by LB deposition. Panel (a) reprinted with permission from Ref. [56]; panel (b) reprinted with permission from Ref. [57].

micelles, whereas pure PS-PF showed three-dimensional (3-D) aggregates with some spherical micelles of irregular size. In the case of blends of these two copolymers, as the fraction of PS-PF increased the morphologies changed from an hexagonal micelle lattice to a cylindrical shape. The application of an electric field in the plane of the air/water interface also caused the structures to compact further and to produce a mesh.

8.3.2
Pattern Formation During LB Transfer

In addition to lateral structures being formed directly at the air/water interface and then transferred onto solid substrates, the LB transfer process itself can be used form patterns close to the three-phase contact line from an homogeneous Langmuir monolayer. As an example, L-α-dipalmitoylphosphatidylcholine (DPPC) (the chemical structure is shown in Figure 8.6) shows how the pattern can be formed during LB transfer, and how the shape, size, and alignment of patterns can be controlled.

DPPC, which constitutes one of the major lipid components of biological membranes, demonstrates the typical phase behavior of a Langmuir monolayer at the air/water interface. This is characterized by a liquid-expanded (LE) phase, a liquid-condensed (LC) phase, and a LE ↔ LC phase transition, as confirmed by the surface pressure–molecular area (π–A) isotherm and BAM images (see Figure 8.6) [35, 36].

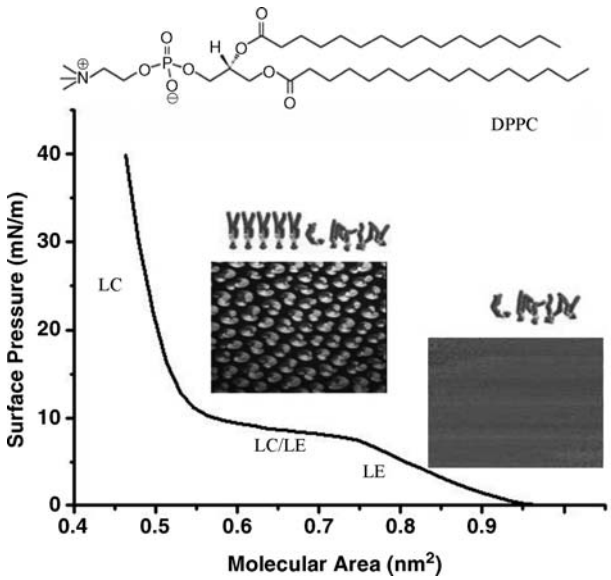


Figure 8.6 Phase behavior of the DPPC monolayer at the air/water interface. Top: Chemical structure of DPPC. Bottom: π–A isotherm of DPPC (~23 °C) and typical BAM images (430 × 537 μm²) for the LE phases and LE ↔ LC phase transition, along with the corresponding conformations of the DPPC molecules. Reproduced with permission from Ref. [23].

In the LE phase, the DPPC monolayer behaves as a quasi-2-D liquid, with the headgroups of the DPPC molecules being translationally disordered and the chains conformationally disordered. On reducing the molecular areas, however, the DPPC molecules begin to condense such that a coexisting phase of LE and crystalline LC occurs at the plateau region of the isotherm. Finally, a homogeneous well-packed condensed monolayer (the LC phase) appears at smaller molecular areas.

When a solid substrate was used to transfer a homogeneous DPPC Langmuir monolayer at the LE phase, a mesostructure which consisted of alternating stripes about 800 nm wide, separated by channels of about 200 nm width, was observed on the mica surface (Figure 8.7b) [60]. Although it is difficult to observe directly the stripe formation *in situ* at the three-phase contact line in this system, it is possible to imagine the process of stripe pattern formation, as depicted schematically in Figure 8.7a. Here, the height difference between the stripes and channels was about 1 nm, and the stripes were composed of condensed (LC phase) DPPC molecules. Considering that the length of a DPPC molecule is about 2 nm, the material in the channels could be attributed to the expanded (similar to LE phase) DPPC molecules, which have a larger tilt angle compared to condensed DPPC molecules in the stripes, as depicted in Figure 8.7c. The origin of the pattern formation was considered due to phase transitions (i.e., substrate-mediated condensation) close to the three-phase contact line during the LB transfer process [60, 61]. During the transfer, a dewetting instability in the vicinity of the three-phase contact line, or meniscus oscillation, caused a switch of DPPC between the expanded phase (the channels) and the condensed phase (the stripes). One possible mechanism for such a switch upon transfer was an oscillation of the meniscus height (i.e., stick–slip model), which

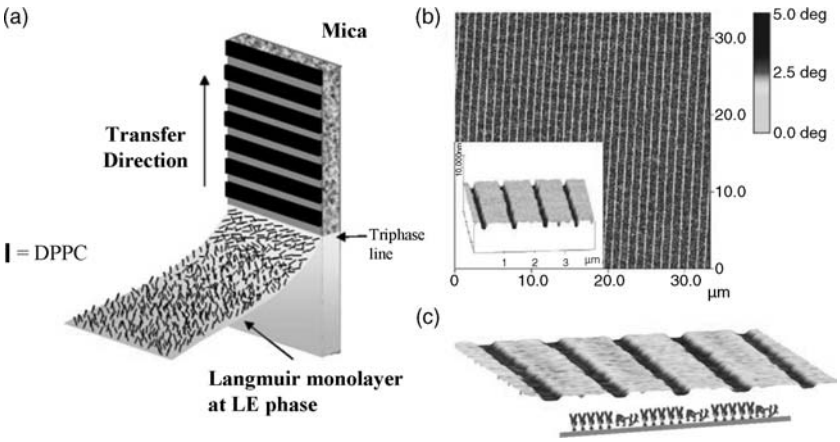


Figure 8.7 (a) Schematic illustration of the process of mesopattern formation; (b) Mesostructures with nanochannels on mica in phase (main figure) and topography (inset) imaging. Experimental conditions: surface pressure 3 mN m^{-1}, transfer velocity 60 mm min^{-1}, temperature 22.5 °C; (c) The composition of DPPC pattern. The DPPC stripe pattern is composed of expanded DPPC molecules in the channels and condensed DPPC molecules in the stripes. Reproduced with permission from Ref. [60].

correlated to the change in interfacial free energies in order to satisfy the Young–Laplace condition [61]. A second possible explanation was a density oscillation in the vicinity of the three-phase contact line.

Importantly, the size and shape of the DPPC patterns can be controlled simply by adjusting the transfer velocity, the surface pressure, the temperature, the substrate chemistry and monolayer composition, and the transfer method. For example, the shape and lateral size of the DPPC stripe pattern from the pure DPPC monolayer depended heavily on the transfer surface pressure and transfer velocity [47, 62]. On mica substrates, at a surface pressure of 3.0 mN m^{-1}, a high transfer velocity of 60 mm min^{-1} induced the formation of horizontal DPPC stripes, parallel to the three-phase contact line (Figure 8.8a and d). In contrast, vertical stripes, perpendicular to the three-phase contact line (Figure 8.8c and f), were obtained at a low transfer velocity (10 mm min^{-1}). At a transfer velocity of 40 mm min^{-1}, a grid pattern that clearly showed the superposition of horizontal stripes and vertical stripes was observed (Figure 8.8b). In general, the horizontal stripes appeared only at the high transfer velocity (60 mm min^{-1}) with a low transfer surface pressure, whilst the pure vertical stripes appeared only at the low transfer velocity and high transfer surface pressure (still in LE phase). Based on such transfer velocity-dependent pattern formation [62, 63], a simple but novel method – termed *LB rotating transfer* – was

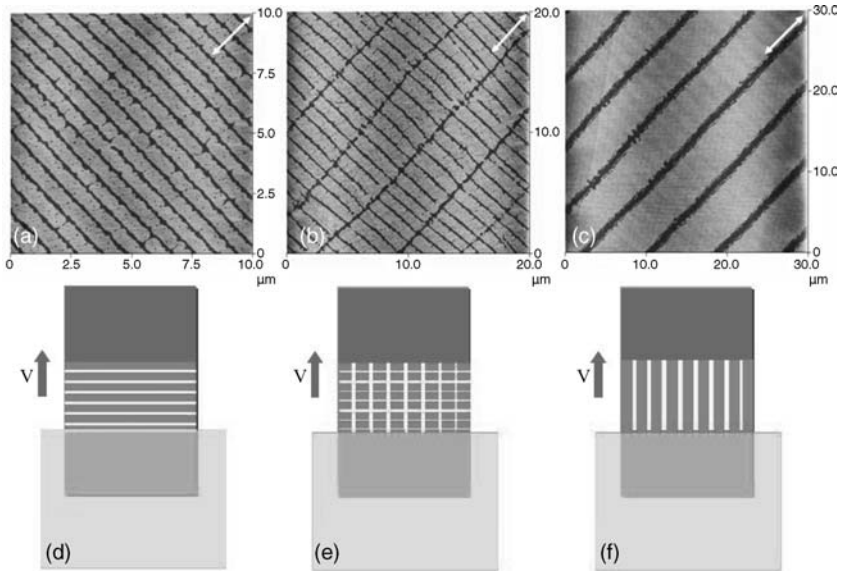


Figure 8.8 The shape and alignment of patterns (pure DPPC) depending on the transfer conditions. (a–c) AFM images of the various pure DPPC patterns on mica surfaces. (a) 60 mm min^{-1} and 3 mN m^{-1}; (b) 40 mm min^{-1} and 3 mN m^{-1}; (c) 10 mm min^{-1} and 3 mN m^{-1}. Double arrows in the AFM images show the axis of film transfer; (d–f) Schematic illustrations for the formation of various patterns during the LB vertical deposition. Reproduced with permission from Ref. [62].

developed to achieve a gradient mesostructure in a well-ordered fashion over large areas [64].

The different hydrophilic substrates can also be used to obtain DPPC mesostructures, although the experimental conditions required for pattern formation will vary due to the different surface properties [65]. One reason for stripe pattern formation is the substrate-mediated condensation of DPPC during the LB transfer; consequently, the molecule–substrate interaction should represent a very important factor in this dynamic self-organization process. For instance, whilst the periodic stripe patterns could be formed on an oxygen plasma-treated silicon surface, the transfer velocity used would need to be slower than that used for transfer onto a mica surface at the same surface pressure and temperature [66].

The addition of a second component to the DPPC monolayer can permit the tuning of DPPC pattern formation, since the miscibility of the various components is important with regard to the phase behavior and stability of the mixed monolayer. For instance, 1,2-di(2,4-octadecadienoyl)-*sn*-glycero-3-phosphocholine (DOEPC) has been selected as an additive component to study the effects of the second component on DPPC pattern formation during LB deposition. This is based on the fact that DOEPC has a similar molecular structure to DPPC, but forms a fully LE phase at the air/water interface under the same conditions. Compared to the pure DPPC monolayer, pattern formation with the mixed monolayer of DPPC/DOEPC (1 : 0.1) shifted to lower velocities and higher surface pressures, while the ability to form horizontal stripes was increased [62]. The grid pattern appeared only at a low transfer velocity (1 mm min^{-1}) and high transfer surface pressures. In general, the size of stripes in the mixed DPPC/DOEPC (1 : 0.1) patterns was about four- to sixfold smaller than that of stripes formed by a pure DPPC monolayer under the same transfer conditions [62]. Other molecules, such as 4-(dicyanomethylene)-2-methyl- 6-(4-dimethylaminostyryl)-4*H*-pyran (DCM) and 2-(12-(7-nitrobenz-2-oxa-1,3-diazol-4-yl)-amino)dodecanoyl-1-hexadecanoyl-*sn*-glycero-3-phosphocholine (NBD), can also be used to generate regular and tunable luminescent stripes with submicrometer-scale lateral dimensions [67]. These dye molecules are uniformly distributed within the expanded DPPC channels, which are in turn separated by condensed DPPC stripes. The width and periodicity of the luminescent stripes can be controlled by adjusting the ratio of dye to DPPC.

8.4
LB Patterning of Nanomaterials

Recent progress has been reported on the close-packed monolayer fabrication of ligand-stabilized nanomaterials on solid substrates [68–74], as one of the most appealing features of the LB technique is the intrinsic control of the internal layer structure down to a molecular level, and the precise control of the resultant film thickness. Unlike these traditional close-packed nanoparticle monolayers on solid substrates, the LB technique itself represents a means of obtaining regular nanoparticles or nanowire pattern arrays on solid substrates [75].

8.4.1
LB Patterning of Nanoparticles

As an example, Heath and coworkers [76] confirmed the formation of aligned, high-aspect ratio nanowires at a low-density Langmuir monolayer film of alkylthiol-passivated silver nanoparticles during film compression. Prior to monolayer compression (surface coverage ~20%, $\pi \approx 0\,\mathrm{nN\,m^{-1}}$), the particles were found to aggregate into circular domains. However, after compression the particle monolayers assembled spontaneously into lamellae or wire-like superstructures with lengths of several micrometers and widths of 20 to 300 nm, which were functions of the solvent and the particle size. The interwire separation distance, as well as the alignment of the wires, could be controlled via compression of the wires. There was, in addition, an agreement between these experimentally acquired data and a computer simulation performed using the standard Metropolis Monte Carlo algorithm [77] (see Figure 8.9). Here, the patterns were described as resulting from competition between an attraction, which makes the particles aggregate, and a longer-ranged repulsion, which limits the aggregation to finite domains. Careful investigations also showed

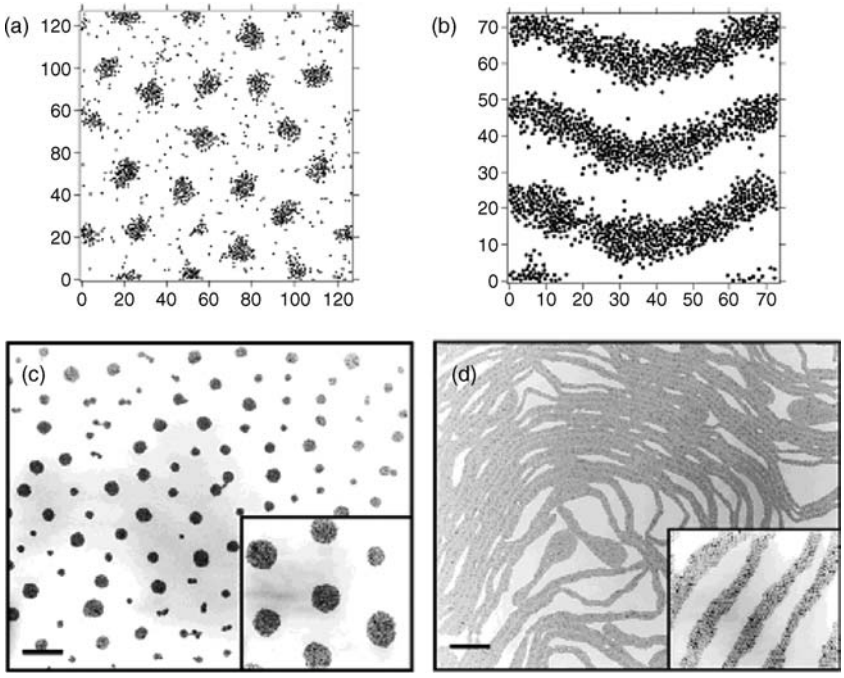


Figure 8.9 Results of the computer simulation (a, b) and the corresponding transmission electron microscopy images (c, d) revealing the spontaneous formation of clusters and stripe-like arrays of alkylthiol-passivated Ag nanocrystals. The solution used to prepare (d) was approximately threefold more concentrated ($\sim 1\,\mathrm{mg\,ml^{-1}}$) than that used for (c). Scale bar = 0.5 μm. Reproduced with permission from Ref. [77].

that an increase in concentration led to a spontaneous reorganization of the self-assembled domains from circular clusters to stripes, as the repulsions between the aggregates became more important than those between the individual particles within them. This phenomenon was considered to be closely related to the transitions to hexagonal and lamellar phases commonly observed in concentrated surfactant solutions, where the locally preferred curvature of micelles is successively "squeezed out" of the systems as the interaggregate repulsions become dominant and the lower-curvature cylinder and bilayer geometries are found to better minimize the overall interaction free energy.

In contrast to the results of Heath *et al.* [76], where higher-order nanoparticulate structures were formed at the air/water interface before being transferred onto a solid substrate, Schmid and Yang *et al.* found the process of LB transfer also to be an efficient means of obtain regular nanoparticle arrays on solid substrates, with the assistance of dewetting during the LB transfer process [78–80]. Schmid *et al.*, by using the LB technique [80], first successfully obtained parallel rows of $Au_{55}(PPh_3)_{12}Cl_6$ clusters, which are quasi-one-dimensional (1-D) structures of quantum dots of about 10 nm width. A modified LB technique (Figure 8.10a), deposited beneath the monolayer at an angle of 20°, was used to generate this type of cluster stripe. Pattern formation was shown to depend mainly on the speed at which the substrate was moved; for example, at speeds of about 10 cm min^{-1} the parallel stripes consisted of three to four cluster rows and were separated one from another by 8 nm (Figure 8.10). The formation of such patterns was attributed to oscillation of the water meniscus at

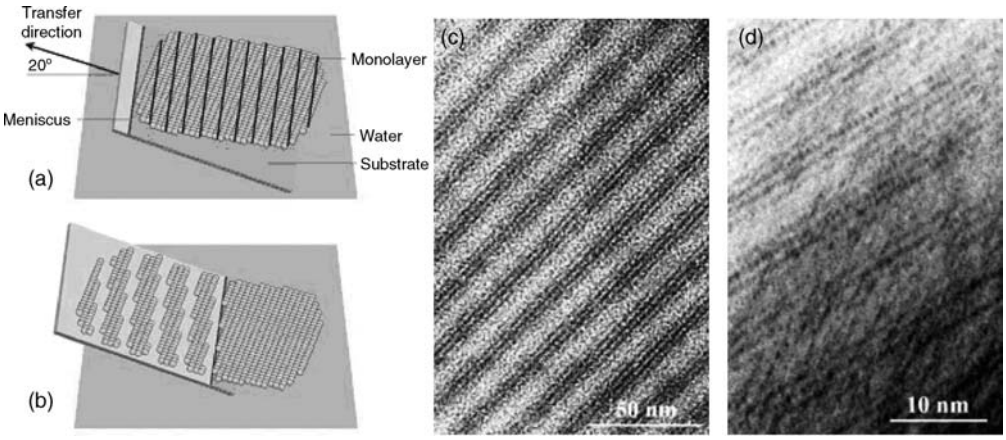


Figure 8.10 (a) Sketch of the formation of cluster stripes from an ordered monolayer. The monolayer is oriented toward the substrate edge and the meniscus, respectively, by a nonpredetermined angle; (b) Owing to the movement of the substrate from the water, and the herewith linked transfer of the monolayer onto the substrate surface, the monolayer is fractured along the black lines due to oscillation of the meniscus. Stripes of three to four rows of clusters lying side by side are formed. The stripes run parallel to the water meniscus; (c) TEM image of cluster stripes consisting of three to four cluster rows; (d) Magnified cutout. The cluster rows consist of equidistantly ordered clusters. Reproduced with permission from Ref. [80].

the substrate, which induced the generation of striped patterns that ran parallel to the meniscus.

Later, Yang *et al*. [79] used the LB technique to generate well-spaced, parallel single particle lines on a substrate from a dilute Langmuir particle monolayer via a stick-slip motion of the water/substrate contact line. In this case, a stick-slip motion was observed *in situ* by optical microscopy, with the three-phase contact line during the transfer process being due to the large interline distance and low density of the Langmuir monolayer at the air/water interface, when compared to the data of Schmid *et al*. [80]. The particle density within the lines could be controlled not only by the particle concentration in the monolayer but also by the pulling speed of the substrate. In this way, lines of a wide variety of materials and sizes, ranging from a few nanometers to a few micrometers, were demonstrated. The ability to assemble nanoparticles into 1-D arrays enables the construction of higher hierarchical device structures. For example, by using gold nanoparticle seeds it is possible to grow vertical single nanowire arrays of silicon, replicating the pattern of single particle lines. The spontaneous formation of ordered gold and silver nanoparticle stripe patterns was identified on dewetting a dilute film of polymer-coated nanoparticles floating on a water surface [78]. However, the difference here was that the nanoparticle stripe patterns were perpendicular to the air/water interface (Figure 8.11), in contrast to the above two examples. The reason for such formation of vertical nanoparticle stripe patterns was considered to be the fingering instability. Taken together, these results showed that the LB technique can provide new avenues for the lithography-free patterning of nanoparticle arrays in a variety of applications, including multiplexed surface-enhanced Raman substrates and the templated fabrication of higher-order nanostructures.

One unique property of the LB technique is that it can control the monolayer composition. To some extent, the morphology of these nanostructures can be controlled by adjusting the parameters that affect the self-assembly process. For example, Hassenkam *et al*. demonstrated the formation of continuous gold nanowires by mixing and spreading the dodecanethiol-capped gold nanoparticles and DPPC at the air/water interface [81]. The unidirectional sintering of particles, which was accompanied by packing into a maze-like structure, was considered due to a template effect of the surfactant at the molecular level. In this case, the amphiphilic DPPC molecules preferred (on an energetic basis) to occupy the entire water surface if left alone, whereas when the hydrophobic gold particles were left alone on the water surface they would form close-packed, floating, 2-D hexagonal rafts. Yet, if a mixture of DPPC and dodecanethiol-capped gold particles were to be placed on the same water surface, the energetic strain between the bare water surface and the hydrophobic particles would be reduced. Since the DPPC molecules can only support single-particle broad lines, this would result in the formation of 1-D aggregates, which is in fact the mechanism of nanowire formation. When Zhang *et al*. described the use of molecular aggregates as templates to assemble water-soluble nanocrystals into branched wire structures at the air/water interface [46], they designed and synthesized a chiral amphiphilic molecule, C12-(L)Cys-(L)Cys-C18, which consisted of two short cysteine peptides as the hydrophilic heads and two hydrophobic alkyl

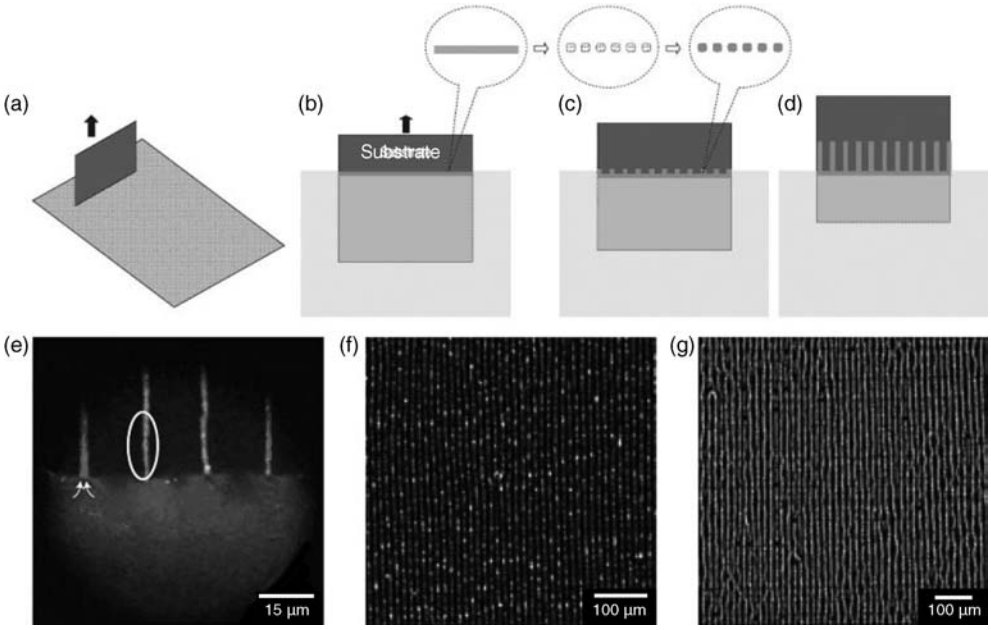


Figure 8.11 Extended stripe pattern formation through dip-coating. (a–d) A schematic drawing illustrating the formation of an aligned gold nanoparticle stripe pattern by vertical deposition (a, b). Only the nanoparticles at the water/substrate contact line (gold dots in b–d) are shown for clarity. The substrate is raised slowly (a, b) so that water is evaporated when a new surface is exposed. The wet contact line containing uniformly dispersed nanoparticles breaks up into aggregates of nanoparticles (b, c), owing to the fingering instability during the initial dewetting stage. These fingertips then guide the further deposition of nanoparticles, finally forming the extended stripe pattern (d); (e) Direct optical microscopy observation of the water front reveals a rapid motion of nanoparticles towards the wet tips (circled area) of the stripes, as indicated by the arrows. This leads to the unidirectional growth of the stripes across the entire substrate as shown in the optical microscopy image in (f); (g) Silver nanoparticle stripes have been obtained in the same fashion. Reproduced with permission from Ref. [78].

chains as the tails, and these formed chiral domains at the air/water interface. This lateral structure, with its chemically active end-groups (thiol groups), was further used for the specific binding of CdTe nanocrystals. After transferring a monolayer of C12-(L)Cys-(L)Cys-C18 via the complexation of CdTe nanocrystals, CdTe nanowires were produced that were 10–15 nm wide, up to several micrometers long, and were branched in a certain fashion.

Polymers may also serve as an important component for fine-tuning the formation of nanoparticle arrays, if there is sufficient attraction between the ligand molecules and the polymer. For instance, poly(vinyl-pyrrolidone) (PVP), which is able to chemisorb $Au_{55}(PPh_3)_{12}C_{l6}$ clusters via the phenyl groups in impressive manner, was added into the subphase to tune the nanoparticle array formation [82]. In the absence of PVP, smaller islands of well-ordered Au_{55} were also formed at the air/water phase

boundary [83]. The wires (which were 30 nm wide and 1 μm long) were connected by junctions of cluster islands to a complete 2-D network on mica or silicon that had been generated via the LB technique when PVP was added. The pattern of cluster-coated polymer molecules indicated that the nanoparticles had acted partly as linking knots between the polymer chains, so as to generate a stable network. Lu *et al*. [84] also used model electrodes fabricated via nanosphere lithography [85] to connect the nanowires of Au_{55}. In this case, the model electrodes were prepared via metal evaporation through a mask of monodispersed latex beads. At the second stage, the silicon surfaces bearing the model electrodes were used as substrates for transferring nanowires that consisted of Au_{55} and had been prepared via the LB technique on the PVP subphase. In this way, by controlling the structure density on the surface, it was possible to obtain both single connections (as shown in Figure 8.12) and multi-connections.

In addition to linear metal nanoparticle arrays or 2-D networks, ring-like CdSe nanoparticle patterns [86] and tree-like fractal aggregates of CdS nanoparticles in amphiphilic oligomers [87] were also observed. The ring-like structures had diameters ranging from 150 to 1200 nm, and were obtained by transferring a mixed monolayer of amphiphilic copolymer poly[(maleic acid hexadecylmonoamide)-*co*-propylene] and CdSe nanoparticles stabilized with polystyrene-poly(4-vinylpyridine) onto solid substrates, using the LB technique [86]. Due to preferential interactions between the polystyrene-functionalized nanoparticles and the polystyrene block of an amphiphilic PS-PEO block copolymer, a highly stable 1-D nanoparticle/polymer was formed at the air/water interface, via synergistic self-assembly, with surface features that included branched nanowires and nanocables up to 100 μm in length [88].

8.4.2
LB Patterning of Nanowires

One-dimensional nanoscale building blocks, such as nanowires, nanorods, and carbon nanotubes (CNTs), can also be ordered and assembled rationally into

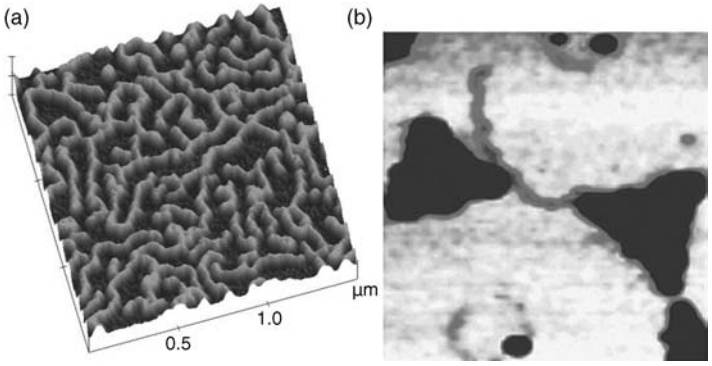


Figure 8.12 Topographical image of network structures of Au_{55} on mica surface ($1.6 \times 1.6\,\mu m^2$) and nanowires of Au_{55} connected with model electrodes ($350 \times 350\,nm^2$). Reproduced with permission from Ref. [84].

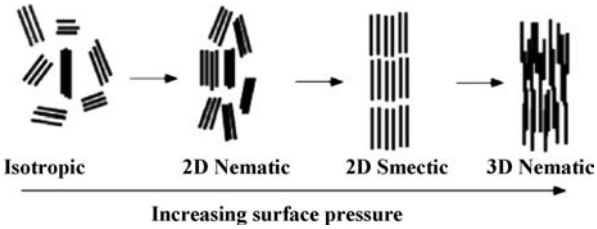


Figure 8.13 Schematic illustration of the pressure-induced phase transition when the nanorods are compressed at the water/air interface. Reproduced with permission from Ref. [89].

appropriate 2-D architectures using the LB technique. Yang *et al*. supported this proposal by using $BaCrO_4$ nanorods, with a low aspect ratio of ~3–5 : 1 and a typical diameter of approximately 5 nm [89] in the pressure-induction of isotropic-2-D → nematic-2-D → smectic-3-D nematic phase transitions, as well producing a transformation from monolayer to multilayer nanorod assembly (as shown in Figure 8.13). At low surface pressure, the $BaCrO_4$ nanorods formed raft-like aggregates (i.e., an isotropic state) that comprised generally three to five rods, with the rods aligned side-by-side due to the effects of directional capillary forces and van der Waals attractions. During the process of compression ($<30\,mN\,m^{-1}$) a monolayer of nanorods was formed in a nematic arrangement, with an orientational order parameter S of 0.83, where the directors of the nanorods were aligned qualitatively, presumably dictated by the barrier of the trough. When the surface pressure was raised to about ~35 $mN\,m^{-1}$, nanorod assemblies with a smectic arrangement are obtained, whilst when the pressure exceeded $38\,mN\,m^{-1}$ there was a transition from monolayer (ordered 2-D smectic arrangement) to multilayer (disordered 3-D nematic configuration).

The thiol-capped gold nanorods (diameters ~8 nm) with similar aspect ratios showed a great tendency to form nanorod ribbon superstructures, where many of the gold nanorods are aligned side-by-side [90]. Compression of these nanorod monolayers did not lead to the same phase evolution as seen in the $BaCrO_4$ system, a difference that might be attributed not only to the much more attractive van der Waals forces and directional capillary interaction among gold nanorods when compared to $BaCrO_4$ nanorods, but also to the polydispersity of the available gold nanorods. In contrast, the organization of $BaWO_4$ nanorods (diameter ~10 nm) with a large aspect ratio (150 : 1) again differed significantly from the assembly superstructures of the short $BaCrO_4$ and Au nanorods [91]. Initially, the nanorods were rather dispersed, with the directors of the nanorods being distributed isotropically and no superstructures being observed. After compression, however, the nanorods were readily aligned in roughly the same direction to form a nematic layer such that, with a strong compression they formed bundles that had almost perfect side-by-side alignment between the included nanorods. The preference for a nematic phase formation upon compression proved to be a distinct characteristic of the assembly behavior of nanorods with a large aspect ratio, a situation also identified for the alternative molecular wire system of $Mo_3Se_3^-$ (diameter 0.8 nm, aspect ratio effectively infinite).

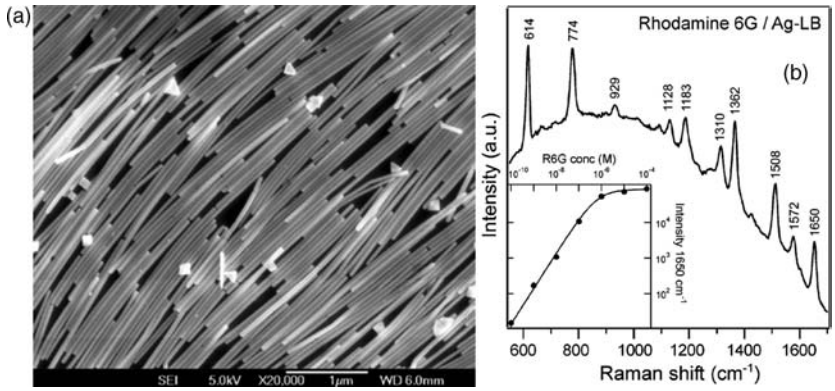


Figure 8.14 (a) Scanning electron microscopy images of the silver nanowire monolayer deposited on a silicon wafer; (b) Surface-enhanced Raman scattering (SERS) spectrum of rhodamine 6G (R6G) on the thiol-capped Ag-LB film (532 nm, 25 mW) after 10 min incubation in a 10^{-9} M R6G solution. The inset shows the linear relationship between the Raman intensity at 1650 cm^{-1} and the R6G concentration. Reproduced with permission from Ref. [92].

Xia and Yang *et al.* also used the LB technique successfully to assemble monolayers (with areas $>20\,cm^2$) of aligned silver nanowires that were $\sim$50 nm in diameter and 2–3 μm long [92]. These nanowires (which had pentagonal cross-sections and pyramidal tips) were close-packed as parallel arrays, with their longitudinal axes aligned perpendicular to the compression direction (see Figure 8.14). The monolayers, which were readily transferred onto any desired substrate, included silicon wafers, glass slides, and polymer substrates, and could serve as simple wire-grid optical polarizers and surface-enhanced Raman spectroscopy (SERS) substrates. The monolayer substrates were shown to behave as major enhancers of electromagnetic fields (factors of 2×10^5 for thiol and 2,4-dinitrotoluene, and of 2×10^9 for Rhodamine 6G), and could be readily used in ultrasensitive, molecule-specific sensing processes by utilizing vibrational signatures. Furthermore, the fact that the observed SERS intensity depended on the polarization direction confirmed the (theoretical) predictions that large electromagnetic fields would be localized within the interstices between adjacent nanowires [93].

Similarly, silicon nanowires can be organized into aligned structures by the LB technique, over large areas [94]. For this, the aligned nanowires were first transferred onto planar substrates via a layer-by-layer process so as to form parallel and crossed nanowire structures (Figure 8.15) that were then transferred onto a substrate. Photolithography was then used to define a pattern over the entire substrate surface, which set the array dimensions and array pitch, after which any nanowires outside the patterned array were removed by gentle sonication. In addition, electrical transport measurements that exhibited linear current versus voltage behavior confirmed that reliable electrical contacts could be made to the hierarchical nanowire arrays prepared via this method. This process offered a flexible pathway for the

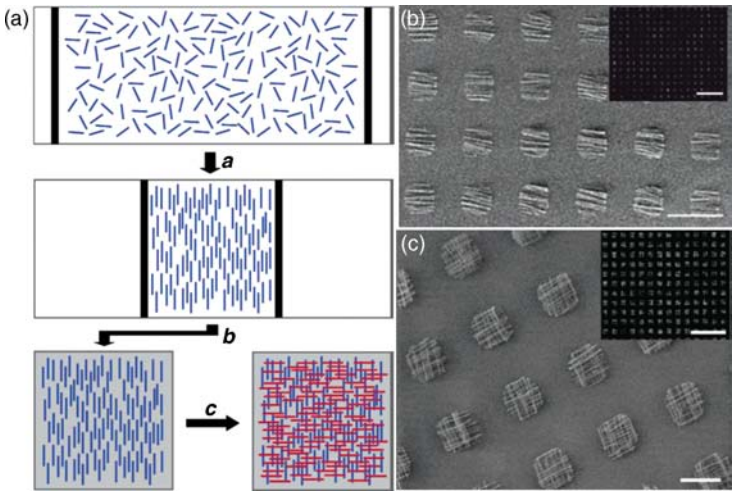


Figure 8.15 (a) Nanowires (blue lines) in a monolayer of surfactant at the air/water interface are compressed (pathway *a*) on a Langmuir–Blodgett trough to a specified pitch. In pathway *b*, the aligned nanowires are transferred to the surface of a substrate to make a uniform parallel array. In pathway *c*, crossed nanowire structures are formed by uniform transfer of a second layer of aligned parallel nanowires (red lines) perpendicular to the first layer (blue lines); (b) Image of patterned $10\,\mu m \times 10\,\mu m$ parallel nanowire arrays. Scale bar $= 25\,\mu m$. The inset shows a large-area dark-field optical micrograph of patterned parallel nanowire arrays (inset scale bar $= 100\,\mu m$); (c) Scanning electron microscopy image of patterned crossed nanowire arrays. Scale bar $= 10\,\mu m$. The inset shows a large-area dark-field optical micrograph of the patterned crossed nanowire arrays (inset scale bar $= 100\,\mu m$). Reproduced with permission from Ref. [94].

bottom-up assembly of virtually any nanowire material into the highly integrated and hierarchically organized nanodevices required for a broad range of functional nanosystems. For example, crossed nanowires arrays might be used as an addressable nanoscale LED source.

8.5
Application of Structures Formed by LB Patterning

8.5.1
Templated Self-Assembly of Molecules and Nanoparticles

As discussed above, the DPPC pattern is composed of expanded DPPC molecules in the channels, and condensed DPPC molecules in the stripes. This chemically striped pattern shows an anisotropic wetting of 1-phenyloctane [95], due to the different interfacial energies for the channels ($\sim$31 mJ m^{-2}) and stripes ($\sim$23 mJ m^{-2}) [96]. As a result, this type of mesostructured surface can be used as a template to guide the self-assembly of molecules and nanoparticles.

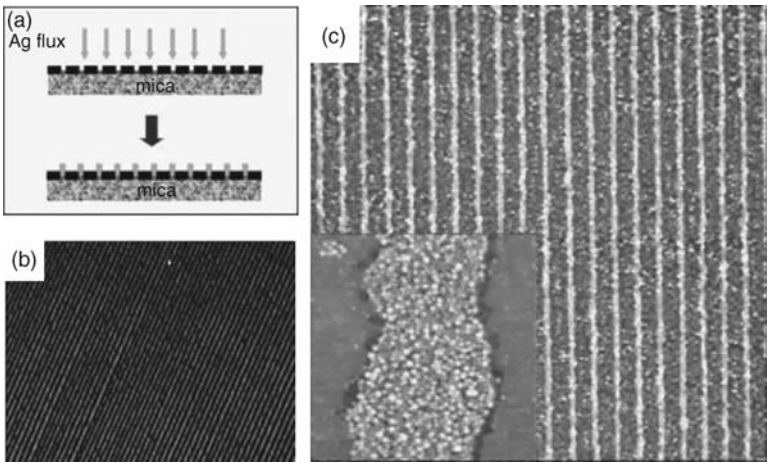


Figure 8.16 (a) Schematic illustration of the process for evaporating silver on structured surface and silver atoms deposit preferably onto channel regions; (b) Optical micrograph representing the regular stripe structure over an area of $80 \times 60\,\mu m^2$. The channels were filled with more silver coating (bright lines), whereas the DPPC stripes appeared dark (less silver); (c) The AFM image ($16 \times 16\,\mu m^2$) indicates that if a small amount of silver is evaporated (<2 nm), then only the channel regions are fully covered, as shown in the inset. Reproduced with permission from Ref. [95].

For example, the $FeCl_3$ molecules which condensed from the vapor phase were adsorbed selectively in the channels, whereas the stripes were not coated when a small droplet of $FeCl_3$ solution was brought onto the structured mica surface. Channels filled with paramagnetic $FeCl_3$ molecules provided a contrast for magnetic force microscopy [60]. In another example, the selective adsorption of thermally evaporated silver (2–3 nm) onto the channels was confirmed using optical microscopy [95] (Figure 8.16). In addition to metals and small molecules that show selective adsorption, Moraille and Badia found that proteins could also be adsorbed selectively onto the nanostriped surface formed by the mixed monolayer of DPPC and L-α-dilauroylphosphatidylcholine (DLPC) [97]. Moraille and Badia confirmed that human blood-plasma proteins (γ-globulin and serum albumin) could be adsorbed selectively to the channels of a nanostructured LB monolayer of DPPC/DLPC, so as to generate well-defined protein and Au nanoparticle/protein patterns. The DPPC mesostructures on oxygen plasma-treated silicon could be used as templates for the directed self-assembly of functional silane molecules to form robust chemical patterns [66]. In this case, a general approach was based on a substitution of the channels and stripes by two different silane molecules (NH_2- and CH_3-terminated silane) that were bound covalently to the surface. As a result, a striped pattern of covalently bound molecules with selective functionality replaced the physisorbed DPPC structure, after which the negatively charged Au_{55} clusters could be adsorbed selectively onto the NH_2-terminated silane stripes, due to an electrostatic interaction [66]. Moreover, the NH_2-terminated silane-striped pattern could be used as a template to assist the electrodeposition of regular arrays of copper nanowires [98].

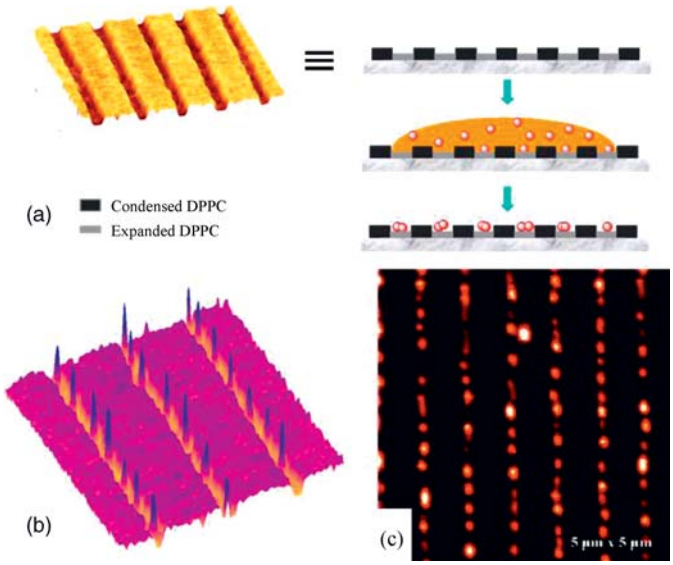


Figure 8.17 (a) Generalized schematic outline of the three steps used to pattern nanoparticles on DPPC stripe pattern. Selective deposition of (b) Au_{55} clusters and (c) CdSe nanocrystals aligned along the channels on a mica surface. Reproduced with permission from Ref. [23].

The DPPC stripe pattern may also serve as a template for the selective deposition of nanoparticles, simply by dropping the 1-phenyloctane solutions of nanoparticles onto the DPPC pattern, as shown in Figure 8.17. The work of adhesion of 1-phenyloctane on the channels was 62.0 mJ m^{-2}, and greater than that of 1-phenyloctane on the stripes (53.7 mJ m^{-2}). As a result, the nanoparticles were found to accumulate in the expanded DPPC channels when the solution was removed from the sample surface after some time. The density of nanoparticle coverage was determined by the concentration of the nanoparticle solution and the duration of exposure to the patterned surface. As an example, quasi 1-D arrays (Figure 8.17b) of Au_{55} clusters stabilized by an organic ligand shell were generated [60]. Semiconductor nanocrystals showed a similar selective adsorption in the channels, as demonstrated by topographic and near-field optical fluorescence measurements (Figure 8.17c) [99, 100]. These examples showed, principally, that nanoparticles could be arranged in 1-D fashion in parallel manner over large areas. Furthermore, the CdSe nanocrystals could be selectively deposited into the green-emitting stripes formed by transferring mixed monolayers of DPPC and 2-(4,4-difluoro-5-methyl-4-bora-3a,4a-diaza-s-indacene-3-dodecanoyl)-1-hexadecanoyl-*sn*-glycero-3-phosphocholine (BODIPY) (0.5 mol%) onto mica surfaces, for which BODIPY molecules are distributed uniformly within the expanded DPPC channels [101]. Based on the photoinduced enhancement of fluorescence of CdSe nanocrystals and the photobleaching of dyes, a hierarchical luminescence pattern will be generated.

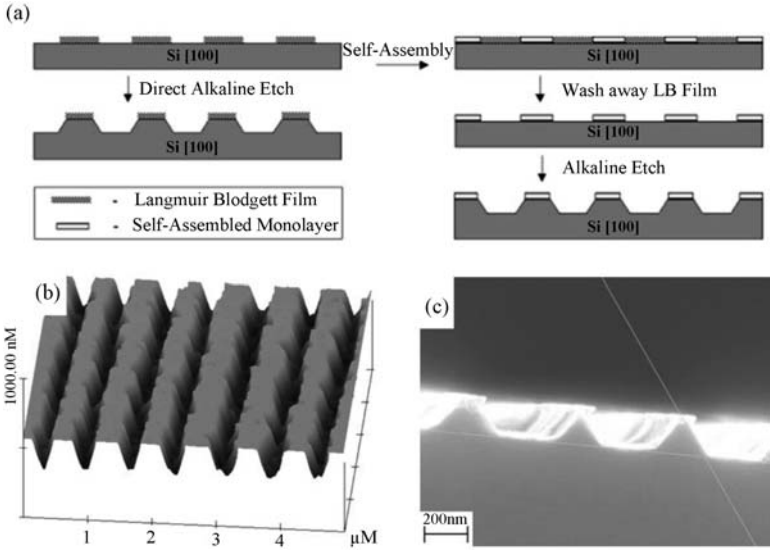


Figure 8.18 (a) Schematic illustration of the chemical-etch process used to transfer LB patterns into topographical features. Groove depth and local periodicity were characterized by (b) AFM, and (c) SEM. Reproduced with permission from Ref. [47].

8.5.2
Pattern Transfer: From Chemical to Topographic Patterns

Both, the self-organized DPPC LB patterns and other LB patterns formed (such as chiral domains) [46, 47] can be used as resistances against wet chemical etching, by employing a very dilute alkaline etchant (e.g., KOH) and a long etching time (~12 h). Such a process, which is referred to as "LB lithography" [47], allows the patterns to be converted into topographic features in silicon (Figure 8.18). In this way, an etch selectivity in excess of 100 (etch depth/resist thickness) can be achieved, while the depth of etching can be controlled at between 20 and 300 nm by varying the etch time.

The topographically patterned silicon obtained with LB lithography can be used as a "master" to generate replicas, by means of nanoimprinting and replica molding (Figure 8.19a) [47, 102]. This is of major interest to certain applications, such as the culturing of biological cells, where it is desirable to mass produce a large number of identical surfaces. In the first step of the process, the silicon master is placed in contact with the polymer under slight pressure, and the system is heated above the glass transition temperature (T_g) of the polymer. After cooling the polymer to below T_g, it is peeled from the master; after which the master can be re-used for the serial production of hundreds of replicas, without any noticeable reduction in quality.

In these studies, surface areas of polystyrene on the order of square centimeters were first topographically patterned using submicrometer-scale grooves, and then used to study the influence of surface texture on the morphology, mobility and differentiation of primary osteoblasts [47, 102]. In this case, cells cultured for 24 h on

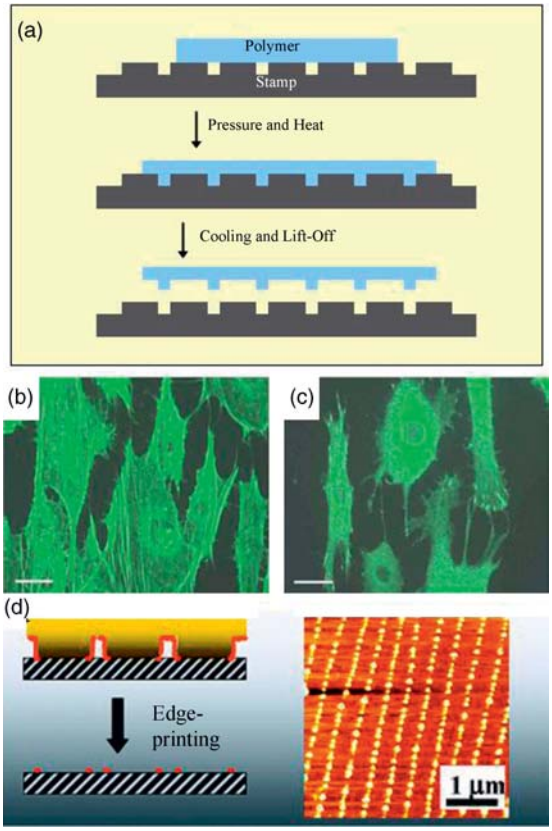


Figure 8.19 (a) Schematic process for pattern transfer from a silicon master to polystyrene. Fluorescence micrographs of osteoblasts aligned on 150 nm-deep grooves, labeled for (b) actin and (c) vinculin. Scale bar = 20 μm; (d) Edge printing of semiconductor nanocrystals by interfacial interaction controlled transport of CdTe nanocrystals. Reproduced with permission from Ref. [23].

grooved polystyrene surfaces with a periodicity of 500 nm were seen to align with the grooves (see Figure 8.19b and c). Notably, the osteoblasts showed a stronger alignment on the deeper grooves, though the numbers of cells attaching to structured surfaces with grooves of different depths (50 nm and 150 nm, and also on a smooth control) seemed to be unaffected by the nanotopography of the surface. Immunohistochemical staining of the aligned cells confirmed the presence of focal adhesions at opposite ends of the aligned cells. A significant anisotropic migration was observed on both the 50 nm- and 150 nm-deep grooves, though to a greater extent on the deeper grooves [102]. In addition to osteoblasts, other types of cell, including the phytopathogenic fungi *Magnaporthe grisea* and *Pucinnia graminis*, were also shown to align on the grooved patterns fabricated by LB lithography [103]. Clearly, the ability to mass produce, on an economic basis, large surface areas patterned with different

nanotopographies will create new opportunities to understand the mechanisms behind contact guidance, and to optimize such surfaces for biological applications.

These structured polystyrene surfaces can also be used as masters for the replica molding of polydimethoxysiloxane (PDMS) [104]. This could be achieved by pouring the PDMS precursors over the polystyrene topographies and curing at 60 °C (well below the T_g of polystyrene) for 2 h; the PDMS stamp could then be readily peeled from the polystyrene master. The main benefit of this two-step process is that the PDMS replica molding can be carried out in parallel, thus allowing a simple and rapid fabrication of numerous, low-cost identical copies. The structured PDMS could then be used for microcontact printing (μCP), for example to pattern CdTe nanocrystals on SiO_2/Si surfaces (Figure 8.19d) [104].

Lieber *et al*. developed a new nanolithographic process based on the LB patterning of aligned nanowires that can be used as masks for etching and deposition, so as to fabricate nanometer-scale lines over large areas (Figure 8.20) [105]. For this, the surfactant-stabilized core–shell nanowires with controlled diameter and shell dimensions were first aligned with nanometer- to micrometer-scale pitches, using the LB technique; they were then transferred onto planar substrates to form uniformly ordered parallel arrays. Following such transfer, reactive ion etching (RIE) with CHF_3 was carried out to remove the oxide shell on the sides and tops of the core–shell nanowires, and to transfer the line pattern to the underlying substrate surfaces. Using the same process, metals can be deposited using the aligned nanowires as shadow masks to create arrays of nanoscale wires. Finally, the nanowire masks are removed by isotropic wet etching and sonication so as to expose the etched or deposited parallel line features. When using this method, the feature sizes are comparable to state-of-the-art extreme UV lithography, and also approach the limits of electron-beam lithography and transfer lithography. The width, length, and pitches of the metal lines can be easily controlled via the synthesis of core–shell nanowires

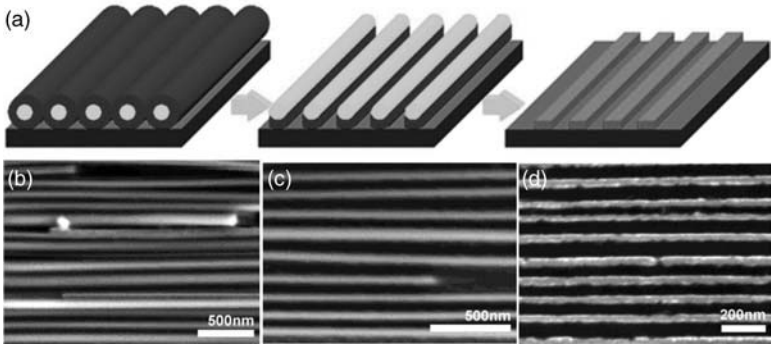


Figure 8.20 (a) Scheme for selective anisotropic etching of the oxide shell of core–shell nanowires and deposition of metal or other materials based on LB patterning nanowires. (b–d) Typical SEM images of (b) close-packed parallel Si-SiO_2 core–shell nanowires on silicon substrate surface; (c) parallel nanowires after selective, anisotropic etching of the SiO_2 shell by reactive ion etching; and (d) 15 nm-thick Cr metal lines following removal of the nanowire mask. Reproduced with permission from Ref. [105].

and a subsequent assembly process. In addition, the nanowires can be assembled in one step over areas of 20 cm^2, which is greater than that possible with most other unconventional lithographic methods. Hierarchical parallel nanowire arrays have also been prepared and used as masks to define nanometer pitch lines in $10 \times 10\,\mu m^2$ arrays, repeated with a 25 μm array pitch over square-centimeter areas. This nanolithographic method represents a highly scalable and flexible route for defining nanometer-scale lines on multiple length scales, and thus has substantial potential for the fabrication of integrated nanosystems.

Using a similar approach, Choi *et al*. employed the structured Ag nanowire patterns [76] formed by the LB technique as resist masks to fabricate a parallel array of poly(methylmethacrylate) (PMMA) wire patterns [106] that were transferred onto PMMA-coated substrates, using a horizontal deposition method. The pattern was then amplified by immersing the substrate in a solution containing decanedithiol ($HSC_{10}H_{20}SH$), followed by immersion of the substrate into a hexane/nanoparticle solution. Such amplification caused a slight increase in the width of the wires, and doubled their height to about 8 nm. Subsequently, 50 nm-wide and 10 nm-high PMMA wire patterns were obtained via spatially selective low-energy electron-beam exposure on the Ag nanocrystal wire shadow mask, and development, and a RIE process was then used to obtain 50 nm-wide silicon wires. This method would appear to represent a low-cost, high-throughput technique for the fabrication of semiconductor, nanometer-scale structures.

As an alternative, Meli *et al*. used the surface pattern formed by a PS-P2VP diblock copolymer as a mask to create an extensive array of nanometer-sized features [107] that were later used as stencil masks to generate quasi-hexagonal 2-D arrays of nanoscale gold islands [108]. These ultrathin masks have an intrinsic topology (which depends on the choice of block copolymer), ranging from about 10 to 100 nm in width and spacing. Straightforward argon ion milling of the gold-coated silicon and mica substrates, which had been covered with the ultrathin masks, resulted in arrays of ~25 nm-diameter gold islands, supported on patterned silicon pillars or gold islands that were directly adhered to a mica substrate. In a similar study, Seo *et al*. used this process to fabricate nickel pattern arrays that could be used to separate DNA [12] and, in particular, to eliminate the need for disposable separation media such as gels or polymer solutions that are susceptible to degradation and difficult to load into small devices due to their inherent high viscosity. By conducting molecular dynamics simulations and experiments, it was shown that this method could simultaneously separate a broad band of DNA fragments, ranging from a few hundred base-pairs to mega-base-pairs, without any loss in resolution. Furthermore, the technique required only very low loading amounts and operating voltages, making it amenable for incorporation into chip-based portable detectors or microarrays.

8.5.3
Integration of Nanomaterial Patterning in Nanodevice Fabrication

One fundamental step in the construction of 1-D nanomaterial devices is transfer of the nanomaterials from their stock to the substrate on top of which the device will be

built. Hence, the correct alignment and controlled positioning of the 1-D nanomaterials are highly desirable, especially with regards to the large-scale (e.g., on a 10-cm wafer) fabrication of parallel device arrays. The integration of LB patterning with device fabrication fits this technological gap.

A general strategy for the parallel and scalable integration of nanowire devices over large areas, without the need to register individual nanowire-electrode interconnects, has been developed by the group of Lieber and involves a combination of the LB technique and photolithography (Figure 8.21) [109]. In this case, organized nanowires with controlled alignment and spacing over large areas were produced using the LB technique [94], and interconnects between the nanowires and electrodes defined by photolithography, in a statistical manner. Because the separation between nanowires assembled with the LB technique had a defined average value, but varied on the local scale, it was possible to achieve a high yield of metal electrode to nanowire contacts simply by setting the average nanowire separation equal to a value that was comparable to the electrode width. In this way, massive arrays containing thousands of single silicon nanowire field-effect transistors (FETs) were fabricated, and shown to exhibit not only a high performance and unprecedented reproducibility, but also a scalability to at least the 100 nm level. Moreover, scalable device characteristics could be demonstrated by interconnecting a controlled number of nanowires per transistor, in "pixel-like" device arrays. It is likely

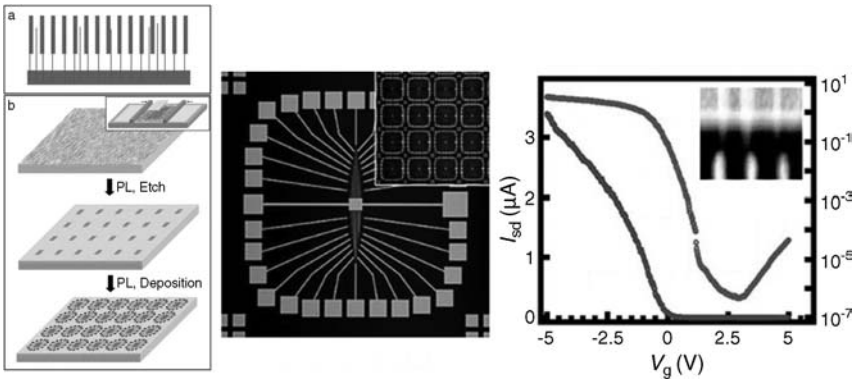


Figure 8.21 Left panel: Parallel and scalable interconnection of nanowire devices without registration. (a) Central electrode region of a single array, emphasizing the high fraction of interconnected nanowires (blue lines) obtained without the registration of individual electrodes; (b) Schematic illustrating key steps of the interconnection approach, including (top) the deposition of aligned nanowires with defined average spacing over the entire substrate, (middle) hierarchical patterning to produce fixed-size and -pitch parallel nanowire arrays, and (bottom) the deposition of a repeating metal electrode array using photolithography. Middle panel: Optical micrograph of integrated metal electrode arrays deposited on top of patterned parallel nanowire arrays defined by photolithography. Right panel: I_{sd} versus gate voltage (V_g) recorded for a typical device plotted on linear (blue) and log (red) scales at a V_{sd} of 1 V. The inset shows SEM images of higher-density nanowire devices, defined by electron-beam lithography. Reproduced with permission from Ref. [109].

that these highly reproducible transistor arrays will find use in many applications, ranging from multiplexed biosensing to information displays. The general applicability of the approach to other nanowire and nanotube building blocks should also enable the assembly, interconnection, and integration of a broad range of functional nanosystems.

In a later report, Dai *et al*. described the creation of densely aligned single-walled carbon nanotubes (SWCNTs), having used the LB technique for device integration [110]. This approach, which enabled the controlled assembly and integration of SWCNTs on a scale far beyond that achieved with an individual SWCNT, will undoubtedly prove important in the future electronics industry. In this procedure, the aligned SWNT monolayers were first transferred onto oxide substrates by the LB technique, followed by the fabrication of arrays of two-terminal devices with a Ti/Au metal source (S) and drain (D) that contacted massively parallel SWNTs in ~10 μm wide S–D regions, with a channel length of ~250 nm (Figure 8.22c and d). Subsequent current versus bias voltage (*I*–*V*) measurements showed that devices created from Hipco SWNTs were over 25-fold more resistive than similar devices prepared from laser-ablation SWNTs, with currents reaching ~0.13 and ~3.5 mA, respectively, at a bias of 3 V through the collective current carrying of SWNTs in parallel (Figure 8.22e and f). The Hipco SWNT devices also exhibited a greater nonlinearity in their *I*–*V* characteristics than did the laser-ablation nanotubes (Figure 8.22e). These characteristics were attributed to diameter differences between the Hipco and laser-ablation materials. The same process was also used for the integration of silver nanowires, where both nanowire density and line spacing can be programmed [111].

8.6 Conclusions and Outlook

Today, the fundamental challenge in the development of nanotechnology is the assembly of building blocks into organized nanostructures, albeit in controllable fashion. The use of LB patterning, as one branch of self-assembly and as discussed in this chapter, provides an effective and distinguished solution to the fabrication of high-ordered structures over larger areas. Compared to other assembly technique, such as self-assembly monolayers, layer-by-layer methods and vacuum molecular deposition, LB patterning incorporates certain unique properties. First, it allows the study of how the structural organization of different molecules at interfaces depends on changing various thermodynamic parameters such as temperature, film pressure, the chemical potential through employing different environments, molecular composition (whether single-component or multi-component), subphase composition, or by the controlled manipulation of certain transfer parameters, such as the speeds of transference and compression, and the method of transfer (whether vertical or horizontal). In addition, controlling the shape, size, and distribution of molecular assemblies in self-organized surface patterns will in turn lead to a controlled construction of 2-D supramolecular architectures. Thus, LB patterning can offer

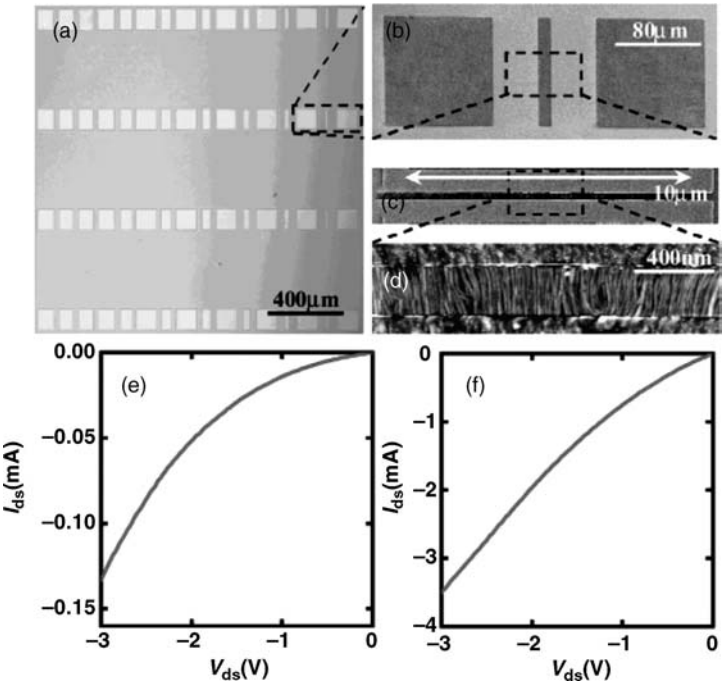


Figure 8.22 Microfabrication patterning and device integration of SWNT LB films. (a) Optical image of a patterned SWNT LB film. The squares and rectangles are regions containing densely aligned SWNTs. Other areas are SiO_2 substrate regions; (b) SEM image of a region highlighted in panel (a) with packed SWNTs aligned vertically; (c) SEM image showing a 10 μm-wide SWNT LB film between source and drain electrodes formed in a region marked in panel (b); (d) AFM image of a region in panel (c), showing aligned SWNTs and the edges of the S and D electrodes; (e) Current versus bias (I_{ds}–V_{ds}) curve of a device made from Hipco SWNTs (10 μm channel width and 250 nm channel length); (f) I_{ds}–V_{ds} of a device made from laser ablation SWNTs (10 μm channel width and 250 nm channel length). Reproduced with permission from Ref. [110].

a great opportunity to achieve control over nanoscale assembly by tuning a macroscopic property, whilst linking the LB technique with patterning and fabrication on a solid substrate.

In addition to expanding the types of building block used in LB patterning, including molecules, nanorods, nanotubes, and nanowires, there remain a number of unanswered questions. The first of these questions involves the orthogonal assembly of multicomponent building blocks with different functions into an integrated system, while the second question relates to the combination of LB patterning with other lithographic techniques, such as soft lithography and DPN, so as to enhance the capabilities of LB patterning. Finally, the integration of LB patterning into functional device architectures remains an important issue for practical applications.

References

1 Aizenberg, J., Black, A.J., and Whitesides, G.M. (1999) *Nature*, **398**, 495–498.
2 Kim, S.O., Solak, H.H., Stoykovich, M.P., Ferrier, N.J., de Pablo, J.J., and Nealey, P.F. (2003) *Nature*, **424**, 411–414.
3 Cavallini, M., Biscarini, F., Gomez-Segura, J., Ruiz, D., and Veciana, J. (2003) *Nano Lett.*, **3**, 1527–1530.
4 Jacobs, H.O., Tao, A.R., Schwartz, A., Gracias, D.H., and Whitesides, G.M. (2002) *Science*, **296**, 323–325.
5 Koide, Y., Wang, Q.W., Cui, J., Benson, D.D., and Marks, T.J. (2000) *J. Am. Chem. Soc.*, **122**, 11266–11267.
6 Kataoka, D.E. and Troian, S.M. (1999) *Nature*, **402**, 794–797.
7 Velev, O.D. and Kaler, E.W. (1999) *Langmuir*, **15**, 3693–3698.
8 Michel, R., Lussi, J.W., Csucs, G., Reviakine, I., Danuser, G., Ketterer, B., Hubbell, J.A., Textor, M., and Spencer, N.D. (2002) *Langmuir*, **18**, 3281–3287.
9 Harrison, D.J., Fluri, K., Seiler, K., Fan, Z.H., Effenhauser, C.S., and Manz, A. (1993) *Science*, **261**, 895–897.
10 Delamarche, E., Bernard, A., Schmid, H., Michel, B., and Biebuyck, H. (1997) *Science*, **276**, 779–781.
11 Volkmuth, W.D. and Austin, R.H. (1992) *Nature*, **358**, 600–602.
12 Seo, Y.S., Luo, H., Samuilov, V.A., Rafailovich, M.H., Sokolov, J., Gersappe, D., and Chu, B. (2004) *Nano Lett.*, **4**, 659–664.
13 Fan, Y.W., Cui, F.Z., Hou, S.P., Xu, Q.Y., Chen, L.N., and Lee, I.S. (2002) *J. Neurosci. Methods*, **120**, 17–23.
14 Xu, C.Y., Inai, R., Kotaki, M., and Ramakrishna, S. (2004) *Biomaterials*, **25**, 877–886.
15 Abbott, N.L., Folkers, J.P., and Whitesides, G.M. (1992) *Science*, **257**, 1380–1382.
16 Kramer, S., Fuierer, R.R., and Gorman, C.B. (2003) *Chem. Rev.*, **103**, 4367–4418.
17 Nyffenegger, R.M. and Penner, R.M. (1997) *Chem. Rev.*, **97**, 1195–1230.
18 Chou, S.Y., Krauss, P.R., and Renstrom, P.J. (1996) *Science*, **272**, 85–87.
19 Xia, Y.N. and Whitesides, G.M. (1998) *Annu. Rev. Mater. Sci.*, **28**, 153–184.
20 Park, M., Harrison, C., Chaikin, P.M., Register, R.A., and Adamson, D.H. (1997) *Science*, **276**, 1401–1404.
21 Schaffer, E., Thurn-Albrecht, T., Russell, T.P., and Steiner, U. (2000) *Nature*, **403**, 874–877.
22 Spatz, J.P., Roescher, A., and Moller, M. (1996) *Adv. Mater.*, **8**, 337–340.
23 Chen, X.D., Lenhert, S., Hirtz, M., Lu, N., Fuchs, H., and Chi, L.F. (2007) *Acc. Chem. Res.*, **40**, 393–401.
24 Franklin, B. (1774) *Philos. Trans. R. Soc.*, **64**, 445.
25 Rayleigh, L. (1890) *Proc. R. Soc.*, **47**, 364.
26 Pockels, A. (1891) *Nature*, **43**, 437.
27 Langmuir, I. (1917) *J. Am. Chem. Soc.*, **39**, 1848–1906.
28 Blodgett, K.B. (1935) *J. Am. Chem. Soc.*, **57**, 1007–1022.
29 Gains, L.G. (1969) *Insoluble Monolayers at Liquid-Gas Interfaces*, Interscience Publishers, New York.
30 Roberts, G. (1990) *Langmuir-Blodgett Films*, Plenum Press, New York.
31 Kjaer, K., Alsnielsen, J., Helm, C.A., Laxhuber, L.A., and Mohwald, H. (1987) *Phys. Rev. Lett.*, **58**, 2224–2227.
32 Losche, M. and Mohwald, H. (1984) *Rev. Sci. Instrum.*, **55**, 1968–1972.
33 Henon, S. and Meunier, J. (1991) *Rev. Sci. Instrum.*, **62**, 936–939.
34 Honig, D. and Mobius, D. (1991) *J. Phys. Chem.*, **95**, 4590–4592.
35 McConnell, H.M. (1991) *Annu. Rev. Phys. Chem.*, **42**, 171–195.
36 Mohwald, H. (1990) *Annu. Rev. Phys. Chem.*, **41**, 441–476.
37 Chi, L.F., Anders, M., Fuchs, H., Johnston, R., and Ringsdorf, H. (1993) *Science*, **259**, 213–216.
38 Chi, L.F., Eng, L.M., Graf, K., and Fuchs, H. (1992) *Langmuir*, **8**, 2255–2261.
39 Kato, T., Kameyama, M., Ehara, M., and Iimura, K. (1998) *Langmuir*, **14**, 1786–1798.
40 Trabelsi, S., Zhang, S.S., Zhang, Z.C., Lee, T.R., and Schwartz, D.K. (2009) *Soft Matter*, **5**, 750–758.

41 Fontaine, P., Goldmann, M., Muller, P., Faure, M.C., Konovalov, O., and Krafft, M.P. (2005) *J. Am. Chem. Soc.*, **127**, 512–513.

42 Zhang, G., Marie, P., Maaloum, M., Muller, P., Benoit, N., and Krafft, M.P. (2005) *J. Am. Chem. Soc.*, **127**, 10412–10419.

43 Maaloum, M., Muller, P., and Krafft, M.P. (2002) *Angew. Chem., Int. Ed.*, **41**, 4331–4334.

44 Huang, X., Li, C., Jiang, S.G., Wang, X.S., Zhang, B.W., and Liu, M.H. (2004) *J. Am. Chem. Soc.*, **126**, 1322–1323.

45 Zhang, Y., Chen, P., Jiang, L., Hu, W., and Liu, M. (2009) *J. Am. Chem. Soc.*, **131**, 2756–2757.

46 Zhang, L., Gaponik, N., Muller, J., Plate, U., Weller, H., Erker, G., Fuchs, H., Rogach, A.L., and Chi, L.F. (2005) *Small*, **1**, 524–527.

47 Lenhert, S., Zhang, L., Mueller, J., Wiesmann, H.P., Erker, G., Fuchs, H., and Chi, L.F. (2004) *Adv. Mater.*, **16**, 619–624.

48 Chunbo, Y., Xinmin, L., Desheng, D., Bin, L., Hongjie, Z., Zuhong, L., Juzheng, L., and Jiazuan, N. (1996) *Surf. Sci.*, **366**, L729–L734.

49 Chi, L.F., Jacobi, S., Anczykowski, B., Overs, M., Schafer, H.J., and Fuchs, H. (2000) *Adv. Mater.*, **12**, 25–30.

50 Powers, E.T., Yang, S.I., Lieber, C.M., and Kelly, J.W. (2002) *Angew. Chem., Int. Ed.*, **41**, 127–130.

51 Iimura, K., Shiraku, T., and Kato, T. (2002) *Langmuir*, **18**, 10183–10190.

52 Imae, T., Takeshita, T., and Kato, M. (2000) *Langmuir*, **16**, 612–621.

53 Zhu, J.Y., Eisenberg, A., and Lennox, R.B. (1991) *J. Am. Chem. Soc.*, **113**, 5583–5588.

54 Zhu, J., Eisenberg, A., and Lennox, R.B. (1992) *Macromolecules*, **25**, 6556–6562.

55 Li, S., Hanley, S., Khan, I., Varshney, S.K., Eisenberg, A., and Lennox, R.B. (1993) *Langmuir*, **9**, 2243–2246.

56 Baker, S.M., Leach, K.A., Devereaux, C.E., and Gragson, D.E. (2000) *Macromolecules*, **33**, 5432–5436.

57 Devereaux, C.A. and Baker, S.M. (2002) *Macromolecules*, **35**, 1921–1927.

58 Gamboa, A.L.S., Filipe, E.J.M., and Brogueira, P. (2002) *Nano Lett.*, **2**, 1083–1086.

59 Seo, Y.S., Kim, K.S., Galambos, A., Lammertink, R.G.H., Vancso, G.J., Sokolov, J., and Rafailovich, M. (2004) *Nano Lett.*, **4**, 483–486.

60 Gleiche, M., Chi, L.F., and Fuchs, H. (2000) *Nature*, **403**, 173–175.

61 Spratte, K., Chi, L.F., and Riegler, H. (1994) *Europhys. Lett.*, **25**, 211–217.

62 Chen, X.D., Lu, N., Zhang, H., Hirtz, M., Wu, L.X., Fuchs, H., and Chi, L.F. (2006) *J. Phys. Chem. B*, **110**, 8039–8046.

63 Lenhert, S., Gleiche, M., Fuchs, H., and Chi, L.F. (2005) *ChemPhysChem*, **6**, 2495–2498.

64 Chen, X.D., Hirtz, M., Fuchs, H., and Chi, L.F. (2007) *Langmuir*, **23**, 2280–2283.

65 Hirtz, M., Fuchs, H., and Chi, L.F. (2008) *J. Phys. Chem. B*, **112**, 824–827.

66 Lu, N., Gleiche, M., Zheng, J.W., Lenhert, S., Xu, B., Chi, L.F., and Fuchs, H. (2002) *Adv. Mater.*, **14**, 1812–1815.

67 Chen, X.D., Hirtz, M., Fuchs, H., and Chi, L.F. (2005) *Adv. Mater.*, **17**, 2881–2885.

68 Markovich, G., Collier, C.P., Henrichs, S.E., Remacle, F., Levine, R.D., and Heath, J.R. (1999) *Acc. Chem. Res.*, **32**, 415–423.

69 Brust, M., Stuhr-Hansen, N., Norgaard, K., Christensen, J.B., Nielsen, L.K., and Bjornholm, T. (2001) *Nano Lett.*, **1**, 189–191.

70 Song, H., Kim, F., Connor, S., Somorjai, G.A., and Yang, P.D. (2005) *J. Phys. Chem. B*, **109**, 188–193.

71 Huang, S.J., Tsutsui, G., Sakaue, H., Shingubara, S., and Takahagi, T. (2001) *J. Vac. Sci. Technol. B*, **19**, 115–120.

72 Chen, S.W. (2001) *Langmuir*, **17**, 2878–2884.

73 Tian, Y.C. and Fendler, J.H. (1996) *Chem. Mater.*, **8**, 969–974.

74 Paul, S., Pearson, C., Molloy, A., Cousins, M.A., Green, M., Kolliopoulou, S., Dimitrakis, P., Normand, P., Tsoukalas, D., and Petty, M.C. (2003) *Nano Lett.*, **3**, 533–536.

75 Tao, A.R., Huang, J.X., and Yang, P.D. (2008) *Acc. Chem. Res.*, **41**, 1662–1673.

76 Chung, S.W., Markovich, G., and Heath, J.R. (1998) *J. Phys. Chem. B*, **102**, 6685–6687.

77 Sear, R.P., Chung, S.W., Markovich, G., Gelbart, W.M., and Heath, J.R. (1999) *Phys. Rev. E*, **59**, R6255–R6258.

78 Huang, J.X., Kim, F., Tao, A.R., Connor, S., and Yang, P.D. (2005) *Nat. Mater.*, **4**, 896–900.

79 Huang, J.X., Tao, A.R., Connor, S., He, R.R., Yang, P.D. (2006) *Nano Lett.*, **6**, 524–529.

80 Vidoni, O., Reuter, T., Torma, V., Meyer-Zaika, W., and Schmid, G. (2001) *J. Mater. Chem.*, **11**, 3188–3190.

81 Hassenkam, T., Norgaard, K., Iversen, L., Kiely, C.J., Brust, M., and Bjornholm, T. (2002) *Adv. Mater.*, **14**, 1126–1130.

82 Reuter, T., Vidoni, O., Torma, V., Schmid, G., Lu, N., Gleiche, M., Chi, L.F., and Fuchs, H. (2002) *Nano Lett.*, **2**, 709–711.

83 Chi, L.F., Rakers, S., Hartig, M., Gleiche, M., Fuchs, H., and Schmid, G. (2000) *Colloids Surf. A.*, **17** (1), 241–248.

84 Lu, N., Zheng, J.W., Gleiche, M., Fuchs, H., Chi, L.F., Vidoni, O., Reuter, T., and Schmid, G. (2002) *Nano Lett.*, **2**, 1097–1099.

85 Haynes, C.L. and Van Duyne, R.P. (2001) *J. Phys. Chem. B*, **105**, 5599–5611.

86 Fahmi, A.W., Oertel, U., Steinert, V., Froeck, C., and Stamm, M. (2003) *Macromol. Rapid Comm.*, **24**, 625–629.

87 Li, L.S., Jin, J., Yu, S., Zhao, Y.Y., Zhang, C.X., and Li, T.J. (1998) *J. Phys. Chem. B*, **102**, 5648–5652.

88 Cheyne, R.B. and Moffitt, M.G. (2005) *Langmuir*, **21**, 10297–10300.

89 Kim, F., Kwan, S., Akana, J., and Yang, P.D. (2001) *J. Am. Chem. Soc*, **123**, 4360–4361.

90 Yang, P.D. and Kim, F. (2002) *ChemPhysChem*, **3**, 503–506.

91 Kwan, S., Kim, F., Akana, J., and Yang, P.D. (2001) *Chem. Commun.*, 447–448.

92 Tao, A., Kim, F., Hess, C., Goldberger, J., He, R.R., Sun, Y.G., Xia, Y.N., and Yang, P.D. (2003) *Nano Lett.*, **3**, 1229–1233.

93 Tao, A.R. and Yang, P.D. (2005) *J. Phys. Chem. B*, **109**, 15687–15690.

94 Whang, D., Jin, S., Wu, Y., and Lieber, C.M. (2003) *Nano Lett.*, **3**, 1255–1259.

95 Gleiche, M., Chi, L.F., Gedig, E., and Fuchs, H. (2001) *ChemPhysChem*, **2**, 187–191.

96 Berger, C.E.H., Vanderwerf, K.O., Kooyman, R.P.H., Degrooth, B.G., and Greve, J. (1995) *Langmuir*, **11**, 4188–4192.

97 Moraille, P. and Badia, A. (2002) *Angew. Chem., Int. Ed.*, **41**, 4303–4306.

98 Zhang, M.Z., Lenhert, S., Wang, M., Chi, L.F., Lu, N., Fuchs, H., and Ming, N.B. (2004) *Adv. Mater.*, **16**, 409–413.

99 Lu, N., Chen, X.D., Molenda, D., Naber, A., Fuchs, H., Talapin, D.V., Weller, H., Muller, J., Lupton, J.M., Feldmann, J., Rogach, A.L., and Chi, L.F. (2004) *Nano Lett.*, **4**, 885–888.

100 Naber, A., Molenda, D., Fischer, U.C., Maas, H.J., Hoppener, C., Lu, N., and Fuchs, H. (2002) *Phys. Rev. Lett.*, **89**, 210801.

101 Chen, X.D., Rogach, A.L., Talapin, D.V., Fuchs, H., and Chi, L.F. (2006) *J. Am. Chem. Soc.*, **128**, 9592–9593.

102 Lenhert, S., Meier, M.B., Meyer, U., Chi, L.F., and Wiesmann, H.P. (2005) *Biomaterials*, **26**, 563–570.

103 Lenhert, S., Semsa, A., Hirtz, M., Chi, L.F., Fuchs, H., Wiesmann, H.P., Osbourn, A.E., and Moerschbacher, B.M. (2007) *Langmuir*, **23**, 10216.

104 Wu, X.C., Lenhert, S., Chi, L.F., and Fuchs, H. (2006) *Langmuir*, **22**, 7807–7811.

105 Whang, D., Jin, S., and Lieber, C.M. (2003) *Nano Lett.*, **3**, 951–954.

106 Choi, S.H., Wang, K.L., Leung, M.S., Stupian, G.W., Presser, N., Chung, S.W., Markovich, G., Kim, S.H., and Heath, J.R. (1999) *J. Vac. Sci. Technol. A*, **17**, 1425–1427.

107 Meli, M.V., Badia, A., Grutter, P., and Lennox, R.B. (2002) *Nano Lett.*, **2**, 131–135.

108 Meli, M.V. and Lennox, R.B. (2003) *Langmuir*, **19**, 9097–9100.

109 Jin, S., Whang, D.M., McAlpine, M.C., Friedman, R.S., Wu, Y., and Lieber, C.M. (2004) *Nano Lett.*, **4**, 915–919.

110 Li, X.L., Zhang, L., Wang, X.R., Shimoyama, I., Sun, X.M., Seo, W.S., and Dai, H.J. (2007) *J. Am. Chem. Soc.*, **129**, 4890–4891.

111 Huang, J.X., Fan, R., Connor, S., and Yang, P.D. (2007) *Angew. Chem., Int. Ed.*, **46**, 2414–2417.

9
Surface-Supported Nanostructures Directed by Atomic- and Molecular-Level Templates

Dingyong Zhong, Haiming Zhang, and Lifeng Chi

9.1
Introduction

Shrinking the size of devices to nanoscale, and discovering new phenomena of objects at the nanoscale, represent the two main goals of nanotechnology in both academia and industry. Unfortunately, the traditional photolithographic techniques that have been used widely to create patterns of characteristic size in the micrometer scale, encounter bottlenecks when attempts are made to further reduce the size to the nanoscale, due to the diffractive effect of light. Many so-called "top-down" methods of fabrication, which include electron-beam lithography (EBL), imprinting, contact printing, and scanning probe lithography, have been developed during the past two decades. Whilst such techniques have the advantage of producing patterns with sizes that, characteristically, are smaller than 100 nm, their low throughput and/or poor reproducibility limits their application.

In contrast to the top-down techniques, "bottom-up" techniques utilize the self-assembly of nanosized objects such as atoms, molecules and nanoparticles, from which ordered structures in the nanoscale are built. Moreover, by adjusting the interactions between the nanosized objects and certain kinetic parameters, such as the growth temperature and rate, the process of self-assembly can easily be controlled to produce final phases with desired architectures. The characteristic sizes of these self-assembled architectures, which range from a few nanometers to tens of nanometers, are below the limitations of traditional patterning techniques. Nonetheless, although routine industrial applications might still be a long way off, patterning techniques based on self-assembly continue to show much promise, based on their ability to control the utilization of small entities such atoms, molecules, or nanoparticles.

Self-assembly processes directed by surface-supported templates are widely used to build architectures of functional materials. For this technique, the substrate surfaces are pre-patterned in order to modulate – either thermodynamically or kinetically – the aggregation of certain nanosized objects at specific sites and/or orientations. One type of natural template with an atomic order is the crystalline surface (Figure 9.1), where the atomic periodicities of an original surface are several angstroms, whilst reconstructions and reconstruction-induced patterns have a

Nanotechnology, Volume 8: Nanostructured Surfaces. Edited by Lifeng Chi

ISBN: 978-3-527-31739-4

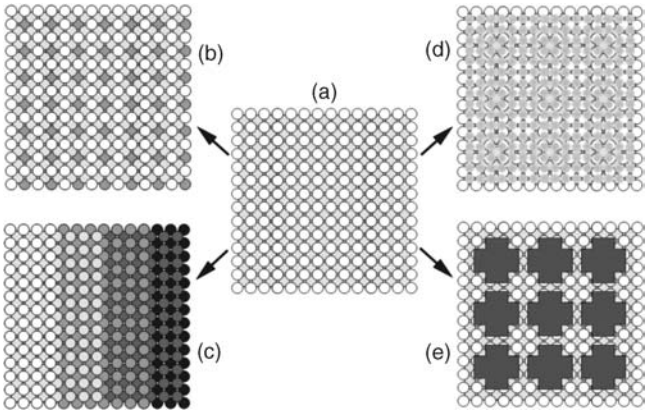


Figure 9.1 Atomic and molecular level templates on surfaces. (a) Original surface with 2-D lattice; (b) Reconstructed surface; (c) Vicinal surface; (d) Strain-relief epitaxial layer; (e) Supramolecular network.

periodicity of several nanometers. Furthermore, the vicinal surfaces with a periodic succession of terraces and steps may have a periodicity ranging from several nanometers to tens of nanometers. These natural templates, as well as self-assembled supramolecular structures and strain-relief epitaxial layers on surfaces, are used to direct the formation of further surface-supported nanostructures.

Scanning tunneling microscopy (STM) is a powerful tool that can be used to visualize nanostructured surfaces. In this process, a sharp conductive tip is positioned near the surface of a conductive sample. With a typical gap of 0.5–1 nm, the electron wave functions of the tip and the surface overlap, such that a tunneling current is created when a voltage bias is applied between the tip and the sample. The tunneling current is relevant to the local density of states on the surface, and is also very sensitive to the gap width. By keeping the current constant and varying the height of the tip, with the help of a piezoelectric system and feedback electronics, the tip is allowed to scan the sample surface such that a topographic image of the surface is obtained with $\sim$0.1 nm lateral resolution, and $\sim$0.01 nm height resolution.

For simplicity, this chapter contain two parts detailing two-dimensional (2-D) atomic- and molecular-level templates, and nanostructure formation as directed by the templates. Initially, the different types of template, including reconstructed surfaces, vicinal surfaces, strain-relief epitaxial layers and supramolecular assemblies, will be introduced, after which the formation of surface-supported nanostructures guided by the templates will be discussed. Throughout the chapter, emphasis will be placed on organic and organic–inorganic hybrid nanostructures.

9.2
Atomic- and Molecular-Level Templates on Surfaces

The root of atomic- and molecular-level templates is the periodic arrangement of atoms and molecules in crystalline structures. Such periodic arrangement is reflected

by the 2-D lattices of single crystalline surfaces. In order to template the formation of nanostructures, a "clean" surface is required under a well-controlled environment, for example under ultra-high-vacuum (UHV) conditions, or in specific solutions. Methods to obtain ordered and clean surfaces include the cleavage of certain crystals (notably layered materials such as graphite, mica and MoS_2), sputtering–annealing cycles for metals, and high-temperature flashing for semiconductors, such as silicon.

9.2.1
Surface Reconstructions and Reconstruction-Related Patterns

Similar to the structure in the bulk, the atomic periodicities of an original surface are several angstroms. However, due to the absence of neighboring atoms on one side of a surface, the first few layers of atoms beneath the surface will be reorganized so as to achieve an equilibrium state with lower surface free energy, and this will result in surface reconstruction. Several well-known surface reconstructions, including Si(111)-7×7, Au(111)-$22 \times \sqrt{3}$ and Cu(110)-(2×1)O, are briefly introduced in the following subsections.

9.2.1.1 Si(111)-7×7

The Si(111)-7×7 was the first surface reconstruction to be visualized by STM [1]. To prepare the Si(111)-7×7 reconstruction, the Si(111) surface is etched with HF solution to remove the native oxide layer on the surface. The Si(111) surface is then heated repeatedly to about ~1200 K under UHV conditions, which causes the atoms of the outermost atomic layers to reorganize and to form a structure with a hexagonal unit cell, the lattice constant of which is 2.7 nm. The surface atoms are rearranged in such a way that the number of dangling bonds is minimized. The unit cell of the reconstruction contains 12 adatoms, nine dimers, and a stacking fault layer based on the DAS model proposed by K. Takayanagi *et al.* [2, 3].

9.2.1.2 Au(111)-$22 \times \sqrt{3}$ and the Herringbone Pattern

In the Au(111)-$22 \times \sqrt{3}$ reconstruction, the first hexagonally arranged atomic layer of Au(111) is compressed with 23 atoms stacked on 22 bulk lattice sites along the [01-1] direction. As a result, alternative stripes of domains with face-centered cubic (*fcc*) and hexagonal-close-packed (*hcp*) stacking styles running along the perpendicular [−211] direction are formed (Figure 9.2). The two types of domain are gradually transited with a height difference of 0.02 nm. The reconstruction has a rectangle unit cell with a periodicity in [01-1] of 6.34 nm and in [−211] of 0.416 nm. Corresponding to the threefold symmetry of the Au(111) surface, there are three equivalent orientations of the stripes by 120° rotation. The orientational domains coexist, and their boundaries merge in such a way that transition from one domain to another occurs through a correlated periodic bending of the parallel stripes, and this results in a zigzag or "herringbone" pattern. The herringbone structure changes the orientation over distances of up to 25 nm as a result of long-range elastic lattice strain [4, 5].

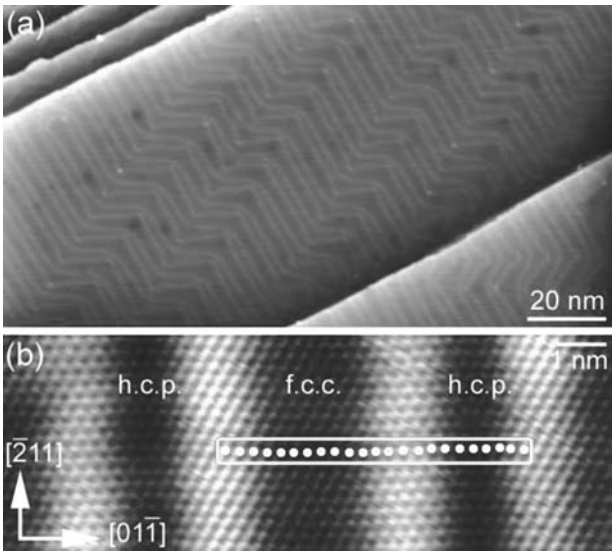


Figure 9.2 Au(111) $22 \times \sqrt{3}$ reconstruction. (a) Large-scale STM image showing the herringbone pattern; (b) High-resolution STM image showing the *hcp* and *fcc* regions. The unit cell of the reconstruction is denoted by the rectangle.

9.2.1.3 Cu(110)-(2 × 1)O and the Stripe Pattern

Although no reconstruction exists at a clean Cu(110) surface, the controlled condensation of oxygen atoms on the surfaces results in a 2×1 added-row reconstruction. The O atoms are chemisorbed on the surface, while the Cu adatoms evaporate from steps and diffuse across the terraces of the substrate surface [6, 7]. In this construction, Cu atom rows are added along the [001] direction with a distance of 0.51 nm, and O atoms are adsorbed at the long-bridge sites in the rows (Figure 9.3a). Depending on the O atom coverage, full or partial coverage of the adsorbate-induced reconstruction can be obtained. By exposing a clean Cu(110) surface to 4–6 Langmuir ($1\,\text{Langmuir} = 10^{-6}\,\text{Torr}\cdot\text{s}$) of oxygen at 625 K, the stripe pattern with alternating stripes of bare Cu and (2×1)O reconstructed regions along the [001] direction are obtained, as shown in Figure 9.3b and c [8, 9]. The dimensions of the stripe structure may be adjusted by controlling the oxidation process; that is, the temperature and exposure dose. The width and distance of the stripes are increased when the process temperature is increased, whereas an increase in exposure will result in narrower stripes. For instance, stripe patterns with long-range order on the length scale of several hundred nanometers consisting of bare Cu troughs 2.0 ± 0.3 nm wide, separated by Cu−O regions 5 ± 2 nm wide have been obtained (Figure 9.3b and c).

9.2.2 Strain-Relief Epitaxial Layers

In heteroepitaxial systems exhibiting large lattice mismatch, the substrate lattice is not strictly followed by the epitaxial layer, and this results in the generation of dense

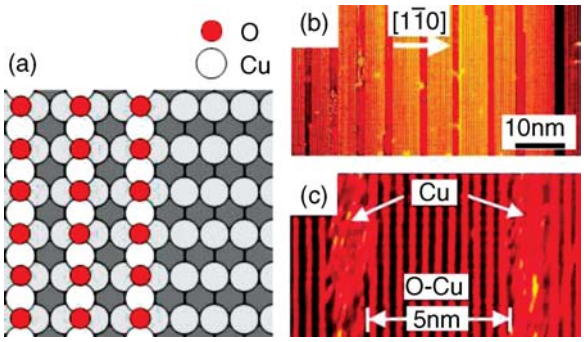


Figure 9.3 Cu(110)-(2 × 1)O: Nanostripes of Cu(110) and Cu(110)-(2 × 1)O reconstruction regions. (a) Schematic of the oxygen-induced added-row reconstruction; (b, c), STM images of the stripe patterns containing alternative bare Cu(110) regions and oxygen-induced 2 × 1 reconstruction regions, respectively [8].

dislocations to relieve large strain in the systems. Due to interactions between the epitaxial layer and the substrate lattice, and the high mobility of dislocations located at the surfaces, the dislocations quite often arrange into highly ordered periodic patterns. In other words, the epitaxial layer prefers to form a large superstructure that is commensurate with the substrate. Such superstructures, showing periodicities of several nanometers, will further modulate surface processes such as adsorption, diffusion, and the nucleation of adatoms or molecules deposited onto the strain-relief epitaxial layers.

9.2.2.1 **Boron Nitride Nanomesh**

Analogous to carbon solids, covalent-bonded boron nitride exists in a variety of structures, including graphitic hexagonal boron nitride (h-BN) and diamond-like cubic boron nitride (c-BN). Ultrathin h-BN films are formed on certain metallic surfaces by chemical vapor deposition (CVD). For example, large terraces of h-BN films with monolayer thickness are formed on Ni(111), which shows a lattice mismatch of 0.4% with the h-BN layer [10]. The honeycomb-structured h-BN films show little corrugation with the N atoms located on top of the outermost Ni atoms and the B atoms on *fcc* adsorption sites of the Ni(111) surface. In case of substrates with large lattice misfit, an obvious corrugation appears in the epitaxial h-BN film, resulting in ordered superstructures with a periodicity of several nanometers [11, 12]. By exposing a Rh(111) surface kept at a temperature of 1070 K to a borazine $(HBNH)_3$ vapor pressure of 3×10^{-7} mbar for about 2 min (40 L), an h-BN film with a regular nanomesh would be formed (as shown in Figure 9.4) [11]. The hexagonal nanomesh has a periodicity of 3.22 nm, with pores of 2 nm diameter; this corresponds to a commensurate lattice of (13 × 13) h-BN units on a (12 × 12) Rh lattice spacing. The mismatch between the h-BN layer (0.248 nm) and the Rh(111) surface (0.269 nm) plays a key role in the nanomesh formation. Based on experimental results and density functional theory (DFT) calculations, it has been proposed that the nanomesh would consist of one layer of h-BN that would be

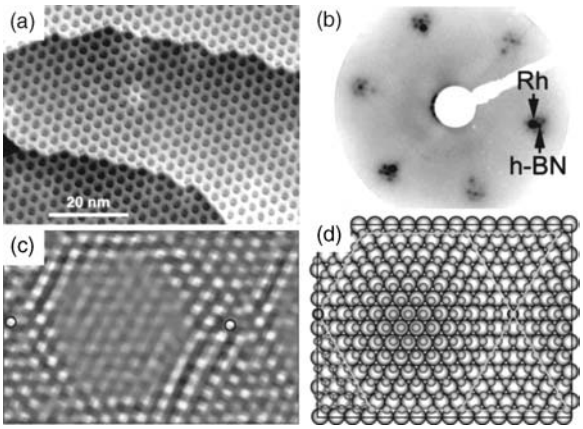


Figure 9.4 Hexagonal boron nitride (h-BN) nanomesh. (a) Large-area constant-current STM image (−1.0 V, 2.5 nA, at room temperature) of the boron nitride nanomesh formed by high-temperature decomposition of borazine on a Rh(111) surface; (b) Low-energy electron diffraction (LEED) pattern from h-BN nanomesh (40-L exposure) on Rh(111). The principal diffraction spots of the Rh(111) surface and h-BN (arrowed) are accompanied with satellite nanomesh superlattice spots. Electron energy: 92 eV; (c) Low-frequency filtered STM image (−2 mV, 1 nA, at 77 K) showing the atomic corrugation. Bright protrusions indicate N atoms; (d) Atomic structures of the h-BN nanomesh on Rh(111) (Green: N; orange: B; gray: Rh) [11, 14].

strongly corrugated due to mismatch with the substrate [13, 14]. The N atoms of the h-BN layer in holes occupy the on-top sites, which are the energetically favorable sites for h-BN grown on a metal surface, whereas the locations of the N atoms from the surrounding of holes will deviate from the on-top sites, owing to the large lattice misfit (6.7%) (see Figure 9.4d). Besides Rh(111), similar nanomesh has been observed on Ru(0001) [12].

9.2.2.2 **Ag/Pt(111)**

The double atomic-layer of Ag on Pt(111) is another example of superstructures induced by mismatched epitaxy. Since the lattice constant of Ag is 4.3% larger than that of Pt, the first commensurate monolayer of Ag on Pt(111) is coherently strained. The second monolayer of Ag, on the other hand, forms an ordered trigonal network of dislocations to partially relieve the compressive strain. Domains with *fcc* and *hcp* stacking coexist in the second monolayer. The unit cell of the superstructure with trigonal symmetry has an average periodicity of 4.5 nm consisting of a large quasi-hexagon (*fcc* stacking) and two triangles with opposite orientation (*hcp* stacking) [15, 16]. The strained stacking styles in the epitaxial Ag film show effect on certain processes such as surface diffusion, nucleation, and reaction activity [15, 17]. The stability of the strain-relief pattern is enhanced by removing about 0.1 monoloayer (ML) of the Ag top layer of this surface structure by He- or Ar-ion sputtering. As a result, a hexagonally well-ordered, room-temperature-stable array of 1 ML-deep holes with a tunable size of about 4 nm^2 and a fixed spacing of 7 nm was fabricat-

ed [18]. In this process, the *hcp*-stacked Ag atoms in the triangle-shaped domains were removed. A similar hole array has been obtained by exposing a strained monolayer-thick Ag layer on Ru(0001), in which a superstructure with a periodicity of about 4–6 nm was formed due to the 7% lattice mismatch, to sulfur at room temperature. In this way, the Ag atoms at dislocated regions could be selectively replaced by S atoms [19, 20].

9.2.3 Vicinal Surfaces

Vicinal surfaces (also known as "high-index surfaces") on which a periodic succession of terraces and steps of monoatomic height exist, are obtained by cutting crystals along a plane that deviates by a small angle from a low-index plane. Such structural anisotropy may be represented in macroscopic physical properties [21]. In comparison to a low-index surface, the high step density increases the surface energy and makes the vicinal surfaces metastable, and tends towards to faceting with the formation of macroscopic low-index surfaces in low-temperature limits. Nevertheless, the stepped vicinal surfaces may be available at low temperatures if the kinetics of reordering is too slow (owing to the reduced temperatures), so as to allow a transition from a stepped-back to a faceted morphology. At high temperature, the entropy of the steps provides a substantial contribution to the balance of free energy, and the formation of an ordered vicinal surface may minimize the total surface free energy. The entropy, which provides a mechanism for step–step repulsion, also plays an important role in the regularity of the step arrangement on a vicinal surface [22]. However, a roughing transition may take place if the temperature exceeds a critical transition point, which is variable for different systems. For instance, the transition temperature is 290 K for vicinal Cu $(11n)$ $(n = 13, 19, 79)$ surfaces [23] and 465 K for Ag(115) surfaces [24]. To an annealing temperature above 900 K, faceting takes place on a Pt(997) vicinal surface [25].

Besides the step–step interactions, the surface morphology of vicinal surfaces is affected by thermal kink creation energies [26]. For example, highly ordered step arrays are obtained on a vicinal Si(111) surface about 1° miscut toward $[-1\ -1\ 2]$. As the kink width is half a 7×7 unit cell (2.3 nm), the energy barrier for creating a kink is very high and the step edges are atomically straight up to 2×10^4 lattice sites [27]. Furthermore, adsorbate-induced faceting takes place on certain vicinal surfaces. The adsorption of O on a vicinal Ag(110) misoriented 2° toward [001] or $[3\ -3\ 1]$ induces faceting of the surface, with the formation of large (110) facets and step-bunching [28]. Step-doubling takes place on a Pt(997) surface by annealing the surface to 700 K at an oxygen atmosphere for a few minutes [25].

9.2.3.1 Vicinal Au(111) Surfaces

The vicinal Au(111) surfaces are briefly introduced as an example in this section (for detailed discussions, see Ref. [30]). Two types of steps running along $[0\ -1\ 1]$, with {111} and {100} microfacets, are obtained by miscutting toward $[-2\ 1\ 1]$ and $[2\ -1\ -1]$,

respectively. Among these vicinal surfaces misoriented with an angle up to 12°, the (322), (755), and (233) surfaces are stable with unreconstructed terraces, while the (788) and (11 12 12) surfaces are stable with reconstructed terraces. However, some vicinal surfaces, including (455), (577), and (12 11 11), are unstable and undergo faceting with "hill-and-valley" morphology. As a result, facets with two or more stable orientations coexist on the surface may be self-organized into a larger periodicity from 10 nm to a few hundreds of nanometers. The Au(788) and Au(11, 12, 12) surfaces, which are misoriented by 3.5° and 2.3° with respect to the (111) towards the [−211] direction, are regularly stepped surfaces with monoatomic {111} microfacets and reconstructed terraces, which are 3.9 and 5.8 nm in width, respectively (Figure 9.5). In this way, a very narrow terrace width distribution with a 0.85 nm full width at half-maximum (FWHM) has been obtained [30]. The discommensuration lines on the terraces run perpendicular to the step edges. Due to the partial release of the stress by the steps, the reconstruction along the steps is greater on the vicinal area than that on wide Au(111) terraces. Meanwhile, the discommensuration lines are disturbed by the steps and no longer run perpendicular to the close-packed direction, forming a "V" shape (see Figure 9.5d). The *fcc* domain width decreases when approaching the upper part of the step, and this is consistent with the fact that the *fcc* domain should be larger near the bottom of the step, since it is the bulk stacking of gold. Although the periodicity along the [−211] direction – that is, the terrace width – is dependent on the miscutting angle, the periodicity along the [0 −1 1] direction originated from the reconstruction, 7.2 nm, is almost invariant.

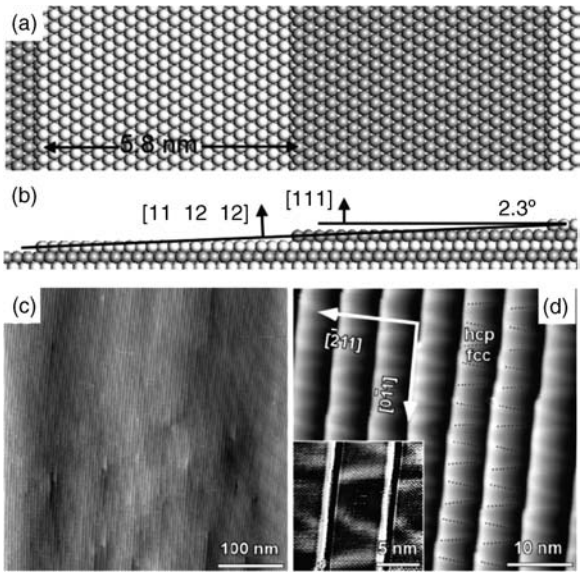


Figure 9.5 Au(11 12 12) vicinal surface with 2.3° miscut towards [−211] with respect to (111) surface. (a) Top view and (b) cross-section of Au(11 12 12); (c, d) STM images of Au(11 12 12) surface. The inset in panel (d) shows a zoomed image [29].

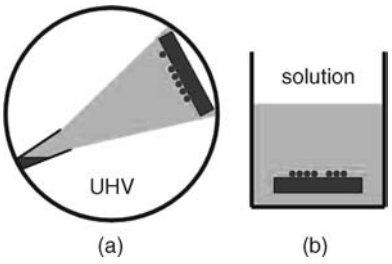


Figure 9.6 Preparation techniques for supramolecular templates. (a) Organic molecular beam deposition (OMBD) under UHV conditions; (b) Liquid-phase deposition in solution.

9.2.4 Surface-Supported Organic Supramolecular Assemblies

The above-discussed templates are either covalent-bonded or metal-bonded inorganic surfaces. Hence, organic supramolecular architectures that are self-assembled on surfaces bound by noncovalent interactions, will be introduced in the following subsections.

9.2.4.1 Preparation Techniques

Two methods are used for preparing surface-supported supramolecular architectures, namely organic molecular-beam deposition (OMBD) under UHV conditions, and liquid-phase deposition at solid–liquid interfaces (see Figure 9.6).

In OMBD, the organic compound is loaded into a crucible under UHV conditions and heated to temperatures at which the organic molecules will either sublime or evaporate onto the substrate surfaces. This technique has the following advantages:

High purity: Prior to OMBD growth, a high purity of organic materials is achieved by thermal gradient sublimation. For this, both the organic source and substrate are maintained under UHV with a background vacuum (typically on the order of 10^{-10} mbar) so that contamination from the environment is avoided.

Well-controlled kinetics: Both, the energetic parameters (i.e., the intermolecular and molecule–substrate interactions) and the kinetic parameters (including growth temperature and deposition rate) play key roles in determining the final structure of the supramolecular assembly. Notably, these kinetic parameters are adjustable during the deposition process.

Multicomponent deposition: Supramolecular structures consisting of multiple components may easily be produced by using more than one source in the UHV chamber. In addition, the temperatures of the sources can be controlled independently, such that it is possible to deposit different organic molecules or metal atoms at the same time or in sequence.

Combination with various characteristic methods: In addition to STM, other surface-analysis techniques, such as low-energy electron diffraction, photoemission spectroscopy and electron energy loss spectroscopy (EELS) are useful when investigating the structure and properties of surface-supported supramolecular

assemblies. These surface-analysis techniques can be integrated with the OMBD preparation chamber.

When using OMBD it is necessary to use sublimable or vaporizable molecules with a vapor pressure of $\sim 10^{-8}$ mbar at a temperature in the range from 60 to 600 °C, although this limits the application of OMBD to molecules that do not sublime, nor are thermally stable. In contrast, for liquid-phase deposition the molecules are deposited from solutions onto the substrate surfaces, which are themselves immersed in a solution of the molecules. Liquid-phase deposition is easier to use than OMBD; moreover, in biological systems it is also possible to mimic the environment when investigating biological molecules. Nonetheless, inert substrates such as highly oriented pyrolitic graphite (HOPG) and Au(111) will be required when using this technique, and the kinetic processes cannot be well controlled.

9.2.4.2 **0-D and 1-D Nanostructures**

The formation of molecular structures, ranging from single molecules to few-molecule aggregates, linear wires, extended 2-D compact layers and open networks up to 3-D architectures, on surfaces is dependent on the intermolecular interactions and the growth parameters. Well-defined zero-dimensional (0-D) and one-dimensional (1-D) structures are normally obtained using a submonolayer regime via OMBD, under UHV. At low temperature, the surface diffusion ability of molecules will be inhibited such that aggregation processes which originate from certain attractive interactions (e.g., van der Waals forces) are prevented, and this results in the presence of isolated molecules on the surfaces. Such aggregation processes are also prevented in the case of molecule–substrate systems by repulsive interactions, such as charge transfer [32–34]. The assembly of molecules with directed and selective noncovalent interactions, such as H-bonding and metal–ligand coordination coupling, is determined by the specific anisotropic feature of the interactions. A careful placement of functional groups that are capable of participating in directed noncovalent interactions will allow the rational design and construction of a wide range of supramolecular architectures assembled on surfaces. Few-molecule aggregates will be formed if the active interaction sites are self-saturated [31] (also D.Y. Zhong, unpublished results), whilst linear supramolecular wires will be formed in the case that the molecules possess two active interaction sites with anti-parallel directions. When molecules possess three or more active interaction sites in plane, 2-D networks will be formed.

Figure 9.7 shows the STM study of substituted porphyrin molecules adsorbed on a Au(111) surface. Cyanophenyl substituents, which possess directed dipole–dipole and H-bonding interactions, are used to control the molecular aggregation. Isolated single molecules, trimers and tetramers are observed when depositing the substituted porphyrin molecules without, with one, and with two cyanophenyl groups at the *cis* sites, respectively. However, when depositing the molecules with two cyanophenyl groups substituted at the *trans* sites, the anti-parallel configuration between the cyanophenyl substituents results in a linear arrangement and the formation of supramolecular wires [31].

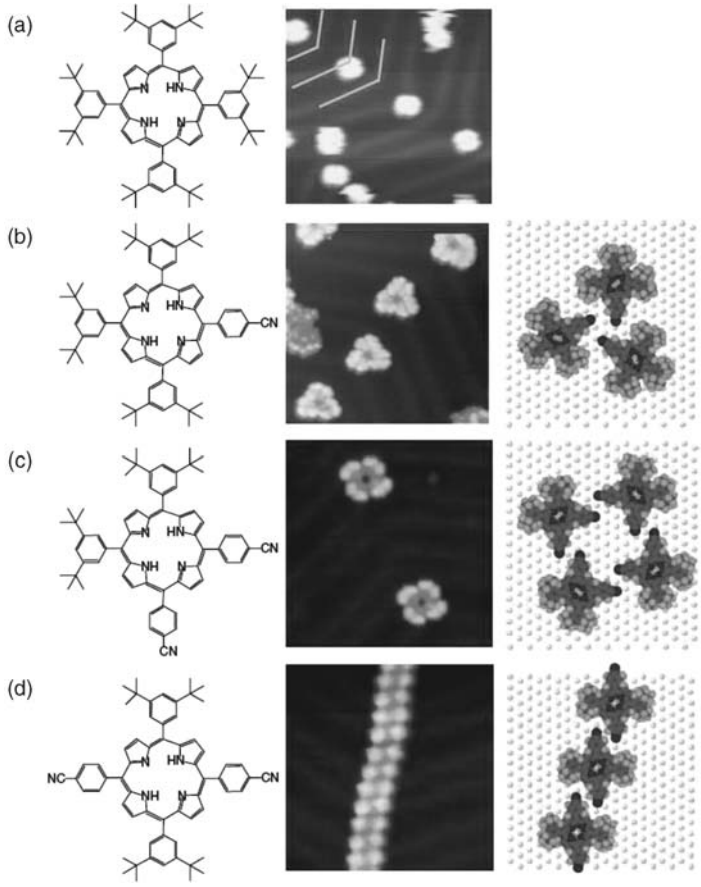


Figure 9.7 0-D and 1-D molecular nanostructures. (a) Monomers of the porphyrin derivative without cyanophenyl substituent; (b) Trimers of the porphyrin derivative with one cyanophenyl substituent; (c) Tetramers of the porphyrin derivative with two *cis* cyanophenyl substituents; (d) 1-D molecular wire of the porphyrin derivative with two *trans* cyanophenyl substituents. STM image area = $20 \times 20\,nm^2$ [31].

The formation of supramolecular wires is also directed by H-bonding and metal–ligand coordination [35–45]. The formation of double-row molecular assembly of 4-[*trans*-2-(pyrid-4-yl-vinyl)]benzoic acid (PVBA), on a Ag(111) surface results from the head-to-end OH...N bonding, as well as the CH...OH bonding between the rows [35, 45]. A similar coupled double-row architecture has been observed on the assembly of L-methionine on a Ag(111) surface [38, 39]. Here, the regular molecular gratings, which are formed at intermediate coverage with tunable periodicity at the nanometer scale by varying the methionine surface concentration, can serve as a template to guide the formation of metal atom lines via surface-state confinement at the Ag(111) surface [38]. In general, by optimizing the growth parameters, molecules with strong anisotropic interactions – including H-bonding, dipole–dipole coupling,

metal–ligand coordination, as well as van der Waals interactions – prefer to form 1-D structures. Yet, the strong anisotropy may induce the formation of 1-D chains that extend for over several hundred nanometers [40].

9.2.4.3 **2-D Molecular Patterns**

The self-assembly of organic molecules on surfaces is determined by the subtle balance between intermolecular and molecule–substrate interactions. In general, close-packed configurations are favorable for supramolecular assembly on surfaces due to the attractive van der Waals interactions between molecules, and between molecules and substrates. The latter effect is dominant in some cases, even though the molecules may possess repulsive interactions, the reason being that those configurations with more molecules adsorbed onto the surfaces will reduce the system total energy. However, by introducing certain directional and selective noncovalent interactions, such as H-bonding and metal–ligand coordination, it is possible to form open networks with voids of substrates. These porous networks attract intrigue in host–guest chemistry, since by combining the different noncovalent interactions with various strengths it is possible to form complicated structures that include hierarchical or multilevel architectures. Moreover, besides the energetic consideration of the systems, certain growth parameters – including growth temperature, rate and coverage, as well as the concentration in liquid-phase deposition – will play important roles, with such experimental parameters affecting the diffusion–aggregation and adsorption–desorption processes of adsorbed molecules. As the polymorphic feature is a common property of most molecule–substrate systems, different metastable structures are attained, depending on the growth parameters.

Molecular Design Due to the large variety of organic molecules, it is relatively easy to alter the supramolecular structures by molecular design. By adjusting the intermolecular and molecule–substrate interactions, it is possible to tune the final supramolecular structures which are at energetic minimum states by self-assembly under quasi-equilibrium conditions. On the basis of this concept, there are two main routes: (i) to alter the geometry of the molecules, and also the packing behaviors of the molecules adsorbed on surfaces via long-range and nondirectional van der Waals interactions; and (ii) to introduce directional and/or selective noncovalent interactions, such as H-bonds and metal–ligand coordination coupling. In this case, different H-bonds, including OH...O [46–49], OH...N [50, 51], NH...O [42, 43, 52–59], NH...N, [52–61], and CH...O [37, 50, 51, 62, 63], can be introduced in surface-supported supramolecular assemblies. Frequently, more than one H-bond is formed on each molecule, whilst a cooperative effect may enhance the binding between molecules [52]. Coordination coupling between transition metals and organic ligands is also used to adjust the supramolecular architectures on the surfaces [64, 65], such that the metal atoms or ions are either codeposited on the surface [66–75], from thermally activated diffusing adatoms on a metal substrate [36, 76–78], or from the solution in the case of liquid-phase deposition [41, 79].

Among these molecules, benzoic acid and its derivatives have been investigated intensely as model systems. Although the carboxyl group exhibited abundant

Figure 9.8 Noncovalent interactions of (deprotonated) carboxyl groups.

behaviors in the self-assembly process (Figure 9.8), it is also possible to form H-bonds between the carboxyl groups, and between the carboxyl groups and benzene rings. In this way, dimers or trimers are formed with either two or three H-bonds between two or three carboxyl groups, respectively. According to DFT calculations, hydrogen bonds in the cyclic dimer of benzoic acid, as are present in the "chicken-wire" network, are 1.5 kcal mol^{-1} stronger than those in the cyclic trimer in the gas phase, as are present in the "flower" structure [80]. On the other hand, the deprotonated species of carboxyl (carboxylate) act as ligands to coordinate with transition metal atoms.

The adsorption of benzoic acid (Figure 9.9, **1a**), which contains one carboxyl group substituting one of the six H atoms on a benzene ring, prefers an upright configuration at high coverage; that is, with the benzene ring perpendicular to the surface [81–85]. On TiO_2(110) surfaces, the oxygen atoms of the deprotonated species (benzoate) are bonded with the fivefold-coordinated Ti^{4+} cations to form an ordered overlayer, which is mainly determined by the relatively strong adsorbate–substrate interaction. Attractive interactions between the aromatic rings of the benzoates lead to the formation of dimerized benzoate rows along the [001] direction [81–85]. The flat-down configuration (with the benzene ring parallel to the surface) only exists at a very low coverage [81–85]. Similarly, on Cu(110) surfaces, H-bonded flat-down dimers are formed at a temperature below 150 K, while deprotonation takes place at a temperature above 150 K and benzoate dimers coordinating with a Cu atom in between are formed. At saturation coverage, however, the adsorption of benzoic acid on Cu(110) between 300 and 350 K results in benzoate species which is oriented perpendicular to the surface with a c(8×2) periodicity [83].

Increasing the number of substituting carboxyl group on the benzene ring, or substituting the benzene ring with a pyridine ring [iso-nicotinic acid (INA); **1b** in Figure 9.9], will have a dramatic influence on the assembly. In this case, it is possible to form more H-bonds in the flat-down configuration with a reduced system total energy, which makes the flat-down configuration more favorable than the upright configuration. By introducing the pyridine ring, the molecular self-assembly of INA on Ag(111) adopts a head-to-tail hydrogen-bond configuration from ring nitrogen atom to hydroxyl hydrogen, with the aromatic rings parallel to the surface plane [86]. Here, there are two sets of H-bonds: the primary head-to-tail H-bonds (OH. . .N) link molecules into long chains along the Ag[11-2], while the secondary hydrogen-bonding interactions link the carbonyl oxygen of one INA chain with the aromatic

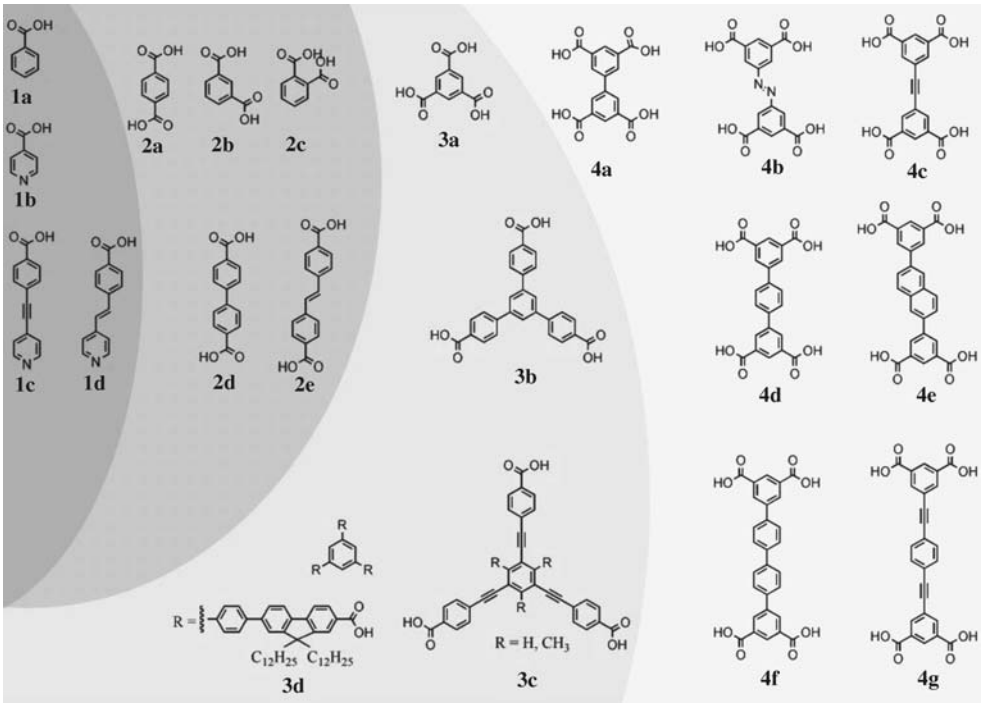


Figure 9.9 Benzoic acid derivatives. **1a**–**1d**, **2a**–**2e**, **3a**–**3d**, and **4a**–**4g** contain one, two, three, and four carboxyl groups, respectively. Abbreviations and references: **1a** [81–85]; **1b** (INA) [86]; **1c** (PEBA) [50]; **1d** (PVBA) [50, 87]; **2a** (TPA) [69, 72, 73, 88–94]; **2b** [90]; **2c** [90]; **2d** (BDA) [63, 73, 74, 87, 88, 95]; **2e** (SDA) [63, 75]; **3a** (TMA) [36, 46, 48, 49, 51, 68, 76, 80, 92, 93, 96–101]; **3b** (BTB) [99]; **3c** [102]; **3d** [47]; **4a** [103]; **4b** (NN4A) [104]; **4c** [103]; **4d** (TPTC) [105]; **4e** [106]; **4f** [106]; **4g** [103].

hydrogen on an adjacent chain. The benzoic acid derivatives containing two carboxyl groups have three isomers, depending on the two substitution sites: terephthalic acid (TPA; **2a** in Figure 9.9), iso-phthalic acid (**2b** in Figure 9.9) and phthalic acid (**2c** in Figure 9.9). The self-assembly of TPA on the Au(111) surface at room temperature under UHV conditions forms a highly ordered, close-packed layer at high coverage. The head-to-tail hydrogen bonds between the carboxyl groups of neighboring molecules are the dominant interactions in the molecular monolayers, and result in linear molecular chains, analogous to the bulk structure [91]. However, the length of the H-bridges exceeds that in the TPA bulk phase due to modulation of the substrate surface. A similar structure, with linear molecular chains stabilized by H-bonds, is formed by the assembly of TPA at liquid/solid interfaces [90, 91]. By comparison, the assembly of iso-phthalic acid molecules results in the zigzag configuration owing to H-bond formation. No ordered structure has been found for the assembly of phthalic acid [90, 91].

The assembly behaviors are changed by further increasing the number of carboxyl groups. Two coexisting phases – the "chicken-wire" and "flower" structures, both of

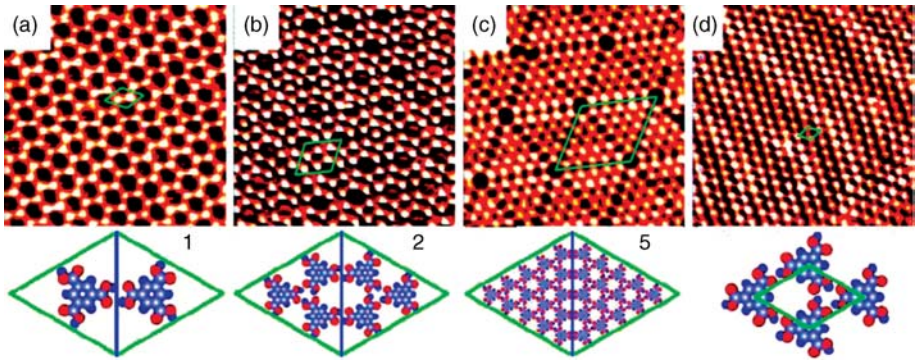


Figure 9.10 1,3,5-Benzenetricarboxylic (trimesic) acid (TMA) on Au(111). The distance of the pores increases by increasing the coverage. (a) "Chicken-wire" structure with only dimeric H-bonding; (b) "Flower" structure with mixed dimeric and trimeric H-bonding; (c) Porous structure similar to the "flower" structure, but with a larger distance between the pores; (d) Close-packed structure with only trimeric H-bonding. STM image area = $16.5 \times 16.5\ nm^2$ [46].

which are induced by directional hydrogen bonding – are formed by the adsorption of 1,3,5-benzenetricarboxylic (trimesic) acid (TMA; **3a** in Figure 9.9) to graphite or Au (111) surfaces under UHV conditions [49]. In the "chicken-wire" structure (Figure 9.10a), all molecules are connected by dimeric H-bonds (two H-bonds involving two carboxyl groups), to form sixfold molecular rings. In contrast to the "chicken-wire" structure, the "flower" structure can be seen as a closed packing of the sixfold rings (Figure 9.10b). Within the rings, the hydrogen bonds are formed in the same way as in the "chicken-wire" structure, whilst between the neighboring sixfold rings trimeric H-bonds are formed that involve three molecules. In both structures, all of the H-bonds have a length about 3 Å, which is well within the range of OH. . .O bonds (2.7–3.1 Å).

Consider, then, the situation in which the symmetry of TPA and TMA molecules is retained, while the molecular size is increased. The molecule 4,4′-biphenyl dicarboxylic acid (BDA; **2d** in Figure 9.9) shows the same twofold symmetry as TPA, but with two phenyl rings as the backbone. The behavior of BDA molecules assembled on Au(111) surface is similar to that of TPA, forming linear molecular chains stabilized by head-to-tail H-bonding between the carboxyl groups of neighboring molecules [95]. The case with threefold rotational symmetry as TMA is 4,4′,4″-benzene-1,3,5-triyl-tribenzoic acid (BTA), in which three 4′-benzoic acid groups are arranged around the central benzene core. In this case, 2-D supramolecular honeycomb networks of BTA with cavities of a larger internal diameter of 2.95 nm are formed by self-assembly on a Ag(111) surface at room temperature [51]. The configuration of the networks is similar to the "chicken-wire" structure of TMA self-assembled on graphite [49]. The symmetry rule is obeyed in the assembly of even more complicated molecules [47].

The benzoic acid derivatives with four carboxyl groups can form both a parallel network and a Kagomé network on the surfaces (Figure 9.11) [103, 105, 106]. In the

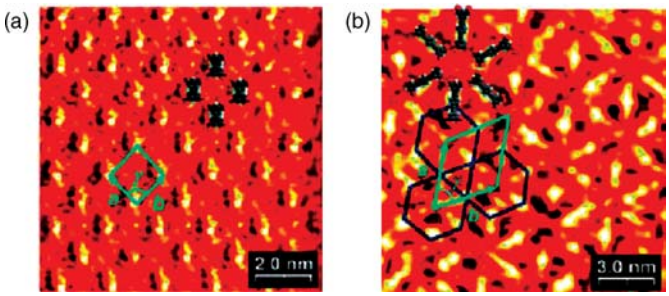


Figure 9.11 Tetra-acids self-assembled on HOPG (deposition from heptanoic acid solution, −1.5 V, 50 pA). (a) Parallel configuration of **4a** (see Figure 9.9); (b) Kagomé network of **4g** (see Figure 9.9) [103].

parallel network, each molecule forms four double H-bonds with four neighboring molecules at the corners, and all molecules show the same orientation. In the Kagomé network, three molecules, which are rotated $\pm 60^{\circ}$ with respect to each other, are H-bonded head-to-tail with a triangle-like feature. The molecules with short (e.g., biphenyl, **4a** of Figure 9.9) and long (e.g., four phenyl rings, **4f** of Figure 9.9) backbones adopt single configurations with long-range periodic order, while those with intermediate backbone lengths, such as *p*-terphenyl-3,5,3′,5′-tetracarboxylic acid (three phenyl rings, **4d** of Figure 9.9) and 5,5′-(1,2-ethynediyl)bis(1,3-benzenedicarboxylic acid) on graphite form mixed networks coexisting both parallel and Kagomé configurations [103, 105]. The specific molecular lengths allow their growth of one polymorph to be interrupted by a smooth transition to the alternative polymorph, without necessarily introducing defects in which molecular building blocks are missing, improperly oriented, or unable to form the optimal number of hydrogen bonds with neighbors [103]. As a result, a random, nonperiodic, entropically stabilized, rhombus tiling is formed in a 2-D molecular network of *p*-terphenyl-3,5,3′,5′-tetracarboxylic acid adsorbed onto graphite [105]. Similar 2-D nonperiodic tiling has been observed in the self-assembly of rubrene molecules in supramolecular pentagons, hexagons, and heptagons on a Au(111) surface [107].

In addition to the H-bond-dominated supramolecular assembly of benzoic acid derivatives described above, another important intermolecular interaction is that of metal–ligand interaction. The deprotonated species of carboxyl groups is able to form metal–ligand coordination with certain transition metal atoms and, in the presence of such interactions, the assembly of organic molecules on surfaces may be changed. Except for the H-bonded "chicken-wire" and "flower" patterns of TMA self-assembled on graphite and Au(111) [46, 76], the deposition of TMA on Cu(100) surfaces at 300 K results in the formation of cloverleaf-shaped $Cu(TMA)_4$ complexes, which are organized into a regular array at high coverage (Figure 9.12). The four deprotonated TMA molecules are bonded with the centered Cu atom through the carboxylate groups. The carboxylate ligands of TMA do not point directly towards the central Cu atom, and the oxygen atoms in the respective COO^{-} moieties are not equivalent, which indicates a unidentate coordination of the Cu atom [76]. For comparison, the

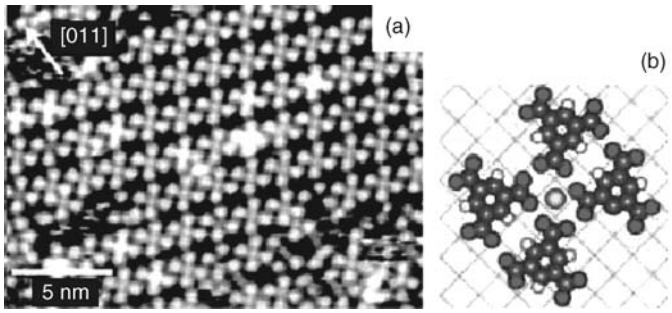


Figure 9.12 $Cu(TMA)_4$ complex formed by depositing TMA on Cu(100) at 300 K. (a) Regular array of $Cu(TMA)_4$ complex; (b) Model of the complex showing the coordination of the carboxylate groups of four TMA molecules with the central Cu atom [76].

deposition of TMA on a Cu(110) surface results in 1-D metal–organic coordination chains of TMA and Cu atoms, owing to the anisotropy of the substrate [36]. Dicarboxylic acids containing one (TPA), two (TDA) and three (DBA) phenyl rings show similar behaviors as TMA. Although these molecules, when self-assembled on Au(111) surface, adopt a H-bonded head-to-tail configuration [91, 95], porous coordination networks of TPA or TBA and iron atoms with variable cavity sizes are formed on the Cu(100) surface [72]. The codeposition of DBA molecules and Mo atoms, followed by annealing to 400 K, results in carboxylate–Mo coordination and the formation of rectangular-shaped or ladder-like networks, depending on the ratios of the deposited amounts of BDA molecules and Mn atoms [74]. In addition, the coordination interaction between the O atoms of carboxylate groups and Fe atoms has been confirmed by comparing the chemical shift of O 1s and Fe 3p peaks in X-ray photoelectron spectroscopy (XPS) measurements with various Fe : TPA ratios [69].

When compared to H-bonds and metal–ligand coupling, van der Waals interactions – which exhibit an indispensable effect on multilevel structures of biomolecules – are much weaker. In some cases, the van der Waals interactions have a dramatic influence on supramolecular assembly, even in those systems where H-bonds or coordination coupling are predominant. For example, a slight variation of a submolecular alkyl group, which is not involved directly in the hydrogen bonding, may have a pronounced effect on the self-assembled surface nanostructures, thus causing a change in the molecular nanostructures obtained, from extended periodic rows to localized chains and polygonal clusters. This change is attributed to subtle differences in van der Waals interactions between the alkyl side chains found on some of the compounds, an insight that can be employed to control the formation of self-assembled molecular surface nanostructures when multifunctional groups are involved in the molecular building blocks [108]. Similarly, the 2-D pattern formation of hydrogen-bonding iso-phthalic acid derivatives at the liquid/solid interface has been investigated by using STM. By varying the location and nature of the alkyl substituents on the aromatic core, in combination with the intrinsic hydrogen-bonding properties of the iso-phthalic acid units, the 2-D supramolecular ordering has been controlled, leading to several different motifs [59].

Complicated supramolecular structures may occur in the systems with only van der Waals interactions. In such systems, subtle intermolecular and molecule–substrate interactions result in a superstructure with several and/or up to several tens of molecules, the relative orientation and positions of which are different from each other. A long-range orientational order of C_{60} monolayers on Au(111) has been observed with a unit cell comprised of 49 molecules adopting 11 different orientations [109]. In this case, intermolecular interactions play a major role in stabilizing the superlattice, while the substrate induces minute changes in the orientation of the C_{60} molecules. Diferrocene molecules, which contains two ferrocenyl groups bridged by an alkyl chain, show van der Waals intermolecular interactions and are weakly bound on certain surfaces, such as Cu(110). The assembly of diferrocenes on Cu(110) exhibits 2-D multilevel structures up to quaternary, resulting from the subtle balance of intermolecular and molecule–substrate interactions that have comparable strength, as well as the mismatch of the molecular packing and the surface atomic periodicity. In such a multi-periodicity modulated system, neither the intermolecular nor molecule–substrate interactions solely dominate the molecular assembly, and this results in complicated supramolecular architectures. The multilevel assemblies demonstrate site-selective properties for the adsorption of guest molecules (D.Y. Zhong, unpublished results).

Substrate Effects Substrates are not only the support for the assembly of organic molecules, but also play a key role in supramolecular assembly. On the one hand, the surface atomic potential with specific periodicities will affect the arrangement of adsorbed molecules. The molecule–substrate interactions include long-range van der Waals interactions, dipole–dipole interactions, and covalent bonding in some cases, which cause the molecules to prefer specific adsorption sites and orientations. On the other hand, molecule–substrate interactions will alter the features of molecules themselves, such as their structure, conformation, and electronic state. For example, charge transfer and substrate-induced dissociation are common phenomena in molecule-on-surface systems. The substrate-induced modification of molecules will further influence intermolecular interactions, and finally the supramolecular assembly. The assembly of BDA (**2d** of Figure 9.9) on Au(111) and Cu(100) surfaces are compared in Figure 9.13. On the Au(111) surface, linear molecular chains are formed at room temperature, with H-bonding between the carboxyl groups of neighboring molecules, analogous to PBA [95]. On a more reactive Cu(100) surface, however, the carboxyl groups are deprotonated owing to the strong molecule–substrate coupling, and this results in a square-type structure through the H-bonds between the carboxylate and the benzene ring, with the molecules alternatively 90° rotated [63].

Growth Parameters The *temperature* of the molecular assembly, including that during growth and post-growth annealing, represents one of the most important experimental factors to directly affect the supramolecular assembly on surfaces [51, 110, 111]. The temperature affects not only the diffusion ability of molecules on surfaces, but also the ability of conformation change and bond dissociation [112]. In this case, 2-D supramolecular honeycomb networks of BTA with cavities of internal

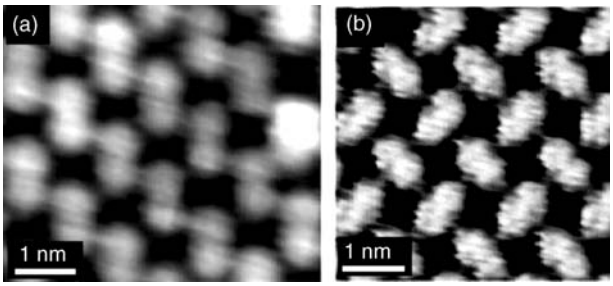


Figure 9.13 Substrate effect: BDA on Au(111) and Cu(100). (a) Linear molecular chains of DBA on Au(111); (b) Orthogonal configuration of deprotonated DBA molecules on Cu(100) [63, 95].

diameter 2.95 nm are formed on a Ag(111) surface at room temperature, in analogy to the "chicken-wire" structure of TMA self-assembled on graphite. However, annealing to a higher temperature results in two sequential phase transformations into closer-packed supramolecular arrangements, which are associated with a stepwise deprotonation of the carboxylic acid groups.

Coverage is another factor which affects molecular self-assembly [46, 60, 113]. As noted above, 0-D and 1-D nanostructures are both formed at lower coverage. However, by varying the coverage from 0.3 up to 1 ML, when TMA is assembled on an Au(111) surface under UHV conditions it forms a series of supramolecular structures, among which "chicken-wire" and "flower" are two special cases (see Figure 9.10). All observed assembling structures formed hexagonal porous networks that are well-described by a unified model in which the TMA molecules inside the half-unit cells (equilateral triangles) are bound via trimeric H-bonds, and all half-unit cells are connected to each other via dimeric hydrogen bonds. These porous networks possess pores of 1.1 nm diameter, and the interpore distance is tunable from 1.6 nm at a step size of 0.93 nm [46].

Solvent Effects In the case of molecular systems self-assembled at liquid–solid interfaces, the solvent concentration should be considered [99, 114–120]. A systematic study on the concentration-dependent formation of surface-confined 2-D networks at the interface of HOPG and 1,2,4-trichlorobenzene shows the competition of two polymorphs. Here, the building blocks are alkoxylated dehydrobenzo[12]annulenes (DBAs), which form 2-D porous networks when the alkoxy chain is less than 12 carbon atoms long [115], and preferentially form close-packed linear structure DBAs with longer alkoxy chains. By adjusting the DBA concentration in solution, the ratio of the two polymorphs can be controlled such that either a regular 2-D porous honeycomb network (at low concentrations) or a dense-packed linear network (at high concentrations) is formed [114]. The competition is related to the chemical potentials of the two surface phases and solution.

Multicomponent Assembly Supramolecular assemblies containing more than one component can be prepared by both OMBD and liquid-phase deposition [98, 99, 121, 122]. In this case, bicomponent and multicomponent assemblies usually result in

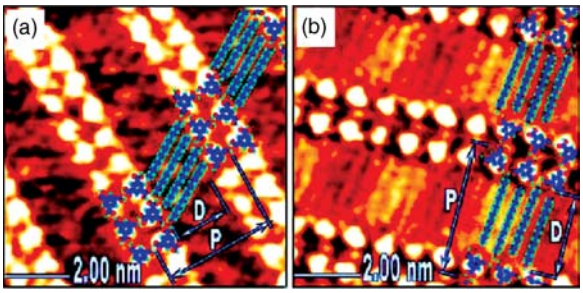


Figure 9.14 STM images of bicomponent assembly of TMA and normal aliphatic alcohols at the heptanoic acid solution/HOPG interface. These architectures contain alternative hydrophilic tapes and hydrophobic inter-tape spaces with adjustable dimensions. (a) TMA and 1-hexadecanol (−1.4 V, 200 pA). Distance between two successive tapes D = 1.7 nm; periodicity of the pattern P = 3.4 nm; (b) TMA and heptadecanol (−1.1 V, 200 pA). D = 2.4 nm, P = 4.1 nm [98].

architectures that differ from the structures assembled from single components. The complexity of the supramolecular structures is dramatically increased, which makes it more feasible to obtain desired structures by the careful molecular design of each component participating in the assembly.

The hydrogen bond-associated assembly of TMA and a variety of normal aliphatic alcohols have been studied [98]. The motif of the mixed assembly (Figure 9.14), which consists of alternative double-row TMA molecular tapes and aliphatic chains running parallel to each other, is distinctly different from the hexagonal patterns normally favored by TMA itself [49]. Its periodicity is proportional to the length of the alcohol, and thus can be modulated with high predictability by changing a readily available component. Very different properties of the tapes (hydrophilic) and inter-tape spaces (hydrophobic) create an opportunity to guide the position-specific adsorption of other atoms and molecules [98].

The coadsorption of benzenetribenzoic acid and trimesic acid at the liquid–solid interface in two different solvents (heptanoic and nonanoic acid) results in six nondensely packed monolayer phases with different structures and stoichiometries stabilized by intermolecular hydrogen bonding between the carboxylic acid functional groups, depending on the concentrations in the binary solutions. Moreover, phase transitions of the monolayer structures, accompanied by an alteration in the size and shape of cavity voids in the 2-D molecular assembly, could be achieved by *in situ* dilution. The emergence of the various phases could be described by a simple thermodynamic model [99, 121].

9.2.4.4 **Porous Networks**

Surface-supported porous networks are related to the supramolecular architectures with ordered separated voids where the substrate surface is exposed. Due to the different physical and chemical reactivities at the nanosized voids and molecule links, such porous networks offer a platform to perform site-selective physical and chemical processes at molecular scale. A detailed discussion of porous networks

combined with nanostructure fabrications using the porous networks as template will be discussed in Section 9.4.2. A review of supramolecular nanoporous networks self-assembled on surfaces has recently been produced [123].

9.3
Surface-Supported Nanostructures Directed by Atomic-Level Inorganic Templates

The formation of nanostructures guided by inorganic templates, including surface reconstructions, stepped surfaces, and strain-relief patterns is discussed in the following subsections. Supramolecular network-related nanostructures will be discussed in Section 9.4.

9.3.1
Reconstruction

Au(111)-22 $\times \sqrt{3}$ Herring-Bone Pattern Ordered arrays of Ni nanodots are formed due to preferential nucleation on the elbow sites of the Au(111) herring-bone reconstruction patterns [124]. Associated by the ordered arrays of transition metal nanodots, complicated organic–inorganic hybrid nanostructures have been fabricated on reconstructed Au(111) surfaces [125–129]. Meanwhile, single-molecule or few-molecule aggregates can also be anchored on the elbow sites of the herring-bone patterns, without metal nanodots [130–132]. Depending on the geometry and the intermolecular interactions, the elbow sites and the *fcc* regions may serve as the active sites for the growth of organic nanostructures. Écija *et al*. have investigated the crossover of site-selectivity in the adsorption and self-assembly of phenyl-C61-butyric acid methyl ester (PCBM) on the herringbone-reconstructed Au(111) surface as a function of the coverage (Figure 9.15) [133]. Initially, the molecules nucleated at the elbow sites, but with increasing coverage long, parallel, isolated zigzag 1-D wires were formed exclusively at the *fcc* regions. However, a compact arrangement of molecules with double-molecular rows was formed and the site-selectivity lost when the coverage was further increased.

Cu(110)-(2 $\times$ 1)O and the Stripe Pattern In comparison to the oxygen-adsorbed regions, the bare Cu(110) regions are normally more active in adsorbing organic molecules. The template effect of the stripe pattern for the fabrication of 1-D organic nanostructures has been demonstrated by Besenbacher and coworkers [8], with well-ordered arrays of long molecular chains of "Single Lander" molecules being self-assembled on the Cu(110) stripes. By controlling the width of the nanotemplate, it was possible to select the adsorption orientation of the molecules, and thereby steer their alignment along the specific direction of the template (Figure 9.16). A similar template effect of the stripe patterns has been also applied for the assembly of α-quinquethiophene [134], rubrene [135], and *para*-Sexiphenyl [136].

Si(111)-7 $\times$ 7 The use of semiconductor substrates is more attractive than metal substrates for potential applications in the field of nanoelectronics. The Si(111)-7 $\times$ 7

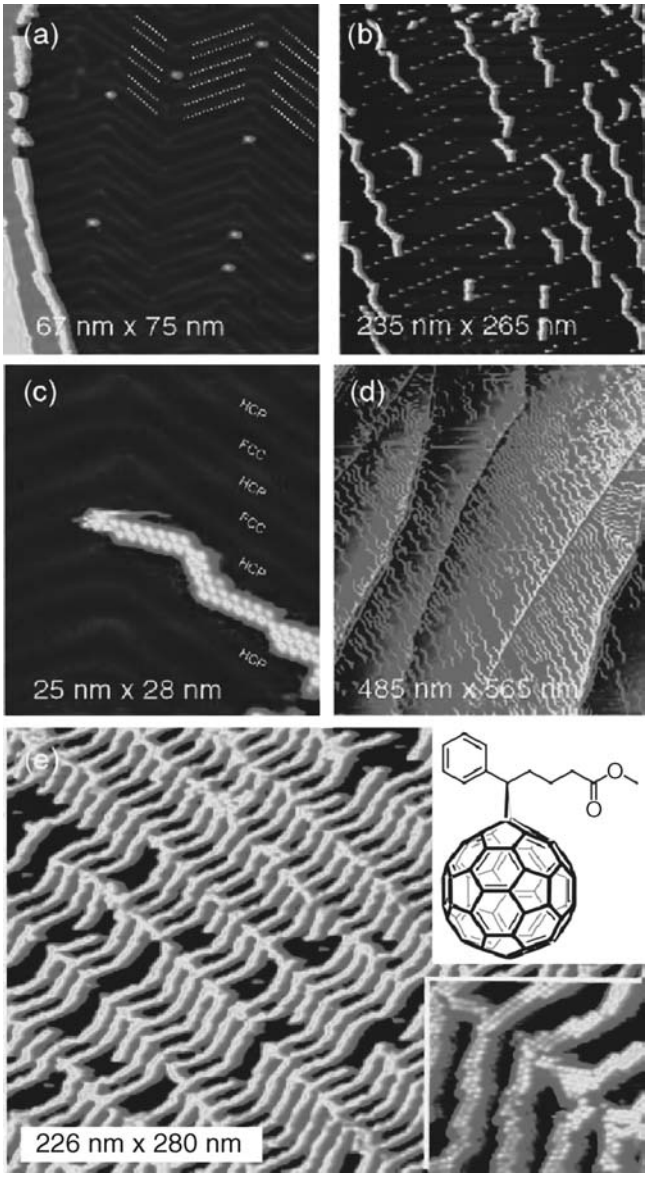


Figure 9.15 Nanostructures of PCBM guided by Au(111) herringbone template with different coverages. (a) <0.1 ML; (b) 0.1 ML; (c) A zigzag structure formed at the *fcc* region; (d) 0.3 ML; (e) 0.4 ML. Upper inset shows the molecular formula of PCBM; Lower inset shows high-resolution image of 2-D network of PCBM molecules (refer to Figure 9.2 for clean Au(111) surface) [133].

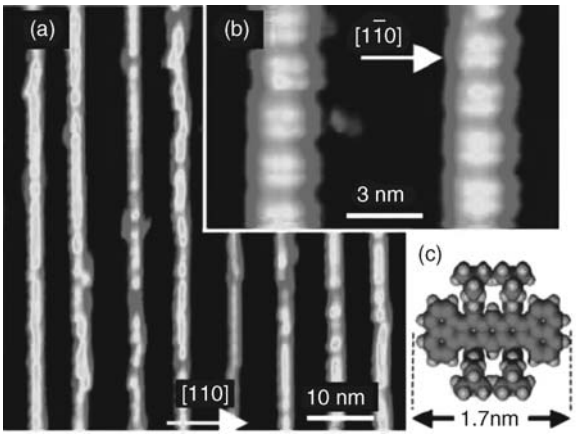


Figure 9.16 Molecular chains formed by the deposition of a "Single Lander" molecules on Cu (110)-(2 × 1)O stripe pattern. Molecules adsorb exclusively on the bare Cu stripes. (a, b) STM images; (c) Model of the "Single Lander" molecule with a polyaromatic hydrocarbon central board and four 3,5-di-*tert*-butylphenyl substituents [8].

is a well-known reconstruction with a surface lattice constant of 2.7 nm. By carefully controlling the kinetic parameters, including the growth temperature and the flux, well-ordered indium nanocluster arrays with identical size could be formed [137]. However, the template effect of the Si(111)-7 × 7 surface cannot simply be applied to the assembly of organic nanostructure, due to the strong interactions of the Si dangling bonds with organic molecules [138]. The diffusion ability of the deposited molecules is also restricted by the surface reaction, such that the molecules are normally immobilized before they move to an energetically favorable site. For example, the deposition of C_{60} molecules on the Si(111) surface resulted in a disordered first layer [139, 140], despite the faulted half and unfaulted half being the most favorably adsorbed sites in the case of a single molecule [141].

One way to avoid adsorbate–surface interactions on Si(111)-7 × 7 surfaces is to use zwitterionic molecules; these are neutral, but carry formal positive and negative charges on different atoms. The negative site of a zwitterionic molecule acts as a electrostatic shield to prevent any reaction of the electron-deficient Si adatoms with the electron-rich carbon atoms of the organic molecules [142]. This strategy has been verified by Makoudi *et al.*, who chose 4-methoxy-4′-(3-sulfonatopropyl)stilbazolium (MSPS) as a model zwitterionic molecule. The MSPS molecules are terminated by a negatively charged SO_3^- group, which acts as an electrostatic shield that protects the organic molecules against the dangling bonds of the surface. At low coverage, the molecules prefer to adsorb at the faulted half cells with a star-shaped configuration (containing three MSPS molecules), indicating a more or less site-selectivity of the reconstructed surface [142, 143].

The reactive surface may also be passivated by silver or boron atoms, although in most cases passivation of the reactive surface may result in a loss of site-selectivity [139, 144, 145]. When Xu and coworkers demonstrated the feasibility of control-

lable growth of ordered molecular nanostructures on passivated Si(111)-7 × 7 surfaces, ordered 2-D Cu clusters were first formed at the faulted half-cells, before the deposition of organic molecules. The thiophene molecules were shown to bind preferentially to the copper clusters through the S−Cu interaction, and this resulted in large-scale 2-D thiophene molecular nanostructures that followed the ordered Cu cluster pattern [146].

9.3.2 Strain-Relief Epitaxial Layers

The dislocations in strain-relief layers grown on lattice-mismatched surfaces often repel adsorbates diffusing over the surface, such that they may serve as templates for the confined nucleation of nanostructures from adsorbed atoms or molecules. Highly ordered, 2-D Fe and Ag nanostructure arrays are fabricated on the strain-relief patterns of double layers of Ag on the Pt(111) surface [15]. At a reduced growth temperature (110 K), the Ag atoms are preferentially nucleated within the *fcc* areas (distorted hexagons), owing to a stronger binding to the *fcc* areas than to the *hcp* areas [147]. The ordered nucleation is accompanied by an enhanced size uniformity. The repulsive nature of the dislocations and the attraction towards specific sites within the unit cell represent the key properties for transferring the periodicity of the dislocation network to a highly ordered 2-D island superlattice.

The application of the same templates to organic materials was achieved by Aït-Mansour *et al*. [18, 148], who deposited C_{60} molecules on strain-relief patterns induced by two monolayers of Ag on Pt(111). At room temperature, nucleation of the C_{60} islands took place on both the *fcc* and *hcp* domains of the template. Moreover, the C_{60} molecules were sufficiently mobile on the template surface to cross the dislocations and to self-assemble into large, hexagonally close-packed 2-D islands [148]. The loss of site-selectivity for C_{60} implied that the well-established repulsive character of the crossing dislocations was not sufficiently strong to prevent significant molecule diffusion across the discommensuration lines from *fcc* sites into *hcp* sites at room temperature. In order to organize the organic molecules to a regular structure that followed the strain-relief pattern, a technique was developed to fabricate a new type of nanotemplate surface that consisted of a well-ordered hexagonal array of one monolayer-deep holes, with a tunable size of about 4 nm^2 and a fixed spacing of 7 nm, based on the strain-relief trigonal network formed in the 2 ML Ag on Pt(111) system [18]. The removal of about 0.1 ML of the Ag top layer of this surface structure by He- or Ar-ion sputtering, led to the formation of nanoholes at specific domains of the trigonal network, which were stable at room temperature. The regularly distributed C_{60} nanoclusters trapped in the holes, replicating the periodicity and hexagonal symmetry of the nanohole template surface, were formed by the deposition of about 0.1 ML of C_{60} molecules at room temperature.

Another important strain-relief pattern, which shows a template effect for organic molecules, is the boron nitride nanomesh grown on Rh(111) and Ru(0001) surfaces. Single molecules such as copper phthalocyanine (CuPc) can be trapped in holes of such nanopatterned surfaces, without the formation of strong covalent bonds. The

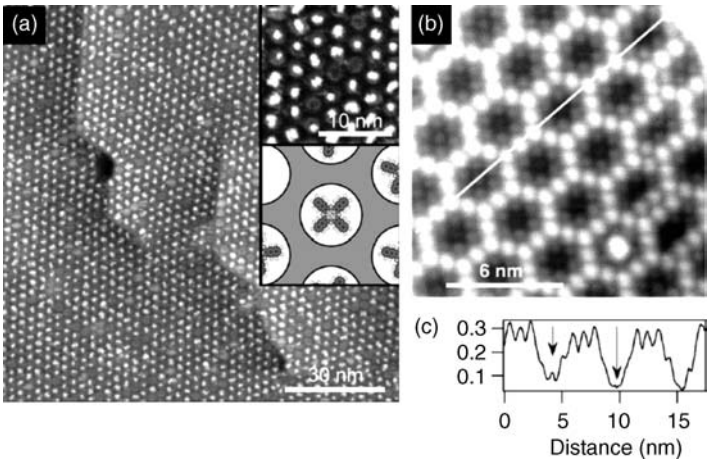


Figure 9.17 Organic molecules on BN nanomesh. (a) Nc molecules located at the center of the holes. The insets show a zoomed STM image (top) and model (bottom). The arrows show two holes with one molecule and empty, respectively; (b) C_{60} molecules decorated on the BN nanomesh; (c) Profile along the line in panel (b) [11, 14].

holes were identified as regions of low work function, with the trapping potential being localized at the rims of the holes [149]. The deposition of planar naphthalocyanine (Nc) molecules (the diameter of which at 2 nm was comparable to that of the nanomesh pores) onto the nanomesh at room temperature resulted in well-ordered arrays with the same periodicity as the nanomesh (3.22 nm) (Figure 9.17) [14]. In analogy to CuPc, the individual Nc molecules became trapped inside the pores with high site-selectivity, such that the molecule–substrate interactions dominated the adsorption behavior and the intermolecular interactions were relatively weak. The trapped molecules exhibited a very low mobility at room temperature, with only rare hopping to a neighboring pore, indicating a rather high trapping potential. In the case of C_{60}, however, the centers of the holes were the least stable adsorption sites [11]. Following the room-temperature deposition of approximately 1 ML of C_{60} molecules, the mesh wires were decorated by lines of individual molecules, and either six or seven molecules were adsorbed inside the holes, while the hole centers remained almost empty and were rarely occupied by one C_{60} molecule. Nevertheless, the periodicity of the mesh supercell was retained. This different behavior of C_{60} on the BN nanomesh might result from its relatively stronger intermolecular interaction compared to CuPc and Nc.

9.3.3 Vicinal Surfaces

Vicinal surfaces as natural templates for the fabrication of 1-D nanostructures have been investigated intensively during the past decades. The systems studied have included the growth of metal nanowires and nanodots [22, 150–154], wide-band-gap

materials such as CaF_2, NaCl, MgO [155], and the alignment of carbon nanotubes (CNTs) [156]. Recently, organic nanostructures guided by vicinal surfaces have been prepared; by carefully controlling the arrangement of kinks on a vicinal surface [22], or with a combination of surface reconstruction patterns, it is possible to obtain ordered 2-D arrays of nanosized objects.

Form an energetic point of view, the step edges are active sites for the nucleation of adsorbates, due to the higher coordinates and/or density of states of electrons. However, the kinetic parameters, including the growth temperature, growth rate and coverage, are also important for controlling the nanostructures. During the growth of metal wires on vicinal metallic or semiconducting surfaces, for example, the optimized growth temperatures should be sufficiently high as to ensure a smooth wire formation, but sufficiently low as to avoid any interlayer diffusion of adatoms [151]. Yet, the surface alloy effect should also be taken account [22]. The solid-state reaction between Fe and Si takes place even at room temperature [157] and, depending on the amount of deposition, either single- or double-row chains of Au and Ag can be grown on the vicinal Si(5 5 12) surface [152].

Vicinal gold surfaces have been used as substrates to grow organic materials [29, 158–162]. In this case, the template effect depends heavily on the chemical and geometric features of the molecules, which determine the interactions with the substrates. It has been reported that C_{60} molecules can recognize the substrate template of Au(433), [159] Au(11 12 12), [29] and Au(788), [158], but that there is no site selectivity on the Au(788) surface for perylene-3,4,9,10-tetracarboxylic-dianhydride (PTCDA) [163].

Xiao *et al.* reported the formation of a regular C_{60} nanochain lattice with long-range order on vicinal Au(11 12 12) surfaces (Figure 9.18a and b) [29]. Here, the C_{60} molecules were sublimated from a Kundsen-cell-type evaporator on the surface at room temperature such that, with a coverage of 0.1 ML ($\sim$0.1 nm^{-2}), well-ordered arrays of short molecular chains containing between two and six C_{60} molecules in each chain were formed. The periodicities of the array were unique to the rectangle superstructure of the vicinal Au(11 12 12) surface; that is, 5.8 nm in the [−2 1 1] direction and 7.2 nm in the [0–11] direction. The chains were located at the lower step edges of the *fcc* domains, where the molecules were preferentially nucleated. The template effect of the surface was lost when increasing the coverage up to 1 ML, however, at which close-packed ordered layer of C_{60} molecules were formed over the entire surface, accompanied by small islands of the second layer.

In the case of the Au(788) surface, a similar template effect has been investigated by Berndt and coworkers (Figure 9.18c and d) [158]. At close to 1 ML coverage, rather than a continuous close-packed molecular layer a well-ordered rectangle array of small single layer noncoalescing islands was observed. The islands, which consisted of approximately 20 molecules, were located at the *fcc* regions of the reconstructed surface, across the step edges. The different ML behavior on the two vicinal surfaces might be attributed to the width of the terraces (see Section 9.2.3). Meanwhile, it was clear that preparation parameters such as deposition rate should also be taken into account. Notably, a smaller deposition rate normally results in more ordered structures due to the efficient diffusion of adsorbates and a full relaxation of the

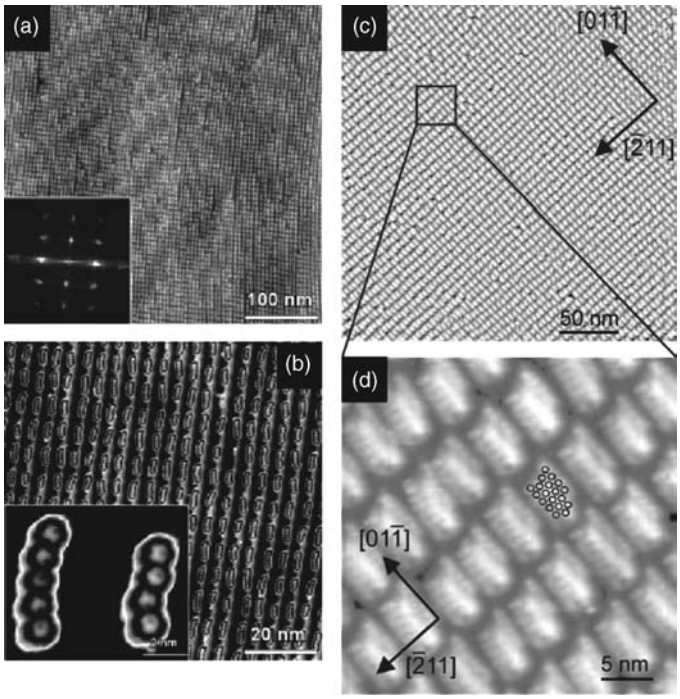


Figure 9.18 C_{60} nanostructures on vicinal Au (11 12 12) and Au(788). (a, b) Highly regular 2-D superlattice of C_{60} nanochains on the Au (11 12 12) template surface after deposition of ~0.1 ML at room temperature. The inset in panel (a) shows the Fourier power spectrum. The inset in panel (b) shows a high-resolution STM image of the nanochains. (c, d) Periodic C_{60} nanomesh on Au(788) at a coverage about 0.9 ML, deposited at room temperature [29, 158].

deposited layer, whereas an increase in the deposition rate might result in a loss of long-range order.

As discussed in Section 9.2.3, faceting takes places on unstable vicinal surfaces, and this results in stable facets with different surface atomic structures and different orientations. In some cases, the facets may be regularly organized, but by carefully selecting the molecules they may show selectivity for the separation of different molecules, due to differences in the binding energies. This concept has been realized by the coadsorption of PTCDA and 2,5-dimethyl-*N*,*N*′-dicyanoquinonediimine (DMe-DCNQI) on the (111) and (221) facets of a Ag(775) substrate [164]. The selectivity was shown to depend on the deposition sequence. When PTCDA was first deposited at room temperature the molecules were adsorbed onto both facets, though with an ordered structure on only the (221) facets. The subsequent deposition of DMe-DCNQI led to a disappearance of the ordered structure. Finally, annealing at 330 K resulted in only PTCDA on the (111) facets, with a mixture of both molecules on the (221) facets. In the second sequence, however, the molecules were deposited in reverse order. When DMe-DCNQI was deposited first at room temperature, no ordered structures were found by low-energy electron diffraction (LEED), but when

PTCDA was deposited next the system was annealed at 340 K. The final result showed ordered monolayers of PTCDA exclusively on (111) facets and DMe-DCNQI exclusively on (221) facets. with LEED analysis at different energies confirming the ordered structures on the two facets. In a final experiment, only DMe-DCNQI was deposited, and the sample was annealed at 318 K; this led to the molecule being adsorbed exclusively on the (221) terraces. Subsequent DFT calculations indicated that the PTCDA was bound more strongly on the (111) facet than on the (221) facet (0.54 versus 0.22 eV). DMe-DCNQI formed even stronger bonds on either of the facets, with a small preference for the stepped (221) facet (1.46 versus 1.36 eV). In fact, the stepped (221) facet proved to be favorable for DMe-DCNQI bonding with N atoms at the step edges, but unfavorable for planar PTCDA, which was longer than the narrow terraces so that the O terminators could not reach the step for extra bonding and the bonding was weakened by the presence of the step.

It is well known that the adsorption of certain species may induce the faceting of surfaces, due to an interplay between the molecules and the substrates [165–170]. Regular nanostructures may be obtained on faceted surfaces induced by adsorbates. A sequence of (115) and (001) nanofacets may be formed on the vicinal Cu(119) by the deposition of pentacene at room temperature, followed by annealing to 150–190 °C [171]. The faceted surface appears as a regular sequence of parallel stripes running along the [−110] direction for up to several tens of nanometers, and consists of a sequence of (115) and (001) facets tilted a few degrees off the (119) plane. The opposite tilt angle of the two facets gives rise to V-shape ripples at the surface, with dimensions in the range 2–5 nm. On (115) facets, the molecules are organized into parallel regular chains, with the long axis aligned along the [−110] direction, while on (001) facets about 50% of the molecules are aligned perpendicular to the [−110] direction. The faceting transition induced by the adsorption of pentacene molecules is thermally activated, and the annealing process is preliminary. Without annealing, long-range-ordered pentacene chains assembled on the Cu(119) vicinal surface have been obtained [172]. Pentacene aligns on the step edges of the Cu(119) vicinal surface, resulting in the formation of a long-range-ordered layer of unidirectional molecular chains. The planar molecule adopts a flat adsorption geometry, with the long molecular axis aligned along the steps, and the benzene units centered on the copper hollow site.

Stepped surfaces are also used as templates for directing the growth of organic–inorganic hybrid nanostructures. For example, PTCDA molecules deposited on a vicinal Ag(10 8 7) surface induce faceting, with the formation of (111) regions and step-bunched regions. The molecules are selectively adsorbed onto the step-bunched region, leaving the (111) region free. However, following the further deposition of about 1.1 ML iron on the striped structure, the iron atoms are adsorbed exclusively onto the PTCDA covered facets and small disk-shaped islands with a diameter of 4–6 nm and height of 0.5 nm are formed [173]. Arrays of iron nanowires have been fabricated by Lin *et al*. in a three-step process [174] where first, a silicon template with a regular array of straight steps was prepared by annealing vicinal Si(111) in a specific temperature sequence. Continuous CaF_2 stripes were then grown on top of a CaF_1/Si (111) surface, after which Fe nanowires in the CaF_1 trenches between the CaF_2 stripes

were formed via the selective adsorption of ferrocene and photolysis into Fe. This method has been observed for a variety of other molecules, and is today emerging as a general technique for growing 1-D nanostructures of transitional metals and other materials using CVD [175].

9.4
Surface-Supported Nanostructures Directed by Supramolecular Assemblies

9.4.1
Polymerization

The 2-D polymerization of organic molecules represents a possible means of fabricating stable, surface-supported nanostructures. These polymerized structures (especially linear structures) may have potential applications in molecular electronics as a form of candidate for nanowires to connect various switching elements. Moreover, polymerized organic thin films have huge potential as materials for use as field-effect transistors (FETs), rectifiers, photoconductors and light-emitting diodes (LEDs) [176, 177]. Studies on the polymerization of organic monomers on single crystal surfaces would be valuable for understanding the formation, propagation, and properties of the polymerized structures. On the basis of its ability to provide unprecedented atomic-resolved information, and to initiate local polymerization by applying a pulsed sample bias, STM has become a powerful tool in studies of low-dimensional polymerization. In the following subsections, attention will be focused on STM studies of 2-D polymerization on single-crystal surfaces.

9.4.1.1 Polymerization of Diacetylenes

The polymerization of diacetylenes is a typical and early reported example of the fabrication of linear polymerized structures on an atomic flat substrate. The critical condition of this reaction is the relative orientation and distance between the adjacent diacetylene monomer units (Figure 9.19). In 1997, Grim *et al.* showed that the polymerization of diacetylenes could be induced at the liquid/substrate interface following the irradiation with UV light of monolayers of diacetylenes containing an isophthalic acid derivative [178]. Monolayers of the isophthalic acid derivative were first prepared on HOPG surfaces, in which the diacetylene functional groups were packed close to each other, with the desired orientation. When, following UV irradiation (254 nm), the monolayers were reinvestigated using STM, some domains in the monolayers were seen to be replaced by polymerized structures, but not *entire* monolayers. This effect was confirmed by the increase in distance between the isophthalic acid groups, and also by the change in contrast of the polydiacetylene region (Figure 9.20).

This type of surface-supported polymerization was also observed by others [179–181]. For example, the groups of both Okawa and Wan reported the polymerization of diacetylenes in the self-assembled monolayers (SAMs) of 10,12-pentacosadiynoic acid on HOPG surfaces after UV irradiation [182, 183]. One interesting

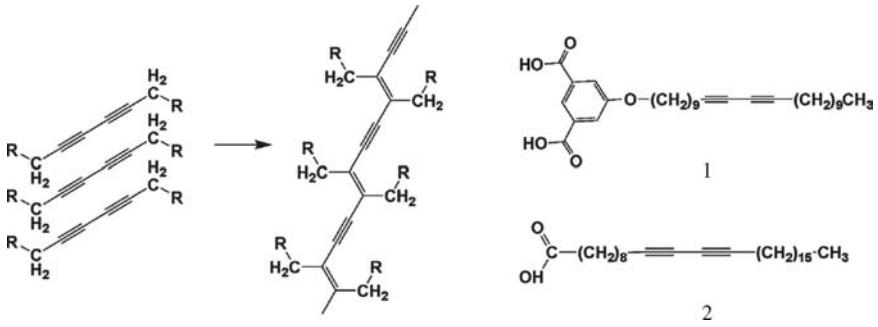


Figure 9.19 Polymerization of diacetylenes. Compounds 1 and 2 are two derivatives of diacetylenes which have been successfully observed after polymerization on single-crystal surfaces.

observation reported by Okawa's group was the controlled behavior of polymerization when using a STM tip [182, 184]. In this case, a negative-pulsed sample bias (−4 V in height, 5 us in width) was applied via the STM tip at a special point on top of the diacetylene groups. STM measurements subsequently demonstrated the presence of a polymerized line that highlighted an enhanced tunneling probability compared to the remainder of the SAMs. This line started from the point where the pulsed sample bias was applied, and terminated at an artificial defect made by a positive-pulsed sample bias (5 V in height, 10 us in width) (Figure 9.21).

In addition to studies on the polymerization of diacetylenes, the electronic properties of the resultant polydiacetylene nanowires on different substrates have been reported recently [185]. Kelly *et al*. revealed that polydiacetylene (PDA) nanowires exhibit intriguing substrate-dependent electronic effects when probed at varying sample bias voltage conditions on HOPG and molybdenum disulfide (MoS_2). On HOPG surfaces, the PDA nanowires exhibited a decreased tunneling probability as the bias voltage was reduced. The height of the PDA nanowires, when measured at

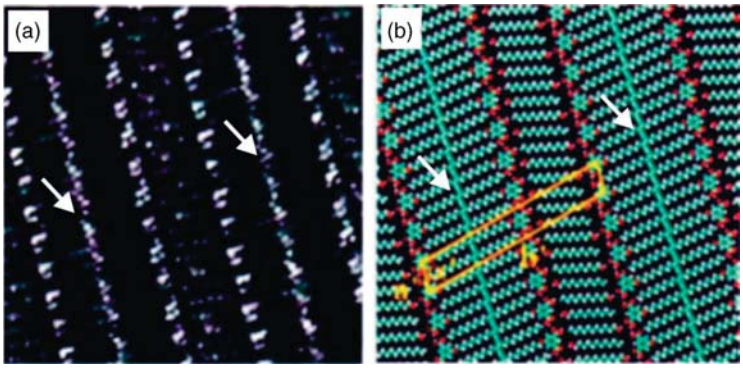


Figure 9.20 STM image of the polymerization of diacetylenes on HOPG. (a) STM image of the polymerization of compound 1. The enhanced tunneling probability in the middle of the molecules marked by white arrows shows the evidence of the polymerized structures; (b) Molecular model of the polymerized structures [179–181].

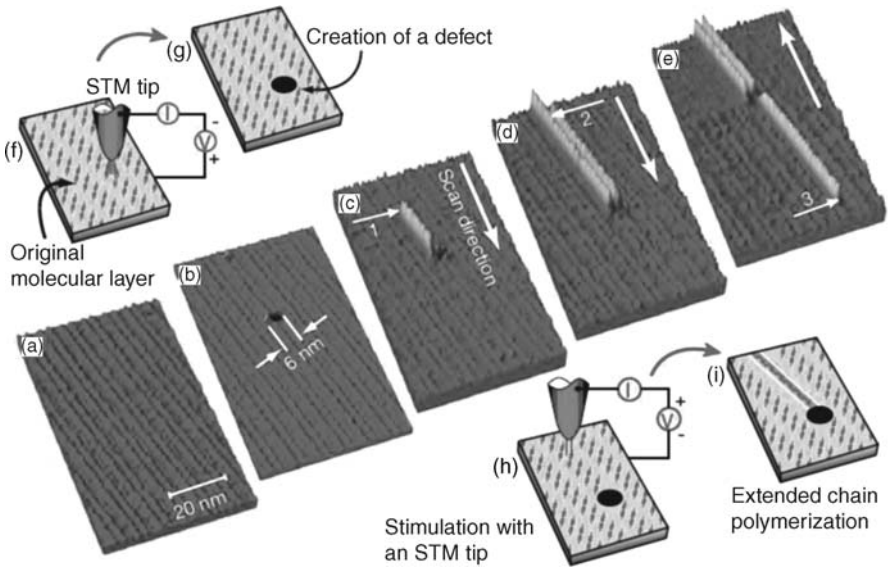


Figure 9.21 STM images and diagrams, showing the controlled polymerization by STM tip. (a) The original SAMs of compound 2p; (b) Artificial defect made by STM tip; (c) First chain polymerization, initiated at arrow (1); (d) Second chain polymerization, initiated at arrow (2); (e) Third chain polymerization, initiated at arrow (3); (f, g) Diagrams showing the creation of an artificial defect; (h, i) Diagrams showing initiation of chain polymerization with an STM tip, and termination of the polymerization at the artificial defect [185].

negative voltages, was substantially higher than that measured at positive voltages. On MoS_2, the PDA nanowires appeared with a much higher contrast on HOPG when imaged under the same negative bias conditions, but could not be visualized under positive bias conditions on MoS_2, despite it being possible still to image the unpolymerized molecules. The authors attributed these phenomena to certain substrate-dependent effects, such as substrate doping, screening, or surface dipole effects

9.4.1.2 **Polymerization by Electrochemical Methods**

Electrochemistry represents a conventional technique for synthesizing polymerized structures on electrodes. A reliance on *in situ* electrochemical scanning tunneling microscopy (EC-STM), and the formation and propagation of linear polymerized structures on single-crystal surfaces, has been revealed at the single-molecule level [186–189]. Yao *et al.* reported the polymerization of an aniline monolayer on a Au(111) electrode in 0.1 M sulfuric acid containing 30 mM aniline [189]. In this case, the ordered aniline monolayer was observed between 0.47 and 0.9 V, versus the reversible hydrogen electrode (RHE). Shifting the potential from 0.9 to 1.05 V led to the polymerization of aniline (see Figure 9.22). Here, the polymerized aniline (PAN) lines were found to propagate preferentially along ⟨112⟩ directions, while the height

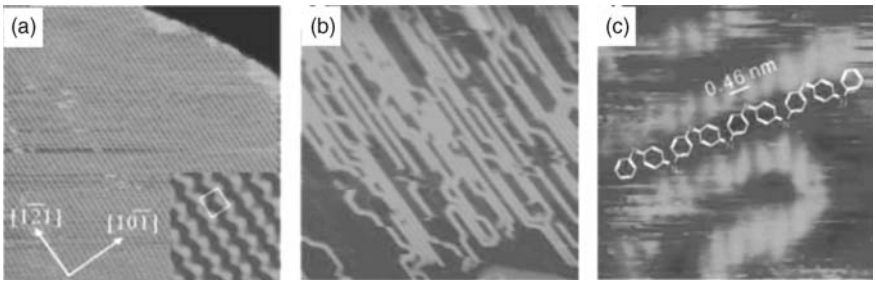


Figure 9.22 Polymerization of aniline on Au(111). (a) *In situ* STM images recorded at 0.9 V; (b) The polymerized aniline structures observed 20 min after the shift of potential from 0.9 to 1.05 V; (c) High-resolution STM image showing the detailed structures of the polymerized aniline [188].

of the polymerized lines was seen to be 0.1–0.4 nm higher than the molecular rows of aniline monomers. According to the high-resolution STM image and the periodic distance of the PAN, the authors proposed that the PAN consisted of aniline molecules linked in a head-to-tail manner (Figure 9.22c).

Other types of electrochemical method have also been reported for producing a single conjugated-polymer wire on a single-crystal electrode. For example, Sakaguchi *et al.* reported the technique of "electrochemical epitaxial polymerization," by which they observed the nucleation and propagation of high-density arrays of single conjugated-polymer wires as long as 75 nm on Au(111) surfaces [188]. In these studies, thiophene derivatives, including 3-butoxy-4-methylthiophene (BuOMT), 3-octylthiophene, 3,3-dibutyl-3,4-dihydro-2*H*-thieno[3,4-b]-[1,4]dioxepine (DBuPDOT) and 3-[(*S*)-2-methylbutoxy]-4-methylthiophene (MBuOMT), were selected to investigate electrochemical polymerization in an iodine-containing electrolyte solution. The electrochemical growth of single-conjugated polymer wires was achieved by applying a given number of positive-voltage pulses (1.4 V, 150 ms, versus Pt) to the Au (111) surfaces. Under the oxidation potential of 1.4 V, the monomer of thiophene derivatives was oxidized to the cation radical, which was the reaction source for the propagation of conjugated polymer. Both, the length and density of single-polythiophene wires were found to depend on the number of pulses of the applied voltage. The application of 15 pulses led to the Au(111) surfaces being almost fully covered by the conjugated polymer, while the iodine-covered Au(111) surfaces acted as a form of molecular template to guide the propagation of the conjugated polymer, as the polythiophene wires appeared along three specific directions (Figure 9.23).

The same method was also reported available for preparing heterojunctions of conjugated copolymers on iodine-covered Au(111) surfaces [187]. For this, two types of thiophene monomer – 3-octyloxy-4-methylthiophene (C8OMT) and 3-octyl-4-methylthiophene (C8MT) – were used as building blocks to create heterowires. Cyclic voltammography showed the oxidation potential of C8OMT to be about 0.4 V lower than that of C8MT, which makes available a multistep electrochemical epitaxial polymerization (ECEP) to prepare the heterowires of conjugated polymers. Several linkage types, including diblock, triblock, and multiblock have been observed using

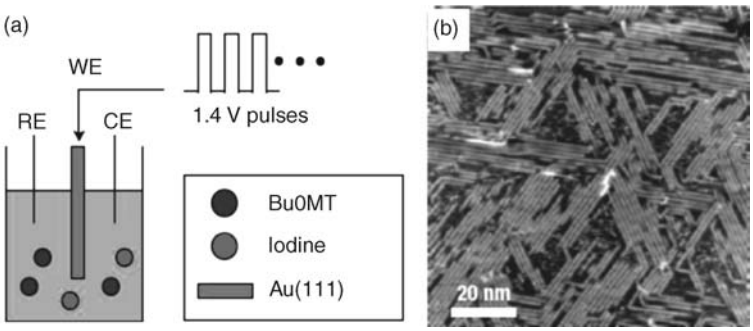


Figure 9.23 (a) Experimental set-up of electrochemical epitaxial polymerization; (b) STM image of the resulting structures after applying 15 voltage pulses (1.4 V, 150 ms) in the BuOMT (10 mM) iodine (0.1 mM)NBu_4PF_6(0.1 M) DCM solution [187].

STM. Subsequent scanning tunneling spectroscopy (STS) studies of the heterowires showed the HOMO–LUMO gaps of each polymer to be the same as those of the wires of each homopolymer.

9.4.1.3 **Polymerization by Thermal Activation**

Thermally initiated polymerization has been used traditionally for the industrial synthesis of polymers. With thermal treatment, reactive radicals can be derived either from the monomer itself or from a small amount of additives, such as organic peroxides. During recent years, thermally initiated polymerization has been applied to the synthesis of covalent connected networks on single-crystal surfaces under UHV conditions [162, 190–192], the aim being to create stable molecular networks with a controlled shape and an efficient electron transport. The critical step in thermally initiated polymerization is the generation of sufficient radicals, and for this two types of method have been reported by Grill and coworkers: (i) where the radicals are produced on the substrate when the sample is heated; or (ii) where they are produced directly in the evaporator [191]. Both methods have been used successfully for the polymerization of molecules containing carbon–halogen bonds.

As shown in Figure 9.24, one type of porphyrin derivative, tetra (4-bromophenyl) porphyrin (Br_4TPP), was selected because: (i) the central part of the molecule is chemically stable, and it is easy to form ordered structures on metal surfaces; and (ii) the radicals may be derived by breaking chemical bonds between bromine and the phenyl group, in a controlled manner. When the evaporator temperature was below 550 K, normal close-packed structures of intact Br_4TPP molecules were found as a result of self-assembly, although covalent connected networks were found on the Au (111) surfaces when the evaporator temperature was higher than 590 K. The pattern of covalent connected structures may be controlled by adjusting the position of the Br substituent. Three types of structure – dimers, chains, and networks – were identified on the surfaces; these corresponded to the self-assembly of one Br substituent, to two *trans*-Br substituents, and to four Br substituents.

Another successful polymerization reaction was reported by Lipton-Duffin and coworkers, who selected 1,4-diiodobenzene and 1,3-diiodobenzene as building

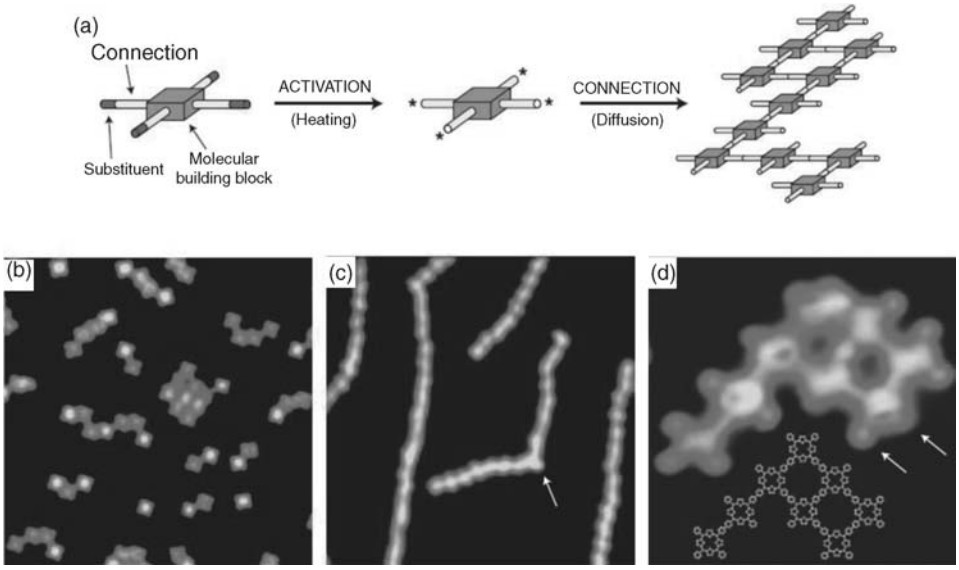


Figure 9.24 (a) Diagram of the polymerization by activated building blocks; (b–d) STM images exhibiting the polymerized structures of dimers, chains and networks, corresponding to the self-assembly of one Br substituent (b), two trans-Br substituents (c), and four Br substituents (d) [192].

blocks [192]. Polymerization reactions were performed on Cu(110) surfaces, as copper may catalyze breaking of the C−I bond. Compared to the bonding strength of the C−Br bond, the weaker C−I bond was easier to break and this allowed the reaction to occur at a relatively lower temperature (ca. 500 K). Subsequent STM studies demonstrated the presence of two types of polymer chain, namely straight and zigzag, which corresponded to the polymerization of 1,4-diiodobenzene and 1,3-diiodobenzene, respectively.

9.4.2
Host–Guest Systems

The host–guest phenomenon is at the origin of supramolecular chemistry. Guest entities (ions, molecules) may be selectively recognized and accommodated by host systems through noncovalent interactions. Recently, these concepts have been introduced for the fabrication of 2-D nanostructures on single-crystal surfaces [72, 193–199], targeting the building of complex ordered structures with nanometer precision. The guest entities, which range from tiny metallic cations [195] to complex organic-molecule-protected metal clusters [200, 201], have been applied to build host–guest structures on the prefabricated molecular template. An overview of recent efforts related to the preparation of molecular templates will be provided in the following subsections, with emphasis placed on the inclusion of guest entities when designing the template.

9.4.2.1 Molecular Template with Porous Networks

A molecular template with porous networks is widely used when studying host–guest systems, as the periodic pores may provide sufficient space for the inclusion of guest entities. The sizes and properties of the pores may be adjusted by changing the molecular structures of the building blocks. Until now, three variant methods have been described for building porous networks on single-crystal surfaces:

- Planar organic molecules with potential to form lateral hydrogen bonds are selected to build hydrogen-bond-connected networks. The pore size may be adjusted by changing the length of the building molecule.
- Under UHV conditions, metal–organic coordination networks may be built by the sequential deposition of organic ligands and metal atoms on preheated substrates.
- At liquid/solid interfaces, using specially synthesized building blocks to self-assemble the porous networks. In this case, the building block consists of a rigid core to ensure the shape of the pore, and lateral length-tunable alkyl chains to adjust the size of the pore.

Hydrogen-Bond-Connected Networks Hydrogen bonding interactions between molecules may cause a remarkable increase in molecule–molecule interactions, increasing the stability of the SAMs. By relying on the selectivity and directionality of the H-bonds, many examples of H-bond-directed molecular templates have been reported, including 1-D lines [35, 50] and 2-D porous networks [49, 193, 202]. As with the porous networks, hexagonal structures are commonly found, and the pore size can be adjusted by changing the building molecules.

The self-assembly of trimesic acid (TMA, **3a** in Figure 9.9) is one the earliest reported examples of forming H-bond-directed 2-D porous networks on single-crystal surfaces [46, 49]. The dimeric H-bonding of the TMA molecules may result in the formation of cavities of diameter about 1.7 nm, which is large enough to accommodate guest molecules such as trimesic acid itself [49], coronene [203], and C_{60} [96]. On the basis of hydrogen bond formation between carboxylic groups, porous networks with larger cavity diameters have been observed by the self-assembly of tetracarboxylic acids (**4e** and **4f** in Figure 9.9) [106] and a tetra-acidic azobenzene molecule (**4b** in Figure 9.9) [104]. However, with increasing molecular length, these larger hexagonal porous networks tend to collapse. As with the self-assembly of tetracarboxylic acids, the molecules prefer to form close-packed structures rather than porous networks on HOPG surfaces [106]. Interestingly, the participation of a guest molecule of coronene significantly enhances the stability of porous networks over the close-packed parallel structures, though this might be attributed to a form of guest molecule-guided template. However, it also shows that the hydrogen bonds between the carboxylic acid are not strong enough to ensure the formation of porous networks. The phase transition, such as from oblique to hexagonal, was also observed in the SAMs of 1,3,5-benzenetribenzoic acid (BTB, **3b** in Figure 9.9) in different solvents at the liquid/substrate interface [116]. This solvent-dependent phase transition implicates the versatility of the SAMs directed by carboxylic hydrogen bonds.

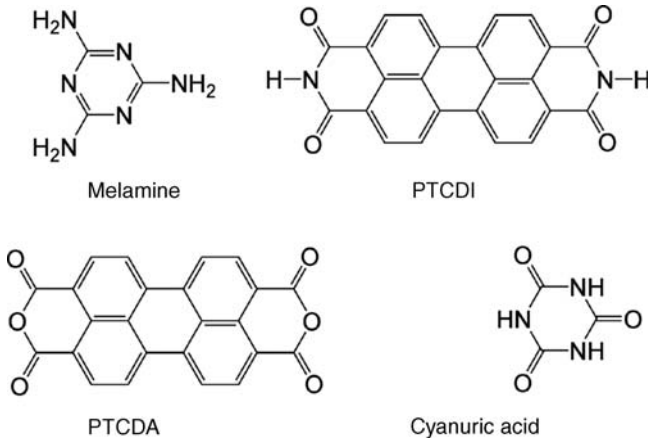


Figure 9.25 Chemical structures of compounds used for building porous networks directed by hydrogen bonds.

To ensure the formation of porous networks, other types of molecule with potential to form stronger hydrogen bonds should be considered. The melamine molecule represents an ideal candidate that may form triple hydrogen bonds with perylene tetracarboxylic diimide (PTCDI) [53, 204–210], perylene tetracarboxylic dianhydride (PTCDA) [211] and cyanuric acid [55, 57, 212], as shown in Figure 9.25. The threefold symmetry of melamine molecule and the strength of the lateral hydrogen bonds (e.g., 15 kcal mol^{-1} between a melamine and a cyanuric acid) [212, 213] ensure the formation of hexagonal porous networks. Robust H-bond-directed porous networks formed by the coadsorption of melamine and PTCDI have been reported both under UHV conditions on silver-covered Si(111) surfaces [204] and on Au(111) surfaces at the liquid/substrate interface [214]. The porous networks formed by melamine and PTCDI were found not only to be capable of accommodating fullerene molecules under UHV conditions [204], but also of patterning the structures of chemisorbed SAMs, such as thiols [214]. Three types of thiol, namely adamantane thiols (ASH), ω-(4′-methylbiphenyl-4-yl) propane thiol (BP3SH) and dodecane thiol (C12SH), were successfully filled in the pores of the networks by forming S–Au bonds. Interestingly, these combined structures, thiols and H-bond-directed networks survived even after the underpotential deposition of copper atoms on the Au substrate, thus demonstrating the adequate stability of the H-bond-directed networks to act as templates in subsequent processes (Figure 9.26).

Porous Networks Directed by Metal–Organic Coordination Bonds The strength of a typical single hydrogen bond is about 5 kcal mol^{-1}, but when multiple hydrogen bonds are formed between self-assembled molecules the molecular interactions will increase correspondingly, giving rise to an enhanced stability of the derived open networks. Compared to metal–organic coordination bonds, however, even multiple hydrogen bonds may be considered only as medium strength interactions, as the typical bond energy of a metal–organic coordination bond amounts to

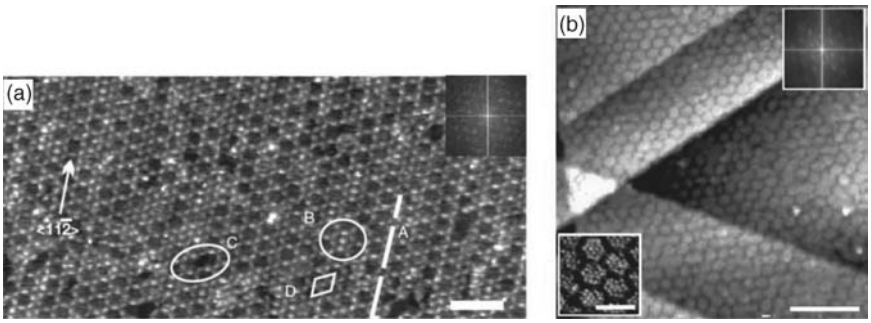


Figure 9.26 (a) STM image showing hexagonal networks formed by melamine and PTCDI; (b) Filled networks by accommodation of guest thiols of ASH [214].

10–30 kcal mol^{-1} per interaction [215]. It may be reasonably anticipated that porous networks directed by metal–organic coordination bonds will possess a greater stability and increased functionality.

The first example of a metal–organic coordination network (MOCN) was reported by the group of Kern, on Cu(100) surfaces under UHV conditions [66]. Here, organic ligands of 1,2,4-benzenetricarboxylic acid (TMLA) and Fe atoms were deposited sequentially onto preheated Cu(100) surfaces (400 K). The structures of the MOCNs could be adjusted by increasing the adsorbate coverage ratio of Fe/TMLA from 1 : 1 to 2 : 1. By using the same method, other organic ligands with carboxylic groups, such as trimesic acid [68], 1,4-dicarboxylic benzoic acid (TPA) and 4,1′,4′,1″-terphenyl-1,4″-dicarboxylic acid (TDA) [72], have been selected to build MOCNs, exhibiting features of chirality and an ability to accommodate guest molecules such as C_{60} [72]. The MOCNs prepared by this method proved to be stable at temperatures up to 500 K under vacuum conditions, which allowed for annealing experiments to investigate the binding strength of the guest molecules. The reversible inclusion of guest molecules, such as cystine, L,L-diphenylalanine (Phe-Phe) and fullerene C_{60}, has been examined by using cavities of MOCNs formed by TMA and Fe as receptors [216] (Figure 9.27).

Those MOCNs formed by the chelation of Fe and carboxylic group often possess twofold or fourfold symmetry. Hexagonal coordination networks may be prepared either from iron centers with linear 4,4′-biphenol ligands, or from cobalt centers with linear 1,4′;4′,1″-terphenyl-4,4″-dicarbonitrile ligands on Cu(100) or Ag(111) surfaces [67]. The fact that the symmetry of the hexagonal structures is independent of the symmetry of the substrate indicates that the strong molecule–molecule interactions predominate over the substrate influences, while the size of the hexagonal pore can be tuned simply by lengthening the size of the ligands. A series of organic linkers (abbreviated NC-Ph$_n$-CN, where n may be three, four, or five) has been synthesized and used to prepare hexagonal networks with Co atoms on Ag(111) surfaces [217]. In this case, hexagonal porous structures were observed with a tunable pore size that ranged from 10 nm^2 ($n = 3$) to 20 nm^2 ($n = 5$). Very recently, the largest size of hexagonal cavity reported when using this method was 24 nm^2, by the

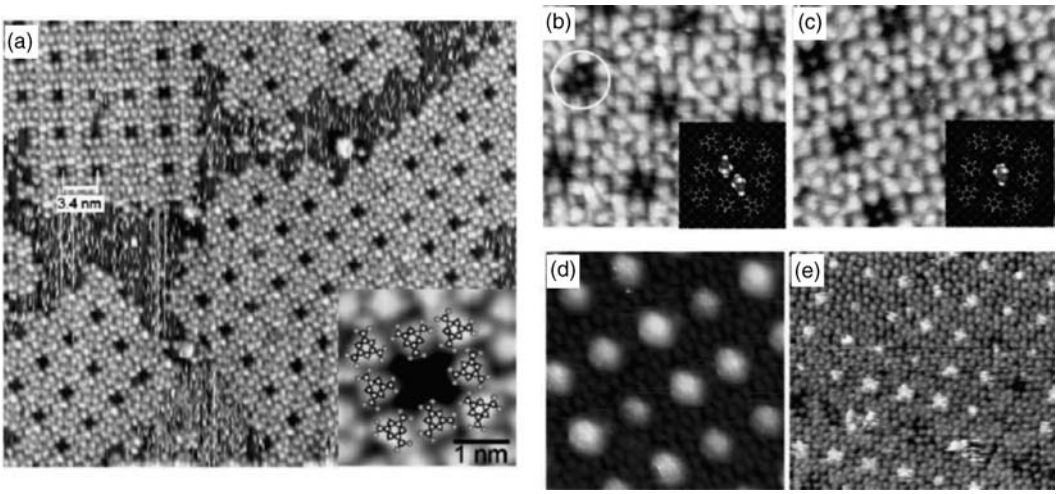


Figure 9.27 (a) Porous networks formed by organic ligands of trimesic acid and metal atoms of Fe; (b) STM image showing two guest cystine molecules anchored in one pore; (c) Upon 430 K annealing, the nanocavities typically accommodate a single cystine guest at the center; (d, e) STM images showing the host–guest system by binding of single C_{60} molecules (d) and Phe-Phe molecules (e) in the cavities [67].

coordination of NC-Ph_6-CN and Co atoms on Ag(111) surfaces [70]. A large area of a single domain was found to cover terraces over μm^2 areas, with a low defect concentration. Under UHV conditions, these hexagonal porous networks have been selected as the template to direct the deposition and the shape of small Fe and Co clusters [218]. These small metal clusters were adsorbed preferentially on top of the organic ligands, for deposition temperatures ranging from 90 to 130 K.

Porous Networks Directed by Van der Waals Forces The porous networks directed by H-bonds and metal–organic coordination bonds usually form rigid structures, due to the properties of molecular interactions. In contrast to rigid porous networks, a type of soft porous network may be prepared at liquid/substrate interfaces by the self-assembly of specially synthesized molecules. Those molecules often have a rigid core to maintain the shape of the networks, and several length-tunable legs (alkyl chains) to adjust the pore size [115, 219–221]. The largest pore prepared in this way was reported to be 7 nm in diameter, and capable of accommodating a giant molecular spoked wheel [222]. In the SAMs of those molecules, the molecular interactions were mainly van der Waals forces. Although the strength of typical van der Waals interactions is less than 1 kcal mol^{-1} between small molecules, adequate molecular interactions to ensure the stability of the SAMs may be obtained by an elongation of the alkyl chains. As the van der Waals interactions are less directional than H-bonds and metal–organic coordination bonds, many factors – such as solvent [115], concentration [114] and guest molecules [223] – may influence the structures of the SAMs.

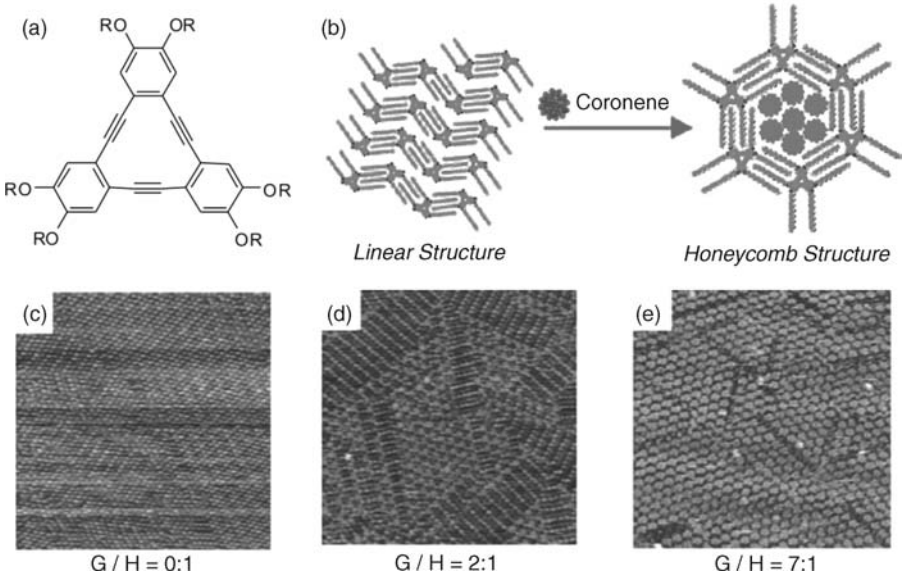


Figure 9.28 (a) Chemical structure of dehydrobenzo[12]annulene (DBA) derivatives; (b) Tentative models of the surface patterns of DBA derivatives with alkyl chain length of C_{14}. Left: linear structure without coronene; right: honeycomb structure capturing at most seven coronene molecules. (c–e) Large STM images of the network structures with or without coronene: (c) guest–host = 0 : 1; (d) guest–host = 2 : 1; (e) guest–host = 7 : 1 [115].

Figure 9.28a shows a molecular building block of dehydrobenzo[12]annulene (DBA) derivatives. The self-assembly of DBA derivatives with alkyl chain length from C_{10} to C_{18} on HOPG surfaces reveals that the structural transformation from honeycomb to a linear structure is related to the length of the alkyl chains [115]. The honeycomb structure predominates in the SAMs of DBA derivatives containing shorter alkyl chains of C_{10}, whereas only the linear structure was found for compounds with chain lengths of C_{14}, C_{16}, and C_{18}. Both, the honeycomb and the linear structure coexist in the SAMs of compounds with alkyl chains of C_{12}. Interestingly, upon the addition of a tenfold excess of guest coronene molecules dissolved in 1,2,4-trichlorobenzene (TCB) to the already-formed linear-type pattern at the liquid/substrate interface, the structures of SAMs formed by compounds with chain lengths of C_{14} were completely converted from the linear structure into the honeycomb structure [223]. Other guest molecules, including hexakis(phenylethynyl) benzene (HPEB), fullerene, 9,10-diphenylanthracene (DPA), chrysene, hexaiodobenzene (HIB), and phthalocyanine (PC), have been checked to explore the phase transition of the host template. Only those planar guest molecules with large π-conjugated cores, such as HPEB and PC, led to the formation of honeycomb networks, whereas the small π-conjugated molecules such as HIB and chrysene, as well as the nonplanar molecules such as DPA and fullerene, had no influence on the linear structure [223]. On the basis of the template for the DBA derivatives, more complex host–guest systems have been observed by adding guest molecules of

coronene, TMA and isophthalic acid (ISA) into the liquid at the 1-octanoic acid/HOPG interface [224]. One coronene molecule with six surrounded ISA molecules was seen to assemble to a complex guest entity that could be accommodated by the molecular template of DBA derivatives, to form a three-component host–guest system.

Those porous networks directed by van der Waals interactions may be prepared at the liquid/substrate interface, which allows for the exploration of the selectivity of the host template by simply adding guest molecules into the liquid. Porous networks formed by the self-assembly of 1,3,5-tris [(E)-2-(3,5-didecyloxyphenyl)-ethenyl]- benzene at the interface of HOPG and 1-phenyloctane may selectively accommodate guest molecules, such as benzo[*rst*] perylene (BPL), coronene, benzo[*rst*]pentaphene (BPP), hexabenzocoronene(HBC), and pentacene [220]. Due to their very large size, only the pentacene molecules were unable to adsorb into the pores of the template at any concentration up to saturation.

9.4.2.2 SAMs of Functional Molecules

The recognition and accommodation of guest molecules on a molecular template with porous networks depend mainly on the structures of the template in space, while the guest entities are restricted to those molecules which can fit into the pores. Besides those porous molecular template, other types of template have been reported to accommodate guest entities, such as metal ions [195], ionic molecules [225], biomolecules [226], C_{60} [227–231] and HBC [232], relying on the properties of the building blocks.

The self-assembly of alkane and alkane derivatives on single-crystal surfaces have been realized for twenty years. At the liquid/substrate interface, these molecules may form spontaneously ordered lamellar structures, with the width of the lamellae being adjustable simply by altering the length of the carbon chains. The well-ordered lamellar structures may be used as molecular templates when the molecules are modified by the functionalized groups [194], The SAMs of fatty acid, such as $C_{19}H_{39}COOH$, were found capable of acting as molecular templates following the addition of guest urea [226]; the molecules of the latter were adsorbed along the boundary of the lamellae by forming hydrogen bonds with carboxylic groups. Other alkane derivatives, such as those modified by amino acid, have been reported to behave in similar fashion when accommodating guest molecules of urea [226]. Besides the interactions between functional groups of the template and the guest molecules, the interactions between alkyl chains may also be utilized to accommodate any guest entities that possess similar structures [200]. This idea originated from the biological proposal that large proteins, when supplied with a long alkyl chain, may be incorporated into the lipid bilayer, during which process the alkyl chain acts as a type of anchor. Based on this strategy, Au_{55} clusters decorated by $C_{18}H_{37}SH$ were prepared as guest entities, whereby a droplet of a mixture containing $C_{14}H_{29}COOH$ and decorated Au_{55} clusters was used to investigate the packing of the Au_{55} cluster on the HOPG substrate. The molecular template of $C_{14}H_{29}COOH$ guided the packing of Au_{55} clusters, leading to strands of linearly packed nanostructures (Figure 9.29) [200].

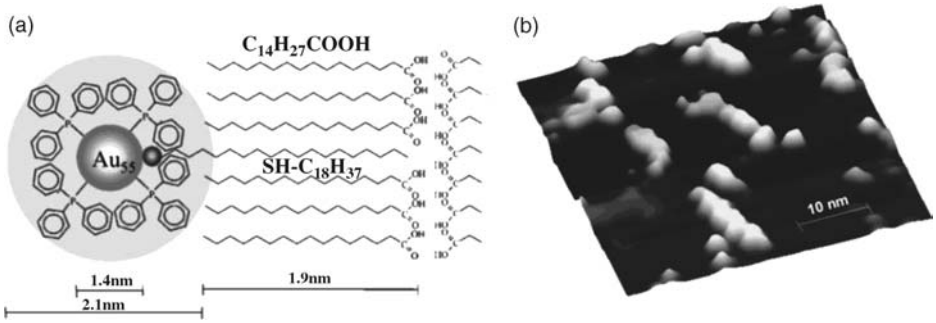


Figure 9.29 (a) Scheme of the exchange mechanism to couple the cluster core with an anchor molecule for the incorporation into a template of tetradecanoic acid; (b) STM image showing the strands of Au_{55} clusters co-adsorbed between $C_{14}H_{29}COOH$ rows [200].

A host template capable of recognizing and accommodating metal ions may be prepared by the self-assembly of functionalized molecules containing a crown ether, such as 15-crown-5-ether-substituted cobalt(II) phthalocyanine (CoCRPc). It has been found that the functional group of 15-crown-5-ether is able to accommodate guest ions such as K^+ and Ca^{2+} in solution. However, the SAMs of CoCRPc prepared on Au(111) and Au(100) surfaces showed different behaviors in the capture of Ca^{2+} ions. In the presence of Ca^{2+}, the SAMs of CoCRPc on Au(111) may capture Ca^{2+} ions in two diagonally located 15-crown-5-ether moieties, whereas the SAMs on Au(100)-(1×1) were unable to capture Ca^{2+} ions. These results suggested that the relationship between the crown moieties and the underlying Au lattice plays an important role in the accommodation of guest ions [195].

Molecules such as 15-crown-5-ether may be considered as a form of "molecular container" the SAMs of which, by relying on their chemical properties, can be used to capture guest entities to form complex host–guest systems. The bowl-shaped calyx [8] arene derivative (Figure 9.30a), OBOCMC8 ($C_{104}H_{128}O_{24}$), is one such type of

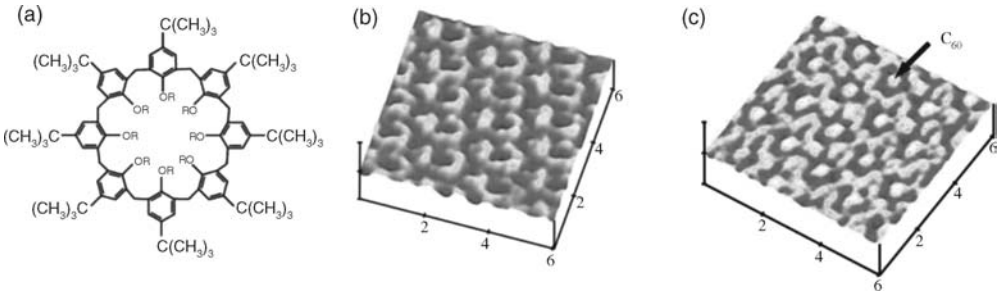


Figure 9.30 (a) Chemical structures of calyx[8] arene derivative (OBOCMC8); (b) STM image showing the SAMs of OBOCMC8. Well-ordered dark depressions are observed which represent the cavities of the OBOCMC8; (c) STM image showing the SAMs of OBOCMC8/C_{60}. The cavities of the OBOCMC8 are filled by guest fullerene molecules, showing bright spots in the center of the cavities [227].

molecular container [227], where the ring size is large enough to accommodate large molecules such as fullerene. In an electrochemical environment, the SAMs of OBOCMC8 were observed when the potential of the Au(111) surfaces was held at 0.6 V (versus RHE), as shown in Figure 9.30b. Here, well-ordered dark depressions were observed that may have been associated with the cavities of the OBOCMC8. Under the same preparation and imaging conditions, the SAMs of OBOCMC8/C_{60} showed different patterns from the SAMs of OBOCMC8. Figure 9.30c shows the high-resolution STM image of the SAMs of OBOCMC8/C_{60}, in which the cavities of the OBOCMC8 were filled by guest fullerene molecules that appeared as bright spots in the center of the cavities. However, not all molecular containers preferentially capture guest entities in the position of the cavities. For example, the molecular template of fully conjugated cyclo[12]thiophene (C[12]T) demonstrated an ability to capture fullerene molecules preferentially on the rim of the C[12]T [233], as a result of the attractive donor–acceptor interactions. Although the guest fullerene molecules might occasionally be captured in the cavities, the interaction between the fullerene and the host C[12]T was found to be very weak, and the fullerene molecules captured in the cavities were easily desorbed during scanning.

9.4.2.3 **Two-Dimensional Chiral Template**

In chemistry, chirality is a concept involving molecular structures, and describes a molecule that cannot be superposed on its mirror image. It is only very recently that the concept of chirality has been introduced to the 2-D building of chiral templates on surfaces. Various types of molecule, including chiral molecules, prochiral molecules and even achiral molecules, have been reported to be successful in the building of 2-D chiral patterns on single-crystal surfaces. The building of chiral templates, and their selective inclusion of guest molecules, will be introduced briefly in the following subsection. More detailed discussions on surface chirality are available in other reviews [234–237].

On occasion, the self-assembly of chiral molecules may cause chirality to be transferred directly from the molecule to SAMs. The forces required to direct such chiral self-assembly include hydrogen-bonding interactions, such as the self-assembly of L(D)-cysteine on Au(110) [238–240], and van der Waals forces, an example being the SAMs of (M)-or (P)-[7]-Helicene on Cu(111) [237, 241, 242]. In differing from chiral molecules, the self-assembly of prochiral molecules can create enantiomers on surfaces, but only when the adsorption of a molecule breaks its symmetry element [45, 50, 243, 244]. The aggregation of pure enantiomers gives rise to the formation of homochiral domains. Besides the self-assembly of chiral and prochiral molecules, chiral packing structures may also be observed, where chiral adsorption geometry is introduced by the adsorbate–substrate interactions, such as the chiral structures observed in the SAMs of normal alkanes [235] and star-shaped molecules [245], although the adsorption of achiral molecules never results in the creation of enantiomers. Achiral molecules can also form chiral structures when several molecules form rotating structures under the directional interactions of hydrogen bonds or metal–organic coordination bonds [63, 67, 207, 212, 246, 247].

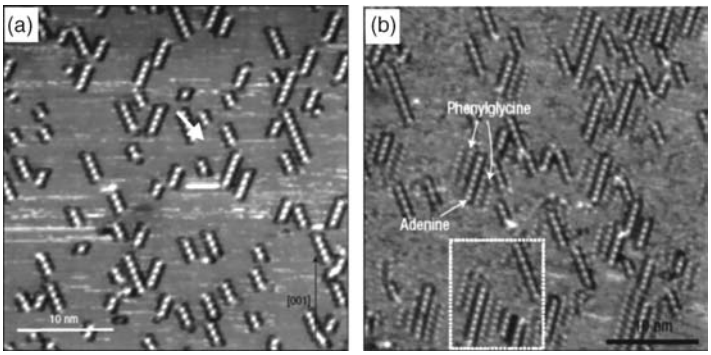


Figure 9.31 Selective adsorptions of guest molecules on basis of chiral structures. (a) STM image of Cu(110) surface with submonolayer coverage of adenine, showing chiral packing molecular chains aligned along [±1,2] directions; (b) STM image of the selective attachment of guest *S*-phenylglycine which were found to adsorb preferentially near the chains aligned along the [1,2] direction [248].

Although many 2-D chiral templates have been realized recently, very few reports have been made on the selective inclusion of guest molecules on basis of the prefabricated chiral template [248]. Molecular dimer chains formed by the self-assembly of adenine on Cu(110) exhibit chiral features due to the preferential adsorption of adenine molecules along [±1,2] directions with respect to the Cu (110) surfaces, as shown in Figure 9.31a. These dimer chains show significant chiral selectivity to the inclusion of guest amino acid molecules. In this case, guest molecules of *S*-phenylglycine were found to adsorb near the chains aligned along the [1,2] direction (see Figure 9.31b), whereas *R*-phenylglycine only attached to the chains aligned along the [−1,2] direction. According to the results of DFT calculations, the coulombic repulsion between the phenylglycine amino group and the DNA base was considered to be responsible for the chiral recognition. The substrate-mediated charge transfer played a critical role in chiral selectivity, whereas the direct molecular interactions such as hydrogen bonds did not [249].

9.5 Summary and Outlook

In this chapter, the recent progress on surface-supported nanostructures directed by atomic- and molecular-level templates has been reviewed, with attention focused on those nanostructured systems that are mainly organic-related and involve organic molecules as either templates or as objects, the organization of which is directed by nanostructured templates. The subject of surface-supported nanostructures that can serve as templates for the further fabrication of nanosized objects was introduced, whereby nanostructures – including surface reconstructions and reconstruction-related patterns, strain-relief epitaxial layers, vicinal surfaces, and surface-supported organic supramolecular assemblies – may serve as naturally formed templates on the

atomic and molecular scale. The organization of nanosized objects directed by templates was then discussed. Despite being unable to include the details of many other reported studies, it is hoped that the present review will provide the reader with basic information regarding recent investigations in this field. Further data are available in other reviews on surface-supported supramolecular assembly [250], and the controlled assembly of organic molecules on 2-D nanotemplates [250, 251].

The size scales that can be approached by either self-assembly or so-called "bottom-up" techniques are far beyond the limit of "top-down" lithographic techniques. The former has the advantage of precisely controlling the position and orientation of nanosized objects. Although knowledge regarding self-assembly continues to be unveiled, there remain major challenges for scientists and engineers to blend the findings of the basic research conducted at academic institutions with applications in industry. For device fabrication, the robust nature of self-assembled nanostructures is vital, since molecules in surface-supported supramolecular architectures, which usually are prepared under UHV or at liquid/solid interfaces, are bound only weakly to the surface and to each other. The ability to improve the stability of these nanostructures during fabrication under ambient conditions remains a major problem; whereas, for electronic devices an insulating substrate is required, the substrates used widely today for supramolecular assembly are crystalline metals or semiconductors. Thus, the expansion of a supramolecular assembly onto an insulating or even amorphous substrate represents a major challenge.

References

1 Binnig, G., Rohrer, H., Gerber, C., and Weibel, E. (1983) *Phys. Rev. Lett.*, **50**, 120.
2 Takayanagi, K., Tanishiro, Y., Takahashi, S., and Takahashi, M. (1985) *Surf. Sci.*, **164**, 367.
3 Takayanagi, K., Tanishiro, Y., Takahashi, M., and Takahashi, S. (1985) *J. Vac. Sci. Technol. A.*, **3**, 1502.
4 Vanhove, M.A., Koestner, R.J., Stair, P.C., Biberian, J.P., Kesmodel, L.L., Bartos, I., and Somorjai, G.A. (1981) *Surf. Sci.*, **103**, 189.
5 Barth, J.V., Brune, H., Ertl, G., and Behm, R.J. (1990) *Phys. Rev. B*, **42**, 9307.
6 Kuk, Y., Chua, F.M., Silverman, P.J., and Meyer, J.A. (1990) *Phys. Rev. B*, **41**, 12393.
7 Coulman, D.J., Wintterlin, J., Behm, R.J., and Ertl, G. (1990) *Phys. Rev. Lett.*, **64**, 1761.
8 Otero, R., Naitoh, Y., Rosei, F., Jiang, P., Thostrup, P., Gourdon, A., Laegsgaard, E., Stensgaard, I., Joachim, C., and Besenbacher, F. (2004) *Angew. Chem., Int. Ed.*, **43**, 2092.
9 Kern, K., Niehus, H., Schatz, A., Zeppenfeld, P., George, J., and Comsa, G. (1991) *Phys. Rev. Lett.*, **67**, 855.
10 Auwarter, W., Kreutz, T.J., Greber, T., and Osterwalder, J. (1999) *Surf. Sci.*, **429**, 229.
11 Corso, M., Auwarter, W., Muntwiler, M., Tamai, A., Greber, T., and Osterwalder, J. (2004) *Science*, **303**, 217.
12 Goriachko, A., He, Y.B., Knapp, M., Over, H., Corso, M., Brugger, T., Berner, S., Osterwalder, J., and Greber, T. (2007) *Langmuir*, **23**, 2928.
13 Laskowski, R., Blaha, P., Gallauner, T., and Schwarz, K. (2007) *Phys. Rev. Lett.*, **98**, 106802.
14 Berner, S., Corso, M., Widmer, R., Groening, O., Laskowski, R., Blaha, P., Schwarz, K., Goriachko, A., Over, H., Gsell, S., Schreck, M., Sachdev, H., Greber, T., and Osterwalder, J. (2007) *Angew. Chem., Int. Ed.*, **46**, 5115.
15 Brune, H., Giovannini, M., Bromann, K., and Kern, K. (1998) *Nature*, **394**, 451.
16 Brune, H., Roder, H., Boragno, C., and Kern, K. (1994) *Phys. Rev. B*, **49**, 2997.

17 Brune, H., Bromann, K., Roder, H., Kern, K., Jacobsen, J., Stoltze, P., Jacobsen, K., and Norskov, J. (1995) *Phys. Rev. B*, **52**, 14380.

18 Ait-Mansour, K., Buchsbaum, A., Ruffieux, P., Schmid, M., Groning, P., Varga, P., Fasel, R., and Groning, O. (2008) *Nano Lett.*, **8**, 2035.

19 Pohl, K., Bartelt, M.C., de la Figuera, J., Bartelt, N.C., Hrbek, J., and Hwang, R.Q. (1999) *Nature*, **397**, 238.

20 Thurmer, K., Hwang, R.Q., and Bartelt, N.C. (2006) *Science*, **311**, 1272.

21 Jaloviar, S.G., Lin, J.L., Liu, F., Zielasek, V., McCaughan, L., and Lagally, M.G. (1999) *Phys. Rev. Lett.*, **82**, 791.

22 Kuhnke, K. and Kern, K. (2003) *J. Phys. Condens. Matter*, **15**, S3311.

23 Giesenseibert, M., Schmitz, F., Jentjens, R., and Ibach, H. (1995) *Surf. Sci.*, **329**, 47.

24 Hoogeman, M.S., Klik, M.A.J., Schlosser, D.C., Kuipers, L., and Frenken, J.W.H. (1999) *Phys. Rev. Lett.*, **82**, 1728.

25 Hahn, E., Schief, H., Marsico, V., Fricke, A., and Kern, K. (1994) *Phys. Rev. Lett.*, **72**, 3378.

26 Barbier, L., Masson, L., Cousty, J., and Salanon, B. (1996) *Surf. Sci.*, **345**, 197.

27 Viernow, J., Lin, J.L., Petrovykh, D.Y., Leibsle, F.M., Men, F.K., and Himpsel, F.J. (1998) *Appl. Phys. Lett.*, **72**, 948.

28 Ozcomert, J.S., Pai, W.W., Bartelt, N.C., and Reuttrobey, J.E. (1994) *Phys. Rev. Lett.*, **72**, 258.

29 Xiao, W.D., Ruffieux, P., Ait-Mansour, K., Groning, O., Palotas, K., Hofer, W.A., Groning, P., and Fasel, R. (2006) *J. Phys. Chem. B*, **110**, 21394.

30 Rousset, S., Repain, V., Baudot, G., Garreau, Y., and Lecoeur, J. (2003) *J. Phys. Condens. Matter*, **15**, S3363.

31 Yokoyama, T., Yokoyama, S., Kamikado, T., Okuno, Y., and Mashiko, S. (2001) *Nature*, **413**, 619.

32 Fernandez-Torrente, I., Monturet, S., Franke, K.J., Fraxedas, J., Lorente, N., and Pascual, J.I. (2007) *Phys. Rev. Lett.*, **99**, 176103.

33 Stadler, C., Hansen, S., Kroger, I., Kumpf, C., and Umbach, E. (2009) *Nat. Phys.*, **5**, 153.

34 Yokoyama, T., Takahashi, T., Shinozaki, K., and Okamoto, M. (2007) *Phys. Rev. Lett.*, **98**, 206102.

35 Barth, J.V., Weckesser, J., Cai, C.Z., Gunter, P., Burgi, L., Jeandupeux, O., and Kern, K. (2000) *Angew. Chem., Int. Ed.*, **39**, 1230.

36 Classen, T., Fratesi, G., Costantini, G., Fabris, S., Stadler, F.L., Kim, C., de Gironcoli, S., Baroni, S., and Kern, K. (2005) *Angew. Chem., Int. Edit.*, **44**, 6142.

37 Pawin, G., Solanki, U., Kwon, K.Y., Wong, K.L., Lin, X., Jiao, T., and Bartels, L. (2007) *J. Am. Chem. Soc.*, **129**, 12056.

38 Schiffrin, A., Reichert, J., Auwarter, W., Jahnz, G., Pennec, Y., Weber-Bargioni, A., Stepanyuk, V.S., Niebergall, L., Bruno, P., and Barth, J.V. (2008) *Phys. Rev. B*, **78**, 035424.

39 Schiffrin, A., Riemann, A., Auwarter, W., Pennec, Y., Weber-Bargioni, A., Cvetko, D., Cossaro, A., Alberto, M., and Barth, J.V. (2007) *Proc. Natl Acad. Sci. USA*, **104**, 5279.

40 Schnadt, J., Rauls, E., Xu, W., Vang, R.T., Knudsen, J., Laegsgaard, E., Li, Z., Hammer, B., and Besenbacher, F. (2008) *Phys. Rev. Lett.*, **100**, 046103.

41 Surin, M., Samori, P., Jouaiti, A., Kyritsakas, N., and Hosseini, M.W. (2007) *Angew. Chem., Int. Ed.*, **46**, 245.

42 Klappenberger, F., Canas-Ventura, M.E., Clair, S., Pons, S., Schlickum, U., Qu, Z.R., Brune, H., Kern, K., Strunskus, T., Woll, C., Comisso, A., De Vita, A., Ruben, M., and Barth, J.V. (2007) *ChemPhysChem*, **8**, 1782.

43 Trixler, F., Market, T., Lackinger, M., Jamitzky, F., and Heckl, W.M. (2007) *Chem. Eur. J.*, **13**, 7785.

44 Ciesielski, A., Schaeffer, G., Petitjean, A., Lehn, J.M., and Samori, P. (2009) *Angew. Chem., Int. Ed.*, **48**, 2039.

45 Weckesser, J., De Vita, A., Barth, J.V., Cai, C., and Kern, K. (2001) *Phys. Rev. Lett.*, **87**, 096101.

46 Ye, Y.C., Sun, W., Wang, Y.F., Shao, X., Xu, X.G., Cheng, F., Li, J.L., and Wu, K. (2007) *J. Phys. Chem. C*, **111**, 10138.

47 Ma, Z., Wang, Y.Y., Wang, P., Huang, W., Li, Y.B., Lei, S.B., Yang, Y.L., Fan, X.L., and Wang, C. (2007) *ACS Nano*, **1**, 160.

48 Ishikawa, Y., Ohira, A., Sakata, M., Hirayama, C., and Kunitake, M. (2002) *Chem. Commun.*, 2652.

49 Griessl, S., Lackinger, M., Edelwirth, M., Hietschold, M., and Heckl, W.M. (2002) *Single Molecules*, **3**, 25.

50 Barth, J.V., Weckesser, J., Trimarchi, G., Vladimirova, M., De Vita, A., Cai, C.Z., Brune, H., Gunter, P., and Kern, K. (2002) *J. Am. Chem. Soc.*, **124**, 7991.

51 Ruben, M., Payer, D., Landa, A., Comisso, A., Gattinoni, C., Lin, N., Collin, J.P., Sauvage, J.P., De Vita, A., and Kern, K. (2006) *J. Am. Chem. Soc.*, **128**, 15644.

52 Otero, R., Schock, M., Molina, L.M., Laegsgaard, E., Stensgaard, I., Hammer, B., and Besenbacher, F. (2005) *Angew. Chem., Int. Ed.*, **44**, 2270.

53 Silly, F., Shaw, A.Q., Porfyrakis, K., Briggs, G.A.D., and Castell, M.R. (2007) *Appl. Phys. Lett.*, **91**, 253109.

54 Silly, F., Shaw, A.Q., Briggs, G.A.D., and Castell, M.R. (2008) *Appl. Phys. Lett.*, **92**, 023102.

55 Perdigao, L.M.A., Champness, N.R., and Beton, P.H. (2006) *Chem. Commun.*, 538.

56 Llanes-Pallas, A., Palma, C.A., Piot, L., Belbakra, A., Listorti, A., Prato, M., Samori, P., Armaroli, N., and Bonifazi, D. (2009) *J. Am. Chem. Soc.*, **131**, 509.

57 Staniec, P.A., Perdigao, L.M.A., Rogers, B.L., Champness, N.R., and Beton, P.H. (2007) *J. Phys. Chem. C*, **111**, 886.

58 Gesquiere, A., Jonkheijm, P., Hoeben, F.J.M., Schenning, A., De Feyter, S., De Schryver, F.C., and Meijer, E.W. (2004) *Nano Lett.*, **4**, 1175.

59 De Feyter, S., Gesquiere, A., Klapper, M., Mullen, K., and De Schryver, F.C. (2003) *Nano Lett.*, **3**, 1485.

60 Stohr, M., Wahl, M., Galka, C.H., Riehm, T., Jung, T.A., and Gade, L.H. (2005) *Angew. Chem., Int. Ed.*, **44**, 7394.

61 Silly, F., Shaw, A.Q., Castell, M.R., Briggs, G.A.D., Mura, M., Martsinovich, N., and Kantorovich, L. (2008) *J. Phys. Chem. C*, **112**, 11476.

62 Pawin, G., Wong, K.L., Kwon, K.Y., and Bartels, L. (2006) *Science*, **313**, 961.

63 Stepanow, S., Lin, N., Vidal, F., Landa, A., Ruben, M., Barth, J.V., and Kern, K. (2005) *Nano Lett.*, **5**, 901.

64 Lin, N., Stepanow, S., Ruben, M., and Barth, J.V. (2009) *Top. Curr. Chem.*, **287**, 1.

65 Stepanow, S., Lin, N., and Barth, J.V. (2008) *J. Phys. Condens. Matter*, **20**, 184002.

66 Dmitriev, A., Spillmann, H., Lin, N., Barth, J.V., and Kern, K. (2003) *Angew. Chem., Int. Ed.*, **42**, 2670.

67 Stepanow, S., Lin, N., Payer, D., Schlickum, U., Klappenberger, F., Zoppellaro, G., Ruben, M., Brune, H., Barth, J.V., and Kern, K. (2007) *Angew. Chem., Int. Ed.*, **46**, 710.

68 Spillmann, H., Dmitriev, A., Lin, N., Messina, P., Barth, J.V., and Kern, K. (2003) *J. Am. Chem. Soc.*, **125**, 10725.

69 Tait, S.L., Wang, Y., Costantini, G., Lin, N., Baraldi, A., Esch, F., Petaccia, L., Lizzit, S., and Kern, K. (2008) *J. Am. Chem. Soc.*, **130**, 2108.

70 Kuhne, D., Klappenberger, F., Decker, R., Schlickum, U., Brune, H., Klyatskaya, S., Ruben, M., and Barth, J.V. (2009) *J. Am. Chem. Soc.*, **131**, 3881.

71 Schlickum, U., Decker, R., Klappenberger, F., Zoppellaro, G., Klyatskaya, S., Ruben, M., Silanes, I., Arnau, A., Kern, K., Brune, H., and Barth, J.V. (2007) *Nano Lett.*, **7**, 3813.

72 Stepanow, S., Lingenfelder, M., Dmitriev, A., Spillmann, H., Delvigne, E., Lin, N., Deng, X.B., Cai, C.Z., Barth, J.V., and Kern, K. (2004) *Nat. Mater.*, **3**, 229.

73 Langner, A., Tait, S.L., Lin, N., Rajadurai, C., Ruben, M., and Kern, K. (2007) *Proc. Natl Acad. Sci. USA*, **104**, 17927.

74 Zhang, Y.F., Zhu, N., and Komeda, T. (2008) *Surf. Sci.*, **602**, 614.

75 Zhang, Y.F., Zhu, N., and Komeda, T. (2007) *J. Phys. Chem. C*, **111**, 16946.

76 Lin, N., Dmitriev, A., Weckesser, J., Barth, J.V., and Kern, K. (2002) *Angew. Chem., Int. Ed.*, **41**, 4779.

77 Tait, S.L., Langner, A., Lin, N., Stepanow, S., Rajadurai, C., Ruben, M., and Kern, K. (2007) *J. Phys. Chem. C*, **111**, 10982.

78 Pawin, G., Wong, K.L., Kim, D., Sun, D.Z., Bartels, L., Hong, S., Rahman, T.S., Carp, R., and Marsella, M. (2008) *Angew. Chem., Int. Ed.*, **47**, 8442.

79 Zhang, H.M., Zhao, W., Xie, Z.X., Long, L.S., Mao, B.W., Xu, X., and Zheng, L.S. (2007) *J. Phys. Chem. C*, **111**, 7570.

80 Nath, K.G., Ivasenko, O., MacLeod, J.M., Miwa, J.A., Wuest, J.D., Nanci, A., Perepichka, D.F., and Rosei, F. (2007) *J. Phys. Chem. C*, **111**, 16996.

81 Lennartz, M.C., Atodiresei, N., Muller-Meskamp, L., Karthauser, S., Waser, R., and Blugel, S. (2009) *Langmuir*, **25**, 856.

82 Guo, Q., Cocks, I., and Williams, E.M. (1997) *Surf. Sci.*, **393**, 1.

83 Frederick, B.G., Chen, Q., Leibsle, F.M., Lee, M.B., Kitching, K.J., and Richardson, N.V. (1997) *Surf. Sci.*, **394**, 1.

84 Dougherty, D.B., Maksymovych, P., and Yates, J.T. (2006) *Surf. Sci.*, **600**, 4484.

85 Chen, Q., Perry, C.C., Frederick, B.G., Murray, P.W., Haq, S., and Richardson, N.V. (2000) *Surf. Sci.*, **446**, 63.

86 Li, H., Xu, B., Evans, D., and Reutt-Robey, J.E. (2007) *J. Phys. Chem. C*, **111**, 2102.

87 Lin, N., Stepanow, S., Vidal, F., Barth, J.V., and Kern, K. (2005) *Chem. Commun.*, 1681.

88 Langner, A., Tait, S.L., Lin, N., Chandrasekar, R., Ruben, M., and Kern, K. (2008) *Angew. Chem., Int. Ed.*, **47**, 8835.

89 Yang, Y.L., Deng, K., Zeng, Q.D., and Wang, C. (2006) *Surf. Interface Anal.*, **38**, 1039.

90 Lackinger, M., Griessl, S., Markert, T., Jamitzky, F., and Heckl, W.M. (2004) *J. Phys. Chem. B*, **108**, 13652.

91 Clair, S., Pons, S., Seitsonen, A.P., Brune, H., Kern, K., and Barth, J.V. (2004) *J. Phys. Chem. B*, **108**, 14585.

92 Kampschulte, L., Griessl, S., Heckl, W.M., and Lackinger, M. (2005) *J. Phys. Chem. B*, **109**, 14074.

93 Lackinger, M., Griessl, S., Kampschulte, L., Jamitzky, F., and Heckl, W.M. (2005) *Small*, **1**, 532.

94 Sheerin, G. and Cafolla, A.A. (2005) *Surf. Sci.*, **577**, 211.

95 Zhu, N., Osada, T., and Komeda, T. (2007) *Surf. Sci.*, **601**, 1789.

96 Griessl, S.J.H., Lackinger, M., Jamitzky, F., Markert, T., Hietschold, M., and Heckl, W.M. (2004) *J. Phys. Chem. B*, **108**, 11556.

97 Li, Z., Han, B., Wan, L.J., and Wandlowski, T. (2005) *Langmuir*, **21**, 6915.

98 Nath, K.G., Ivasenko, O., Miwa, J.A., Dang, H., Wuest, J.D., Nanci, A., Perepichka, D.F., and Rosei, F. (2006) *J. Am. Chem. Soc.*, **128**, 4212.

99 Kampschulte, L., Werblowsky, T.L., Kishore, R.S.K., Schmittel, M., Heckl, W.M., and Lackinger, M. (2008) *J. Am. Chem. Soc.*, **130**, 8502.

100 Su, G.J., Zhang, H.M., Wan, L.J., Bai, C.L., and Wandlowski, T. (2004) *J. Phys. Chem. B*, **108**, 1931.

101 Dmitriev, A., Lin, N., Weckesser, J., Barth, J.V., and Kern, K. (2002) *J. Phys. Chem. B*, **106**, 6907.

102 Gutzler, R., Lappe, S., Mahata, K., Schmittel, M., Heckl, W.M., and Lackinger, M. (2009) *Chem. Commun.*, 680.

103 Zhou, H., Dang, H., Yi, J.H., Nanci, A., Rochefort, A., and Wuest, J.D. (2007) *J. Am. Chem. Soc.*, **129**, 13774.

104 Li, M., Deng, K., Lei, S.B., Yang, Y.L., Wang, T.S., Shen, Y.T., Wang, C.R., Zeng, Q.D., and Wang, C. (2008) *Angew. Chem., Int. Ed.*, **47**, 6717.

105 Blunt, M.O., Russell, J.C., Gimenez-Lopez, M.D., Garrahan, J.P., Lin, X., Schroder, M., Champness, N.R., and Beton, P.H. (2008) *Science*, **322**, 1077.

106 Blunt, M., Lin, X., Gimenez-Lopez, M.D., Schroder, M., Champness, N.R., and Beton, P.H. (2008) *Chem. Commun.*, 2304.

107 Pivetta, M., Blum, M.C., Patthey, F., and Schneider, W.D. (2008) *Angew. Chem., Int. Ed.*, **47**, 1076.

108 Xu, W., Dong, M.D., Gersen, H., Rauls, E., Vazquez-Campos, S., Crego-Calama, M., Reinhoudt, D.N., Laegsgaard, E., Stensgaard, I., Linderoth, T.R., and Besenbacher, F. (2008) *Small*, **4**, 1620.

109 Schull, G., and Berndt, R. (2007) *Phys. Rev. Lett.*, **99**, 226105.

110 Lin, F., Zhong, D.Y., Chi, L.F., Ye, K., Wang, Y., and Fuchs, H. (2006) *Phys. Rev. B*, **73**, 235420.

111 Kong, X.H., Deng, K., Yang, Y.L., Zeng, Q.D., and Wang, C. (2007) *J. Phys. Chem. C*, **111**, 9235.

112 Stepanow, S., Strunskus, T., Lingenfelder, M., Dmitriev, A., Spillmann, H., Lin, N., Barth, J.V., Woll, C., and Kern, K. (2004) *J. Phys. Chem. B*, **108**, 19392.

113 Zhong, D.Y., Lin, F., Chi, L.F., Wang, Y., and Fuchs, H. (2005) *Phys. Rev. B*, **71**, 125336.

114 Lei, S.B., Tahara, K., De Schryver, F.C., Van der Auweraer, M., Tobe, Y., and De Feyter, S. (2008) *Angew. Chem., Int. Ed.*, **47**, 2964.

115 Tahara, K., Furukawa, S., Uji-i, H., Uchino, T., Ichikawa, T., Zhang, J., Mamdouh, W., Sonoda, M., De Schryver, F.C., De Feyter, S., and Tobe, Y. (2006) *J. Am. Chem. Soc.*, **128**, 16613.

116 Kampschulte, L., Lackinger, M., Maier, A.K., Kishore, R.S.K., Griessl, S., Schmittel, M., and Heckl, W.M. (2006) *J. Phys. Chem. B*, **110**, 10829.

117 Lackinger, M., Griessl, S., Heckl, W.A., Hietschold, M., and Flynn, G.W. (2005) *Langmuir*, **21**, 4984.

118 Mamdouh, W., Uji-i, H., Ladislaw, J.S., Dulcey, A.E., Percec, V., De Schryver, F.C., and De Feyter, S. (2006) *J. Am. Chem. Soc.*, **128**, 317.

119 Yang, Y.L. and Wang, C. (2009) *Curr. Opin. Colloid Interface Sci.*, **14**, 135.

120 Florio, G.M., Ilan, B., Muller, T., Baker, T.A., Rothman, A., Werblowsky, T.L., Berne, B.J., and Flynn, G.W. (2009) *J. Phys. Chem. C*, **113**, 3631.

121 Palma, C.A., Bonini, M., Breiner, T., and Samori, P. (2009) *Adv. Mater.*, **21**, 1383.

122 Palma, C.A., Bonini, M., Llanes-Pallas, A., Breiner, T., Prato, M., Bonifazi, D., and Samori, P. (2008) *Chem. Commun.*, 5289.

123 Kudernac, T., Lei, S.B., Elemans, J., and De Feyter, S. (2009) *Chem. Soc. Rev.*, **38**, 402.

124 Chambliss, D.D., Wilson, R.J., and Chiang, S. (1991) *Phys. Rev. Lett.*, **66**, 1721.

125 Clair, S., Pons, S., Fabris, S., Baroni, S., Brune, H., Kern, K., and Barth, J.V. (2006) *J. Phys. Chem. B*, **110**, 5627.

126 Clair, S., Pons, W., Brune, H., Kern, K., and Barth, J.V. (2005) *Angew. Chem., Int. Ed.*, **44**, 7294.

127 Jensen, S. and Baddeley, C.J. (2008) *J. Phys. Chem. C*, **112**, 15439.

128 Mendez, J., Caillard, R., Otero, G., Nicoara, N., and Martin-Gago, J.A. (2006) *Adv. Mater.*, **18**, 2048.

129 Trant, A.G., Jones, T.E., and Baddeley, C.J. (2007) *J. Phys. Chem. C*, **111**, 10534.

130 Gao, L., Liu, Q., Zhang, Y.Y., Jiang, N., Zhang, H.G., Cheng, Z.H., Qiu, W.F., Du, S.X., Liu, Y.Q., Hofer, W.A., and Gao, H.J. (2008) *Phys. Rev. Lett.*, **101**, 197209.

131 Wang, Y.F., Ge, X., Schull, G., Berndt, R., Bornholdt, C., Koehler, F., and Herges, R. (2008) *J. Am. Chem. Soc.*, **130**, 4218.

132 Zhong, D.Y., Blömker, T., Wedeking, K., Chi, L.F., Erker, G., and Fuchs, H. (2009) *Nano Lett.*, **9**, 4387.

133 Ecija, D., Otero, R., Sanchez, L., Gallego, J.M., Wang, Y., Alcami, M., Martin, F., Martin, N., and Miranda, R. (2007) *Angew. Chem., Int. Ed.*, **46**, 7874.

134 Cicoira, F., Miwa, J.A., Melucci, M., Barbarella, G., and Rosei, F. (2006) *Small*, **2**, 1366.

135 Cicoira, F., Miwa, J.A., Perepichka, D.F., and Rosei, F. (2007) *J. Phys. Chem. A*, **111**, 12674.

136 Oehzelt, M., Grill, L., Berkebile, S., Koller, G., Netzer, F.P., and Ramsey, M.G. (2007) *ChemPhysChem*, **8**, 1707.

137 Li, J.L., Jia, J.F., Liang, X.J., Liu, X., Wang, J.Z., Xue, Q.K., Li, Z.Q., Tse, J.S., Zhang, Z.Y., and Zhang, S.B. (2002) *Phys. Rev. Lett.*, **88**, 066101.

138 Suto, S., Sakamoto, K., Wakita, T., Harada, M., and Kasuya, A. (1998) *Surf. Sci.*, **404**, 523.

139 Nakaya, M., Nakayama, T., Kuwahara, Y., and Aono, M. (2006) *Surf. Sci.*, **600**, 2810.

140 Wang, X.D., Hashizume, T., Shinohara, H., Saito, Y., Nishina, Y., and Sakurai, T. (1992) *Jpn. J. Appl. Phys. 2*, **31**, L983.

141 Wang, H.Q., Zeng, C.G., Li, Q.X., Wang, B., Yang, J.L., Hou, J.G., and Zhu, Q.S. (1999) *Surf. Sci.*, **442**, L1024.

142 Makoudi, Y., Arab, M., Palmino, F., Duverger, E., Ramseyer, C., Picaud, F., and Cherioux, F. (2007) *Angew. Chem., Int. Ed.*, **46**, 9287.

143 Makoudi, Y., El Garah, M., Palmino, F., Duverger, E., Arab, M., and Cherioux, F. (2008) *Surf. Sci.*, **602**, 2719.

144 Stimpel, T., Schraufstetter, M., Baumgartner, H., and Eisele, I. (2002) *Mater. Sci. Eng., B - Solid*, **89**, 394.

145 Makoudi, Y., Palmino, F., Arab, M., Duverger, E., and Cherioux, F. (2008) *J. Am. Chem. Soc.*, **130**, 6670.

146 Zhang, Y.P., Yong, K.S., Lai, Y.H., Xu, G.Q., and Wang, X.S. (2004) *Appl. Phys. Lett.*, **85**, 2926.

147 Ratsch, C., Seitsonen, A.P., and Scheffler, M. (1997) *Phys. Rev. B*, **55**, 6750.

148 Ait-Mansour, K., Ruffieux, P., Xiao, W., Groning, P., Fasel, R., and Groning, O. (2006) *Phys. Rev. B*, **74**, 195418.

149 Dil, H., Lobo-Checa, J., Laskowski, R., Blaha, P., Berner, S., Osterwalder, J., and Greber, T. (2008) *Science*, **319**, 1824.

150 Wang, S.C., Yilmaz, M.B., Knox, K.R., Zaki, N., Dadap, J.I., Valla, T., Johnson, P.D., and Osgood, R.M. (2008) *Phys. Rev. B*, **77**, 115448.

151 Gambardella, P., Blanc, M., Brune, H., Kuhnke, K., and Kern, K. (2000) *Phys. Rev. B*, **61**, 2254.

152 Ahn, J.R., Kim, Y.J., Lee, H.S., Hwang, C.C., Kim, B.S., and Yeom, H.W. (2002) *Phys. Rev. B*, **66**, 153403.

153 Lipton-Duffin, J.A., Mark, A.G., MacLeod, J.M., and McLean, A.B. (2008) *Phys. Rev. B*, **77**, 125419.

154 Jalochowski, M. and Bauer, E. (2001) *Prog. Surf. Sci.*, **67**, 79.

155 Tegenkamp, C. (2009) *J. Phys. Condens. Matter*, **21**, 013002.

156 Ismach, A., Segev, L., Wachtel, E., and Joselevich, E. (2004) *Angew. Chem., Int. Ed.*, **43**, 6140.

157 Wawro, A., Suto, S., Czajka, R., and Kasuya, A. (2003) *Phys. Rev. B*, **67**, 195401.

158 Neel, N., Kroger, J., and Berndt, R. (2006) *Adv. Mater.*, **18**, 174.

159 Neel, N., Kroger, J., and Berndt, R. (2006) *Appl. Phys. Lett.*, **88**, 163101.

160 Vladimirova, M., Stengel, M., De Vita, A., Baldereschi, A., Bohringer, M., Morgenstern, K., Berndt, R., and Schneider, W.D. (2001) *Europhys. Lett.*, **56**, 254.

161 Kroger, J., Jensen, H., and Neel, N. (2007) *Surf. Sci.*, **601**, 4180.

162 Treier, M., Ruffieux, P., Schillinger, R., Greber, T., Mullen, K., and Fasel, R. (2008) *Surf. Sci.*, **602**, L84.

163 Kroger, J., Neel, N., Jensen, H., Berndt, R., Rurali, R., and Lorente, N. (2006) *J. Phys. Condens. Matter*, **18**, S51.

164 Du, S.X., Gao, H.J., Seidel, C., Tsetseris, L., Ji, W., Kopf, H., Chi, L.F., Fuchs, H., Pennycook, S.J., and Pantelides, S.T. (2006) *Phys. Rev. Lett.*, **97**, 156105.

165 Zhong, D.Y., Wang, W.C., Dou, R.F., Wedeking, K., Erker, G., Chi, L.F., and Fuchs, H. (2007) *Phys. Rev. B*, **76**, 205428.

166 Jones, T.E., Baddeley, C.J., Gerbi, A., Savio, L., Rocca, M., and Vattuone, L. (2005) *Langmuir*, **21**, 9468.

167 Chen, Q. and Richardson, N.V. (2003) *Prog. Surf. Sci.*, **73**, 59.

168 Chen, Q., Frankel, D.J., and Richardson, N.V. (2001) *Langmuir*, **17**, 8276.

169 Zhao, X.Y., Wang, H., Zhao, R.G., and Yang, W.S. (2001) *Mater. Sci. Eng., C Biomim.*, **16**, 41.

170 Leibsle, F.M., Haq, S., Frederick, B.G., Bowker, M., and Richardson, N.V. (1995) *Surf. Sci.*, **343**, L1175.

171 Fanetti, M., Gavioli, L., and Sancrotti, M. (2006) *Adv. Mater.*, **18**, 2863.

172 Gavioli, L., Fanetti, M., Sancrotti, M., and Betti, M.G. (2005) *Phys. Rev. B*, **72**, 035458.

173 Ma, X., Meyerheim, H.L., Barthel, J., Kirschner, J., Schmitt, S., and Umbach, E. (2004) *Appl. Phys. Lett.*, **84**, 4038.

174 Lin, J.L., Petrovykh, D.Y., Kirakosian, A., Rauscher, H., Himpsel, F.J., and Dowben, P.A. (2001) *Appl. Phys. Lett.*, **78**, 829.

175 Rauscher, H., Jung, T.A., Lin, J.L., Kirakosian, A., Himpsel, F.J., Rohr, U., and Mullen, K. (1999) *Chem. Phys. Lett.*, **303**, 363.

176 Garnier, F., Horowitz, G., Peng, X.H., and Fichou, D. (1990) *Adv. Mater.*, **2**, 592.

177 Sirringhaus, H., Brown, P.J., Friend, R.H., Nielsen, M.M., Bechgaard, K., Langeveld-Voss, B.M.W., Spiering,

A.J.H., Janssen, R.A.J., Meijer, E.W., Herwig, P., and de Leeuw, D.M. (1999) *Nature*, **401**, 685.
178 Grim, P.C.M., De Feyter, S., Gesquiere, A., Vanoppen, P., Rucker, M., Valiyaveettil, S., Moessner, G., Mullen, K., and De Schryver, F.C. (1997) *Angew. Chem., Int. Ed.*, **36**, 2601.
179 Takami, T., Ozaki, H., Kasuga, M., Tsuchiya, T., Mazaki, Y., Fukushi, D., Ogawa, A., Uda, M., and Aono, M. (1997) *Angew. Chem., Int. Ed.*, **36**, 2755.
180 Nishio, S., I-i, D., Matsuda, H., Yoshidome, M., Uji-i, H., and Fukumura, H. (2005) *Jpn. J. Appl. Phys. 1*, **44**, 5417.
181 Okawa, Y. and Aono, M. (2001) *Nature*, **409**, 683.
182 Okawa, Y. and Aono, M. (2001) *J. Chem. Phys.*, **115**, 2317.
183 Wan, L.J. (2006) *Acc. Chem. Res.*, **39**, 334.
184 Takajo, D., Okawa, Y., Hasegawa, T., and Aono, M. (2007) *Langmuir*, **23**, 5247.
185 Giridharagopal, R. and Kelly, K.F. (2008) *ACS Nano*, **2**, 1571.
186 Wen, R., Pan, G.B., and Wan, U.J. (2008) *J. Am. Chem. Soc.*, **130**, 12123.
187 Sakaguchi, H., Matsumura, H., Gong, H., and Abouelwafa, A.M. (2005) *Science*, **310**, 1002.
188 Sakaguchi, H., Matsumura, H., and Gong, H. (2004) *Nat. Mater.*, **3**, 551.
189 Yang, L.Y.O., Chang, C., Liu, S., Wu, C., and Yau, S.L. (2007) *J. Am. Chem. Soc.*, **129**, 8076.
190 Matena, M., Riehm, T., Stohr, M., Jung, T.A., and Gade, L.H. (2008) *Angew. Chem., Int. Ed.*, **47**, 2414.
191 Grill, L., Dyer, M., Lafferentz, L., Persson, M., Peters, M.V., and Hecht, S. (2007) *Nat. Nanotechnol.*, **2**, 687.
192 Lipton-Duffin, J.A., Ivasenko, O., Perepichka, D.F., and Rosei, F. (2009) *Small*, **5**, 592.
193 Lu, J., Lei, S.B., Zeng, Q.D., Kang, S.Z., Wang, C., Wan, L.J., and Bai, C.L. (2004) *J. Phys. Chem. B*, **108**, 5161.
194 Lei, S.B., Wang, C., Fan, X.L., Wan, L.J., and Bai, C.L. (2003) *Langmuir*, **19**, 9759.
195 Yoshimoto, S., Suto, K., Tada, A., Kobayashi, N., and Itaya, K. (2004) *J. Am. Chem. Soc.*, **126**, 8020.
196 Yoshimoto, S., Tsutsumi, E., Fujii, O., Narita, R., and Itaya, K. (2005) *Chem. Commun.*, 1188.
197 Bonifazi, D., Spillmann, H., Kiebele, A., de Wild, M., Seiler, P., Cheng, F.Y., Guntherodt, H.J., Jung, T., and Diederich, F. (2004) *Angew. Chem., Int. Ed.*, **43**, 4759.
198 Spillmann, H., Kiebele, A., Stohr, M., Jung, T.A., Bonifazi, D., Cheng, F.Y., and Diederich, F. (2006) *Adv. Mater.*, **18**, 275.
199 Bonifazi, D., Kiebele, A., Stohr, M., Cheng, F.Y., Jung, T., Diederich, F., and Spillmann, H. (2007) *Adv. Funct. Mater.*, **17**, 1051.
200 Hoeppener, S., Chi, L.F., and Fuchs, H. (2002) *Nano Lett.*, **2**, 459.
201 Lei, S.B., Wang, C., Yin, S.X., Wan, L.J., and Bai, C.L. (2003) *ChemPhysChem*, **4**, 1114.
202 Kong, X.H., Deng, K., Yang, Y.L., Zeng, Q.D., and Wang, C. (2007) *J. Phys. Chem. C*, **111**, 17382.
203 Griessl, S.J.H., Lackinger, M., Jamitzky, F., Markert, T., Hietschold, M., and Heckl, W.A. (2004) *Langmuir*, **20**, 9403.
204 Theobald, J.A., Oxtoby, N.S., Phillips, M.A., Champness, N.R., and Beton, P.H. (2003) *Nature*, **424**, 1029.
205 Perdigao, L.M.A., Perkins, E.W., Ma, J., Staniec, P.A., Rogers, B.L., Champness, N.R., and Beton, P.H. (2006) *J. Phys. Chem. B*, **110**, 12539.
206 Perdigao, L.M.A., Saywell, A., Fontes, G.N., Staniec, P.A., Goretzki, G., Phillips, A.G., Champness, N.R., and Beton, P.H. (2008) *Chem. Eur. J.*, **14**, 7600.
207 Silly, F., Shaw, A.Q., Castell, M.R., and Briggs, G.A.D. (2008) *Chem. Commun.*, 1907.
208 Saywell, A., Magnano, G., Satterley, C.J., Perdigao, L.M.A., Champness, N.R., Beton, P.H., and O'Shea, J.N. (2008) *J. Phys. Chem. C*, **112**, 7706.
209 Weber, U.K., Burlakov, V.M., Perdigao, L.M.A., Fawcett, R.H.J., Beton, P.H., Champness, N.R., , J.H., Briggs, G.A.D., and Pettifor, D.G. (2008) *Phys. Rev. Lett.*, **100**, 156101.
210 Silly, F., Shaw, A.Q., Porfyrakis, K., Warner, J.H., Watt, A.A.R., Castell, M.R., Umemoto, H., Akachi, T., Shinohara, H.,

and Briggs, G.A.D. (2008) *Chem. Commun.*, 4616.
211 Swarbrick, J.C., Rogers, B.L., Champness, N.R., and Beton, P.H. (2006) *J. Phys. Chem. B*, **110**, 6110.
212 Zhang, H.M., Xie, Z.X., Long, L.S., Zhong, H.P., Zhao, W., Mao, B.W., Xu, X., and Zheng, L.S. (2008) *J. Phys. Chem. C*, **112**, 4209.
213 Xu, W., Dong, M.D., Gersen, H., Rauls, E., Vazquez-Campos, S., Crego-Calama, M., Reinhoudt, D.N., Stensgaard, I., Laegsgaard, E., Linderoth, T.R., and Besenbacher, F. (2007) *Small*, **3**, 854.
214 Madueno, R., Raisanen, M.T., Silien, C., and Buck, M. (2008) *Nature*, **454**, 618.
215 Leininger, S., Olenyuk, B., and Stang, P.J. (2000) *Chem. Rev.*, **100**, 853.
216 Stepanow, S., Lin, N., Barth, J.V., and Kern, K. (2006) *Chem. Commun.*, **2006**, 2153.
217 Schickum, U., Decker, R., Klappenberger, F., Zoppellaro, G., Klyatskaya, S., Ruben, M., Silanes, I., Arnau, A., Kern, K., Brune, H., and Barth, J.V. (2007) *Nano Lett.*, **7**, 3813.
218 Decker, R., Schlickum, U., Klappenberger, F., Zoppellaro, G., Klyatskaya, S., Ruben, M., Barth, J.V., and Brune, H. (2008) *Appl. Phys. Lett.*, **93**, 243102.
219 Furukawa, S., Uji-i, H., Tahara, K., Ichikawa, T., Sonoda, M., De Schryver, F.C., Tobe, Y., and De Feyter, S. (2006) *J. Am. Chem. Soc.*, **128**, 3502.
220 Schull, G., Douillard, L., Fiorini-Debuisschert, C., Charra, F., Mathevet, F., Kreher, D., and Attias, A.J. (2006) *Adv. Mater.*, **18**, 2954.
221 Schull, G., Douillard, L., Fiorini-Debuisschert, C., Charra, F., Mathevet, F., Kreher, D., and Attias, A.J. (2006) *Nano Lett.*, **6**, 1360.
222 Tahara, K., Lei, S., Mossinger, D., Kozuma, H., Inukai, K., Van der Auweraer, M., De Schryver, F.C., Hoger, S., Tobe, Y., and De Feyter, S. (2008) *Chem. Commun.*, 3897.
223 Furukawa, S., Tahara, K., De Schryver, F.C., Van der Auweraer, M., Tobe, Y., and De Feyter, S. (2007) *Angew. Chem., Int. Ed.*, **46**, 2831.
224 Lei, S., Surin, M., Tahara, K., Adisoejoso, J., Lazzaroni, R., Tobe, Y., and De Feyter, S. (2008) *Nano Lett.*, **8**, 2541.
225 Tahara, K., Lei, S., Mamdouh, W., Yamaguchi, Y., Ichikawa, T., Uji,-I H., Sonoda, M., Hirose, K., De Schryver, F.C., De Feyter, S., and Tobe, Y. (2008) *J. Am. Chem. Soc.*, **130**, 6666.
226 Hoeppener, S., Wonnemann, J., Chi, L.F., Erker, G., and Fuchs, H. (2003) *ChemPhysChem*, **4**, 490.
227 Pan, G.B., Liu, J.M., Zhang, H.M., Wan, L.J., Zheng, Q.Y., and Bai, C.L. (2003) *Angew. Chem., Int. Ed.*, **42**, 2747.
228 Zhang, H.L., Chen, W., Chen, L., Huang, H., Wang, X.S., Yuhara, J., and Wee, A.T.S. (2007) *Small*, **3**, 2015.
229 Zeng, C.G., Wang, B., Li, B., Wang, H.Q., and Hou, J.G. (2001) *Appl. Phys. Lett.*, **79**, 1685.
230 Huang, H., Chen, W., Chen, L., Zhang, H.L., Sen Wang, X., Bao, S.N., and Wee, A.T.S. (2008) *Appl. Phys. Lett.*, **92**, 023105.
231 Xu, B., Tao, C.G., Cullen, W.G., Reutt-Robey, J.E., and Williams, E.D. (2005) *Nano Lett.*, **5**, 2207.
232 Schmaltz, B., Rouhanipour, A., Rader, H.J., Pisula, W., and Mullen, K. (2009) *Angew. Chem., Int. Ed.*, **48**, 720.
233 Mena-Osteritz, E. and Bauerle, P. (2006) *Adv. Mater.*, **18**, 447.
234 De Feyter, S. and De Schryver, F.C. (2003) *Chem. Soc. Rev.*, **32**, 139.
235 Humblot, V., Barlow, S.M., and Raval, R. (2004) *Prog. Surf. Sci.*, **76**, 1.
236 Barlow, S.M. and Raval, R. (2003) *Surf. Sci. Rep.*, **50**, 201.
237 Ernst, K.H. (2006) *Supramolecular Chirality*, vol. 265, Springer-Verlag Berlin, Berlin, p. 209.
238 Kuhnle, A., Linderoth, T.R., and Besenbacher, F. (2003) *J. Am. Chem. Soc.*, **125**, 14680.
239 Kuhnle, A., Linderoth, T.R., Hammer, B., and Besenbacher, F. (2002) *Nature*, **415**, 891.
240 Kuhnle, A., Linderoth, T.R., and Besenbacher, F. (2006) *J. Am. Chem. Soc.*, **128**, 1076.
241 Fasel, R., Parschau, M., and Ernst, K.H. (2003) *Angew. Chem., Int. Ed.*, **42**, 5178.

242 Parschau, M., Romer, S., and Ernst, K.H. (2004) *J. Am. Chem. Soc.*, **126**, 15398.

243 Vidal, F., Delvigne, E., Stepanow, S., Lin, N., Barth, J.V., and Kern, K. (2005) *J. Am. Chem. Soc.*, **127**, 10101.

244 Cortes, R., Mascaraque, A., Schmidt-Weber, P., Dil, H., Kampen, T.U., and Horn, K. (2008) *Nano Lett.*, **8**, 4162.

245 Schock, M., Otero, R., Stojkovic, S., Hummelink, F., Gourdon, A., Laegsgaard, E., Stensgaard, I., Joachim, C., and Besenbacher, F. (2006) *J. Phys. Chem. B*, **110**, 12835.

246 Schlickum, U., Decker, R., Klappenberger, F., Zoppellaro, G., Klyatskaya, S., Auwarter, W., Neppl, S., Kern, K., Brune, H., Ruben, M., and Barth, J.V. (2008) *J. Am. Chem. Soc.*, **130**, 11778.

247 Mu, Z.C., Shu, L.J., Fuchs, H., Mayor, M., and Chi, L.F. (2008) *J. Am. Chem. Soc.*, **130**, 10840.

248 Chen, Q. and Richardson, N.V. (2003) *Nat. Mater.*, **2**, 324.

249 Blankenburg, S. and Schmidt, W.G. (2007) *Phys. Rev. Lett.*, **99**, 196107.

250 Furukawa, S. and De Feyter, S. (2009) *Top. Curr. Chem.*, **287**, 87.

251 Cicoira, F., Santato, C., and Rosei, F. (2008) *STM and AFM Studies on (Bio) Molecular Systems: Unravelling The Nanoworld*, vol. 285, Springer-Verlag Berlin, Berlin, p. 203.

10
Surface Microstructures and Nanostructures in Natural Systems

Taolei Sun and Lei Jiang

10.1
Introduction

Biosystems in Nature have evolved for billions of years, with their structures and functions having reached an optimized state during the evolutionary process. Natural biomaterials obtained from animals or plants normally exhibit certain unique properties that are far superior to those of artificial biomaterials. On closer examination, however, these properties are found to be determined not only by the intrinsic properties of the materials but, more importantly, they are related to the well-designed topological structures at both the micro level and nano level. The details of some delicate surface nanostructures present in natural systems, based on their contributions to different surface properties, are discussed in the following subsections.

10.2
Surface Nanostructures and Special Wettability

Wettability is a fundamental property of a material surface, and plays important roles in many aspects of human activities. Recently, special wettability has aroused much interest because its great advantages in applications [1]. For example, a superhydrophilic surface with a water contact angle (CA) of about 0°, generated by ultraviolet (UV) irradiation has been used successfully as a transparent coating with antifogging and self-cleaning properties [2]. Likewise, various phenomena that include contamination, snow sticking, erosion, and even the conduction of an electrical current, would be expected to be inhibited on superhydrophobic surfaces [3–6] with a CA larger than 150° and a sliding angle (SA) less than 10° [7]. In Nature, many interesting phenomena occur that are relevant to the special wettability that proves to be highly convenient for the lives of plants and animals. Examples include the self-cleaning effect on the lotus leaf, the super water-repellent force of the water strider's legs, and the anisotropic wetting and dewetting properties of the rice leaf. Yet, various studies have indicated that these phenomena are contributed to not only by the chemical

Nanotechnology, Volume 8: Nanostructured Surfaces. Edited by Lifeng Chi

ISBN: 978-3-527-31739-4

properties of the surface, but more importantly, are governed by the special structural effects at the microlevel and nanolevel.

Although the chemical compositions [8, 9] determine the surface free energy, and thus have a major influence on wettability, the system has certain limitations. For example, a $-CF_3$-terminated surface was reported to possess the lowest free energy and the best hydrophobicity whereas, on flat surfaces, the maximum CA achieved was only about 120° [10]. One other important factor that influences wettability is the surface topographic structure, as described by Wenzel's equation [11]:

$$\cos\theta' = r\cos\theta \tag{10.1}$$

where θ' is the apparent CA on a rough surface, θ is the intrinsic CA on a flat surface, and the surface roughness (r) can enhance both the hydrophilicity and hydrophobicity of the surfaces. Thus, the modified Cassie's equation becomes [12]:

$$\cos\theta' = f\cos\theta - (1-f) \tag{10.2}$$

in which f is the fraction of solid/water interface, while $(1-f)$ is that of the air/water interface. This indicates that when a rough surface comes into contact with water, air trapping in the trough area may occur, which would contribute greatly to the increase in hydrophobicity, and help to achieve superhydrophobicity with a CA larger than 150°. These are the most important mechanisms for the special wettability phenomena in Nature.

However, the structural effect has a much greater role, as it may also alter the properties of the solid/water/air triple contact line (TCL), thus greatly influencing the dynamic aspects of wettability. The wettability also shows a distinct size effect for the nanostructures, and this plays important roles not only in the mesoscale assembly of the bio-units but also the stability of the biostructures.

10.2.1
Self-Cleaning Effect on the Lotus Leaf

The self-cleaning effect exists widely in Nature on the surface of plant leaves, with perhaps the best-known example being that of the lotus leaf (Figure 10.1a); indeed, the "lotus effect" is so-named for this very reason! The effect involves two main aspects: (i) superhydrophobicity (CA > 150°) of the surface; and (ii) a strong anti-adhesive effect towards water (i.e., a small SA). Although initially, Barthlott and Neihuis [13] considered the lotus effect to be induced by the coexistence of wax compounds and micrometer-scale papillae structures on the surface, recent studies have indicated that nanostructures still occur in the micropapillae (Figure 10.1b), with diameters of about 5–9 μm. As shown in Figure 10.1c, each papilla is composed of further nanofibrous structures with an average diameter of about 120 nm, and these are also found on the lower part of the leaf. The static CA and the SA on this surface are about $161 \pm 3°$ and less than 3°, respectively.

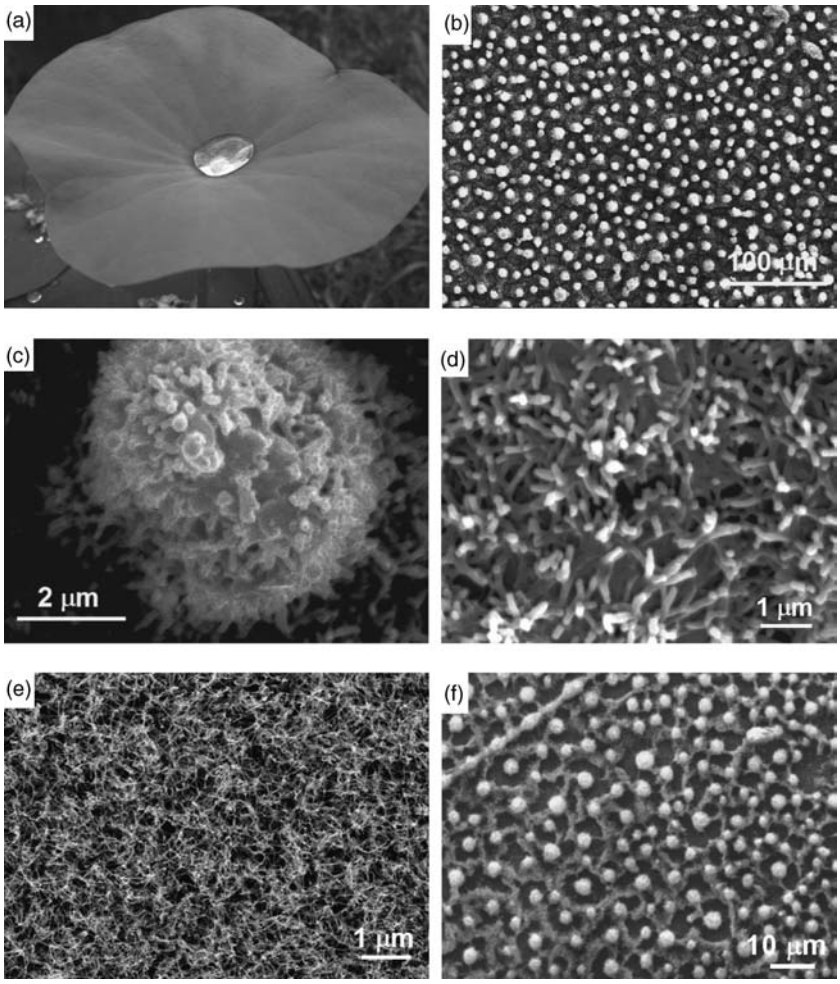


Figure 10.1 Microstructures and nanostructures on the lotus leaf. (a) The lotus leaf; (b) Large-scale scanning electron microscopy (SEM) image of the lotus leaf. Each epidermal cell forms a papilla, and has a dense layer of epicular waxes superimposed on it; (c) Magnified image of a single papilla of panel (b); (d) SEM image on the lower surface of the lotus leaf; (e) Densely packed ACNT film with pure nanostructure (top view); (f) Lotus-like ACNT film with multilevel microstructures and nanostructures. Panels (b–f) adapted from Ref. [4].

According to Adamson and Gast [14], a theoretical model can be built for the relationship between the superhydrophobicity and the multilevel surface structure. In this model, the multilevel structures on lotus leaf can be described as a fractal structure that is similar to the Koch curve [15], where the fractal dimension (D) was used to characterize the roughness of the surface. According to Wenzel's formula [see Eq. (10.1)] and Cassie's equations [Eq. (10.2)], the relationship between the CA

(θ_f) on the rough surface and that (θ_0) on the corresponding smooth surface can be described as:

$$\cos\theta_f = f_s\left(\frac{L}{l}\right)^{D-2}\cos\theta_0 - f_v \tag{10.3}$$

where $(L/l)^{D-2}$ is the roughness factor of the surface, and L and l represent the maximum and minimum sizes, respectively, for surface structures with the fractal behavior. For the lotus leaf surface, these correspond to the diameters of the micro-papillae and the nanofibers in each micro-papilla, respectively. In the Koch curve, D is about 2.282, and (L/l) is 3^n, where n is an integer determined by the exact fractal structure. An increase in n illustrates a corresponding increase in the surface roughness. f_s and f_v ($f_s + f_v = 1$) represent the fractions of the solid/water and air/water interfaces when the surface contacts with water. Thus, a relationship between θ_f on the surface of lotus leaf and the value of n can be obtained, as shown in Figure 10.1d. By using the above results, it is possible to calculate the theoretical diameter of the nanofibers that corresponds to a CA value of about 160°; a value of 128 nm is very close to the experimental value. This analysis shows clearly why Nature would select multiscale structures to achieve a superior self-cleaning effect, but not only microstructures or nanostructures [16].

In order to identify the role of the hierarchical structures in the self-cleaning effect, the present authors' group synthesized aligned carbon nanotube (ACNT) films with and without hierarchical structures, for comparison. The ACNT film [17] with a pure nanostructure (Figure 10.1e) was fabricated using a chemical vapor deposition (CVD) method on a silica substrate with homogeneous catalyst distribution, and demonstrated superhydrophobicity with a CA of about 158 ± 2°; however, the fact that the SA was greater than 30° indicated a relatively large CA hysteresis and a strong adhesion to water. When the ACNT film was fabricated using the same CVD method on a silica substrate, but with a heterogeneous catalyst distribution, similar hierarchical microstructures and nanostructures were seen (Figure 10.1f) as on the lotus leaf, but the CA on the surface was about 166° and the SA only about 3°. A further honeycomb-like ACNT film [18] with a hierarchical structure was also fabricated that showed a large CA of about 163° and a small SA of <5°. A subsequent comparison of the data obtained with these films confirmed that the hierarchical structure not only further improved the hydrophobicity but, most importantly, also provided a small SA.

Similar phenomena are also observed on many other plant leaves, such as Indian Cress or Lady's Mantle (*Alchemilla vulgaris* L.). However, following extensive investigations of the mechanisms involved, these phenomena were shown to be relevant to special microstructures and nanostructures on the leaves, despite differences in the shape and arrangement of the structures, and the corresponding mechanisms. On the basis of these findings, superhydrophobicity and other more specialized properties of wettability have been demonstrated on a variety of artificial functional surfaces [19–22], using several methods. Two examples of artificial superhydrophobic surfaces prepared by the present authors' group are shown in Figure 10.2a and b [23, 24]. Moreover, when a functional surface was combined with well-designed microstructures and nanostructures, a reversible switching between

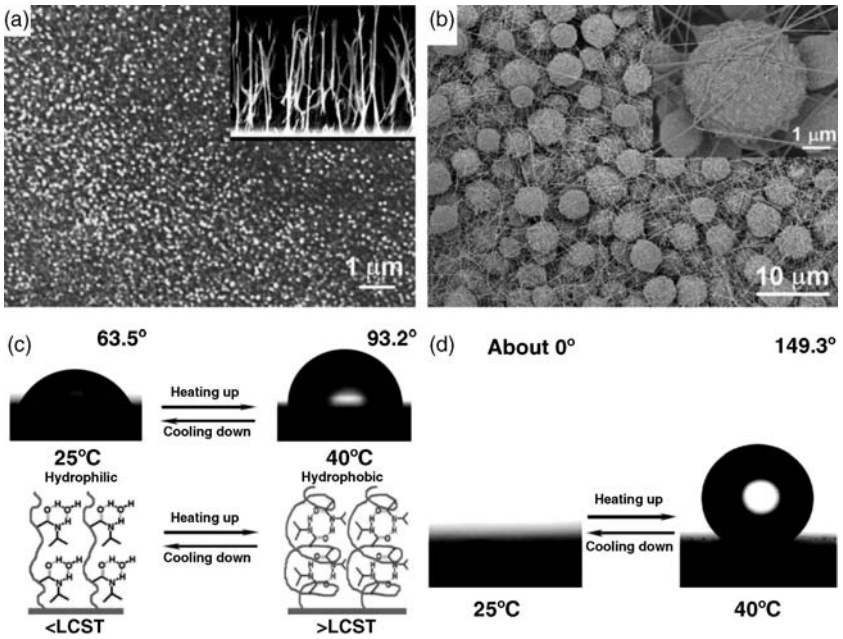


Figure 10.2 Artificial surfaces with special wettability. (a) An aligned polyacrylonitrile (PAN) nanofiber film (top view) with superhydrophobicity. The inset shows a side view; (b) Superhydrophobic polystyrene film with microsphere/nanofiber composite structure; (c) Temperature-sensitive wettability of a poly(*N*-isopropylacrylamide) film on a flat substrate; (d) Reversible switching between superhydrophilicity and superhydrophobicity on a structured substrate. LCST = lowest critical solution temperature (of the responsive polymer). Adapted from Refs [23], [24], and [25].

superhydrophobicity and superhydrophilicity (Figure 10.2c and d) could be conveniently achieved, using a variety of methods that included thermal treatment [25], light irradiation [26, 27], and solvent treatment [28]. Ultimately, the results of such studies opened up a novel research field for biomimetic materials, and greatly extended the application domains of specialized wettability.

10.2.2
Multifunctional Surfaces of Insect Wings with a Self-Cleaning Effect

In addition to plant leaves, self-cleaning effects may also be observed on the wings of many insects, including dragon fly, honeybee, cicada, and moths, where the main benefit is to protect against the wings being wetted by the dew. Notably, the self-cleaning effects of insect wings differs from that of surfaces, in that it is normally accompanied by various other extraordinary functions.

In the case of the cicada (see Figure 10.3a, inset), which lives either in the soil or on trees, the wings show excellent superhydrophobicity, with a CA $>160^\circ$ and superior dewetting properties that are of great convenience to the insect. When investigating the relationship between the wettability and nanostructures located on the cicada

wings [29, 30] (see Figure 10.3a), a well-arranged nanopillar structure with a feature size of approximately 100 nm was identified on the insect wings. Whilst the tops of the nanopillars were of dissimilar heights, and showed an uneven arrangement, the operating principle was seen to be somewhat similar to the multilevel structures on the lotus leaf. This guaranteed a repellent force that prevented water droplets from contacting the surface (as per Wenzel's equation), while simultaneously causing an efficient reduction in the contact area between the water droplet and the surface, and also in the total TCL length. This resulted in a discontinuous TCL arrangement whereby the nanopillar structure and uneven arrangement of the nanopillars, together with the wax components on the wing surface, resulted not only in an outstanding superhydrophobicity but also an anti-adhesive effect against water. The cicada wing is also famous for its ultra-transparency, which results from its excellent anti-reflective properties that have been shown to result from the nanostructure's periodic arrangement. In order to monitor this effect, the present authors' group used a template-based "rolling press" technique to create well-patterned nanopillars (Figure 10.3b) on a polymer surface, thus mimicking the nanostructures on the cicada wings. Subsequently, a remarkable similarity was observed for the nanostructures (see Figure 10.3a and b), while the as-prepared film also showed a similar superhydrophobic property on its surface.

Another example of low-reflective properties can be found in the compound eyes of moths [31], mosquitoes, butterflies, or other insects. In the case of moths, this makes them difficult to be detected by their natural enemies whilst, at the same time, providing perfect antifogging and dewetting properties [32] that help to maintain clear vision in a humid environment. The reason for these benefits also relate to the special nanostructures on the eyes, with studies on mosquito compound eyes having shown the duplex functions to have originated from the papilla structure on the micrometer scale, and from the hexagonally arranged nanostructures on its surface [32].

In summary, both microstructures and nanostructures on the surface of organisms can help to integrate several peculiar functions into a simple system. This in turn provides much insight into the development of novel functional devices. An example was shown in a biomimetic study conducted by Gao *et al.*, who fabricated artificial compound eyes using a soft-lithographic technique and then discussed the

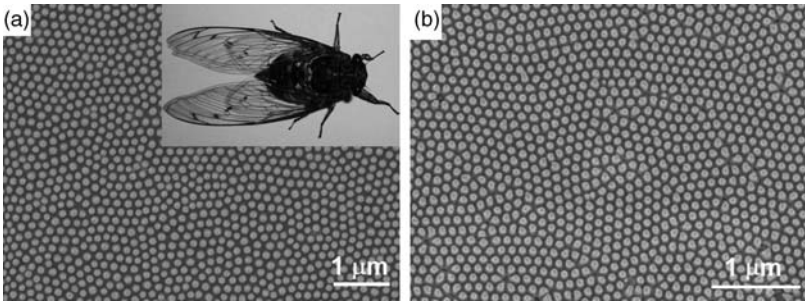


Figure 10.3 Nanostructure (a) of the cicada wing and (b) of the artificial simulation by a template-based method. Adapted from Ref. [29].

influence of the microstructures and nanostructures on the eyes' properties. Gu *et al*. [33] were also successful in creating functional nanomaterials with both coloration properties and superhydrophobicity. Taken together, these results have provided an important theoretical and practical background for the exploration of other biomimetically functional materials.

10.2.3
The Anisotropic Dewetting Property on Plant Leaves

Whilst, on a lotus leaf, water is able to roll freely in any direction over the entire surface, on the rice leaf an anisotropic dewetting property is observed, whereby the water droplet can roll freely in only *one* direction. The scanning electron microscopy (SEM) images in Figure 10.4a and b indicate a hierarchical structure on the rice leaf [16] that is similar to the lotus leaf; accordingly, the surface is superhydrophobic. A different situation arises, however, in the case of the papillae, which are arranged in one-dimensional (1-D) order parallel to the edge of the leaf (arrows 'a' in Figure 10.4). This means that the water drop can roll off freely along this direction, but will move with greater difficulty along the perpendicular direction (arrow 'b' in Figure 10.4). The SAs in these two directions are about 3–5° and 9–15°, respectively. Such a phenomenon is also considered relevant to the anisotropic TCL arrangement. In the case of the rice leaf, the density of the micro-papillae in the parallel direction of the leaf edge is significantly less than in the perpendicular direction, and this will result in an anisotropic arrangement of the TCLs; this is in contrast to the isotropic arrangement on the lotus leaf, which is due to the homogeneous distribution of the papillae. In an attempt to mimic this phenomenon, the present authors' group prepared a rice-like ACNT film (Figure 10.4b) by controlling the surface distribution of the catalyst on which the micro-level ACNT arrays; this was achieved by patterning with different spacings in mutually orthogonal directions. A similar anisotropic dewetting phenomenon was also observed on this film.

An anisotropic dewetting phenomenon also exists on the feathers of water-fowls [34]. These birds live in water and while their feathers are waterproof, any

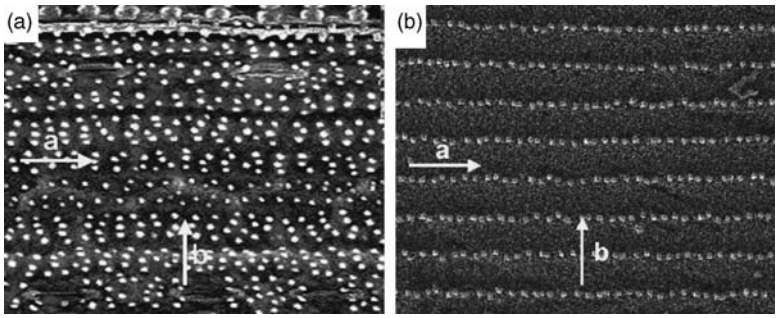


Figure 10.4 Anisotropic structures on a rice leaf (a) and an artificial ACNT film (b). Adapted from Ref. [16].

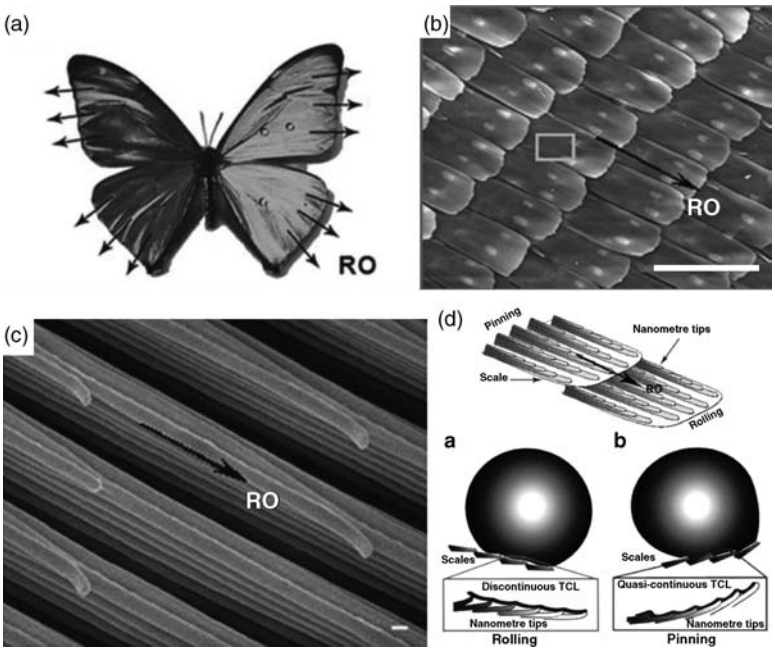


Figure 10.5 (a) Hierarchical anisotropic microstructures and nanostructures on butterfly wings; (b, c) Scanning electron microscopy images of the periodic arrangement of overlapping microscales on the wings and fine lamella-stacking nanostripes on the scales; (d) Schematic diagram of the mechanism of the anisotropic dewetting property on butterfly wings. Adapted from Ref. [35], with permission.

adhered water droplets can be easily removed simply by the bird shaking its wings. This special function is also due to the anisotropically aligned strip-like microstructures and nanostructures of the feather, which can provide an excellent hydrophobicity and water-repellent properties, whilst at the same time helping to preserve a good permeability against the air.

Recently, details were reported of the anisotropic rolling properties of water drops on the wings of the butterfly *Morpho aega* (Figure 10.5a) [35]; in this case, the wings were found to have well-defined multilevel microstructures and nanostructures. Figure 10.5b shows the SEM image of the first level of the overlapping micro-scales with a width of about 40–50 μm on the wings, which are composed of nanostripes (see Figure 10.4c) with a width of about 100 nm. Further lamellar-stacking nanotip structures (Figure 10.5c) for each nanostripe can be observed in the magnified image. This periodic stripe structure and the multilevel structures produce not only beautiful colors (known as "structural colors"; see Section 10.3.2) to the butterfly, but also bring about unique anisotropic dewetting properties. The static CA on the butterfly's wings was found to be $152 \pm 2^{\circ}$, which indicated the wing to have superhydrophobic properties. Perhaps more interestingly, however, whilst the water drop could roll along the outer direction of the microscales, it would stop in the opposite direction,

even if the wings were stood vertically. Such a property could be highly beneficial to protect the insect body against being wetted by water. The mechanism for this effect is shown in Figure 10.5d. Here, when the wing is tilted down, the oriented nanotips on the nanostripes and microscales become separated from one another; this causes any water droplets deposited on the wing to form a discontinuous TCL, which shows a small SA of about 9°. However, when the wing is tilted upwards the nanotips take on a close arrangement so as to form a continuous TCL with the water drop.

10.2.4
The Super Water-Repellent Force of the Water Strider's Legs

The water strider (Figure 10.6a) is an insect which lives in water but can run and jump both rapidly and freely on the water's surface, without wetting its legs. This interesting phenomenon has long attracted much attention, with many investigations being conducted to determine the dynamic reasons that permit the water strider to act in this way [36, 37]. By using a high-speed camera, Bush *et al.* [38] showed that the insect's leg movement created a vortex in the water that in turn created a rapid leg movement on the water's surface. However, these studies failed to meet the crux of the problem, namely, what forces prevented the water strider's legs from piercing the water surface?

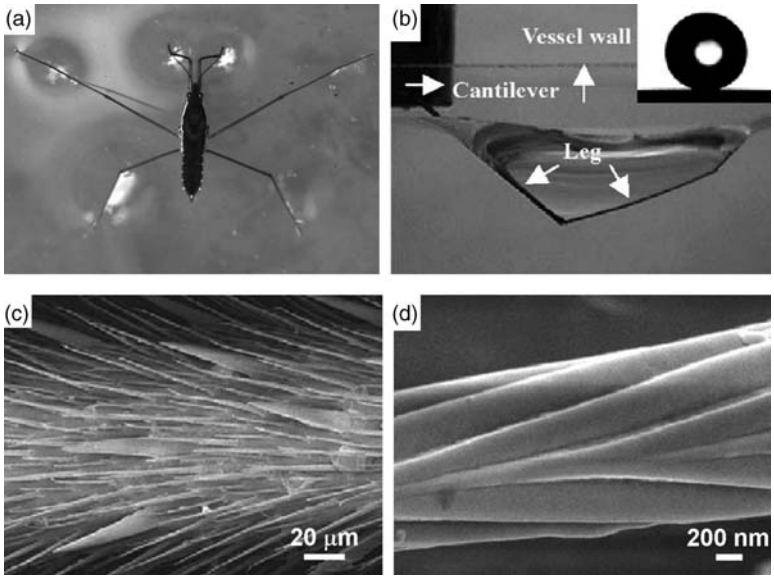


Figure 10.6 Hierarchical microstructures and nanostructures on the legs of a water spider (a); (b) Side view of the maximal dimple just before the leg pierces the water surface. The inset shows the profile of a water drop in the contact angle measurement for the superhydrophobic leg surface; (c) Scanning electron microscopy image of the leg surface, showing numerous oriented spindly microsetae; (d) Nanoscale groove structure on a seta. Adapted from Ref. [39], with permission.

A recent study conducted by Jiang's and coworkers provided some fundamental answers to these questions [39], when it shown that a large water-repellent force (Figure 10.6b) was produced by nanostructures on the water strider's legs. On examination, the surfaces of the legs were shown to bear numerous bristles that had diameters of about 1–3 μm and were arranged to lie in the same direction (see Figure 10.6c). On the surface of each bristle was a further well-defined spiral groove structure (Figure 10.6d). The nanostructure was shown capable of trapping air within the grooves, and this resulted in an excellent superhydrophobicity (see inset of Figure 10.6b) with a static CA >160° on the leg surface. In contrast, according to the relationship between the force provided by the surface tension (γ) at the border of the interface and the length of the border, the water-repellent force (f_r) provided by a leg could be written as:

$$f_r = \gamma \cos \theta_0 \cdot \sum_i l_i \cos \beta_i \tag{10.4}$$

where l_i is the length of the TCL provided by each bristle on the leg, β_i is the angle between the direction of each bristle and the vertical direction of water surface, γ is the surface tension of water, and θ_0 is the intrinsic CA of the wax material on the surface of the bristles. This shows that microbristles, and the nanostructures on them, can greatly increase the total length of the TCL, resulting in a large water-repellent force such that each leg could support about 15 times the insect's body weight. In fact, the load capacity was so high that it could assure the free activities of the water strider on water surfaces, even within some complicated environments. In time, such an effect might bring inspiration to the development of novel water robots or other devices.

10.2.5
Extraordinary Water-Harvesting Ability of the Desert Beetle's Wings

The desert is very dry, yet some insects have developed unique methods to locate water. A typical example is the *Stenocara* beetle [40] (Figure 10.7a), which lives in the Nambi Desert in South Africa. This desert supports a unique sand-dune fauna, and normally experiences high winds, extreme daytime temperatures and dense, early-morning fog, yet with a rainfall that is very low and almost negligible. Yet, in this extremely dry environment, the *Stenocara* beetle is able to collect water droplets from the fog, with assistance from the wind blow. Moreover, the water drops which form on the top elytra (wing cover) subsequently roll down the beetle's outer surface towards its mouthparts.

In examining this phenomenon, Parker and Lawrence [40] reported the existence of two types of randomly distributed arrays of bumps on the beetle's carapace; these are located 0.5–1.5 mm apart, and each bump is about 0.5 mm in diameter (Figure 10.7b). The peaks of the bumps are smooth at the microscopic level and are without any covering; thus, they are highly hydrophilic. In contrast, in the trough area the surface is covered by microstructures coated in wax (Figure 10.7c). These microstructures consist of hexagonally arranged flattened hemispheres with

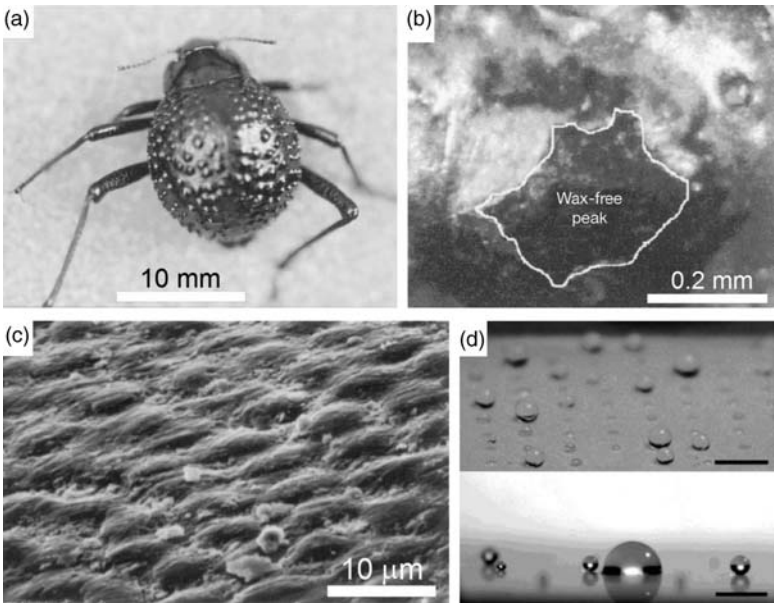


Figure 10.7 The water-capturing surface of the fused overwings (elytra) of the desert beetle. (a) Adult female beetle, dorsal view; (b) A "bump" on the elytra; (c) Scanning electron microscopy image of the textured surface of the suppressed areas; (d) Artificial simulation of the water-capturing process on the alternate hydrophobic/hydrophilic surface pattern. Panels (a–c) adapted from Ref. [40]; Panel (d) adapted from Ref. [41], with permission.

a diameter of about 10 μm, are reminiscent of those on the lotus leaf, and result in these areas having superhydrophobic properties. During foggy weather, tiny water droplets contained in the fog are able to gather on the hydrophilic peak area, where they rapidly form a larger drop. Water coming into contact with the hydrophobic grooves will also be collected by these hydrophilic regions. The water droplets then coalesce until their weight is sufficient to overcome the binding force between the water and the surface, at which point they flow down to the beetle's mouth under the action of wind blow.

This "fog-catching ability" of the *Stenocara* beetle, which is based on an alternative design of the hydrophilic and hydrophobic domains, brings important insights to the development of highly efficient water-harvesting devices that might have numerous applications in the future. For example, synthetic films could be fabricated onto polymer sheets and attached to buildings and tents so as to harvest water vapor, perhaps to serve refugee camps. A similar process might also help in the capture and recycling of water vapor from cooling towers, which would in turn lead to reductions in energy costs. With these possibilities in mind, Zhai *et al*. [41] created hydrophilic patterns on superhydrophobic surfaces by using water/isopropanol solutions of a polyelectrolyte, the aim being ultimately to produce surfaces with an extreme hydrophobic contrast. These surfaces perfectly mimicked the water-capturing mechanism on the back of the *Stenocara* beetle, and could be used to capture very small

drops of water and convert them to larger drops (Figure 10.7d). By using the same technique, it might also be possible to create superhydrophilic canals by applying superhydrophilic multilayers onto hydrophilic stripes on the superhydrophobic surface. These structures might find broader applications in other domains, such as microfluidic devices.

10.3
Structure-Related Special Optical Phenomena in Nature

When a substance is illuminated with white light, a specific color is observed because only a particular range of wavelength of light is reflected and is visible to the eye. There are two ways to eliminate the other wavelengths of light [42]. The first approach is for the substance to absorb the light, and this is the mechanism normally employed when coloring with by pigments [43, 44]. Whilst this is the main route taken by Nature to generate colors, there are many cases where Nature prefers to select structural colors to present its beauty. This is a purely physical process that depends on interactions between the light and the elaborate periodic submicro- or nanostructures with scales that are comparable to the light's wavelength [45, 46]. Compared to pigments, the structural colors have not only a much greater stability (amongst other advantages), but can also help to achieve certain special functionalities such as supertransparency and ultra-low reflection. They are, therefore, very useful in modern materials science, and have attracted much interest during recent decades.

10.3.1
Photonic Crystal Structures in Opal and Opal Analogues

The periodic submicro- or nanostructures that show special optical properties are termed "photonic crystals." Similar to semiconductor crystals, photonic crystals also show an energy band structure that influences the propagation of light (which is also an electromagnetic wave). Photons propagate through these structures, or not, depending on their wavelength. As the basic physical phenomenon is based on diffraction, the periodicity of the photon crystal structure should be of a similar length scale as the half of light wavelength in order for the material to exhibit specific colors.

One representative example in Nature is that of *opals*, which are a natural mineral that has a well-known range of beautiful colors and glosses. Opals are mineraloid gels that are composed of silica nanospheres with different sizes and arrangements. The diverse periodic arrangements of the nanostructures give rise to the changeable colors and glosses associated with opals [47]. With this in mind, a variety of artificial nanostructures have been created (Figure 10.8) that display structural colors and other functionalities [33, 48].

The opal analogue has also been found on scales of certain beetles. For example, Parker *et al*. [49] reported a beetle, *Pachyrhynchus argus* (Figure 10.9a), which was found in the forests of northeastern Queensland, Australia and has a metallic coloration that is visible from any direction. According to the studies conducted,

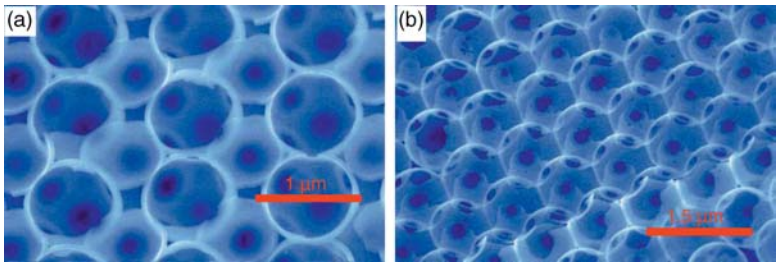


Figure 10.8 Typical scanning electron microscopy images for an artificial inverse opal structure. (a) 110 facet; (b) 111 facet. Adapted from Ref. [48], with permission.

the unique optical property that distinguishes this beetle from others is due to the photonic crystal structure that is analogous to that of opal. The beetle is seen to be covered by scales of about 0.1 mm (Figure 10.9b), which are individually flat, lie parallel with the body, and consist of an outer shell and an inner structure. Both SEM and transmission electron microscopy (TEM) images have shown that the inner structure of the scales is composed of a solid array of transparent spheres, each with a diameter of 250 nm. The nanospheres are arranged precisely in a hexagonal close-packing fashion (Figure 10.9c), thus forming a photonic crystal structure that is very

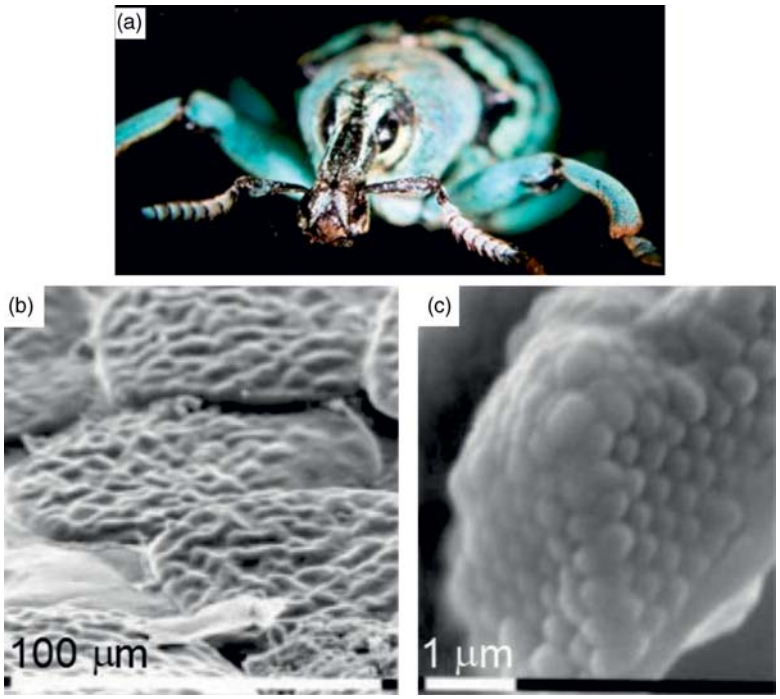


Figure 10.9 Beetle *Pachyrhynchus argus* (a) and the photonic crystal structures on its scales (b, c). Adapted from Ref. [49], with permission.

similar to the common nanostructure found in opal. As the lattice parameter of the structure is close to the half-wavelength of the visible light, the single scale acts as a three-dimensional (3-D) diffraction grating, and this allows the reflection of a narrow range of wavelengths over a wide range of incident angles. In transmitted white light, the scales appear as the negative of the reflected color: yellow-green in reflected light, and purple in transmitted light from most directions. In the spectroscopic analysis results, with white light incident at 20° to the normal direction of the scale surface, the peak reflection occurred at a wavelength of 530 nm, at an angle of 20° to the other side of the normal direction. As shown in Figure 10.9c, the photonic crystal structure is a 3-D structure; however, when compared to the similar 2-D photonic structure that was reported on the sea mouse (*Aphrodita* sp.) [50], it can provide an omnidirectional optical property to the beetles, and this may bring about important insights for the design of novel, high-performance display devices.

10.3.2
Structural Colors in Biological Systems

Almost 300 years ago, Newton [51] noted that the brilliant colors of peacock feathers (Figure 10.10a) were not caused by pigments, but rather were due to a thin-film interference mechanism. Recently, when Zi *et al*. [52] further studied the detailed mechanism, they found the coloration strategy for peacock feathers to be very delicate, to originate mainly from the periodic nanostructures of the cortex layer on barbules. As shown in Figure 10.10b, a barbule of a peacock feather consists of a medullar core of ~3 μm enclosed by a cortex layer. Interestingly, the cortex of all different-colored barbules contains a 2-D photonic crystal structure (Figure 10.10c

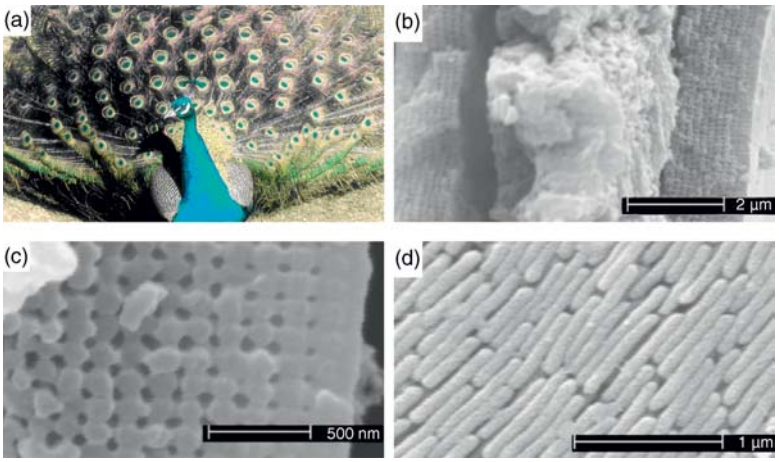


Figure 10.10 Scanning electron microscopy images of the barbule structures on a peacock's feathers (a); (b) Transverse cross-section of a green barbules; (c) Magnification of the image in panel (b); (d) Longitudinal cross-section of a green barbule with the surface layer removed. Adapted from Ref. [52], with permission.

and d), composed of melanin rods connected by keratin. Moreover, the photonic crystal structures in the differently colored barbules were quite similar, whereas the lattice constants and the number of periods varied widely. According to the analysis made by Zi *et al.*, the 2-D photonic crystal structure showed a strong reflection for the light with specific wavelengths along the direction of cortex layer, thus generating colors which were regulated by the lattice constants and the number of periods. The different colors were due to the different lattice constants, which increased regularly for the blue, green, yellow, and brown barbules. The Fabry–Perot interference effect has also been found for the brown barbules, in which the periodic number is the smallest. Due to this effect, an extra blue color was generated and contributed to the final brown color of the barbules.

For the wings of butterflies [53, 54], the periodic structure appears in different ways, and Vukusic *et al.* have undertaken several systematic investigations in this aspect. As has been shown previously [55, 56], the broad wings of butterflies are invariably covered with well-arranged arrays of minute scales, each of which is a thin, flattened, cuticular evagination from an individual cell in the wing epithelium. The scales overlap with each other, much as do roof tiles, and this functions as a quarter-wave interference device by presenting a series of alternating lucent and dense layers. The scales are composed further of complicated, delicately arranged nanostructures (see Figure 10.11), but this differs very much among the butterfly species. The diverse

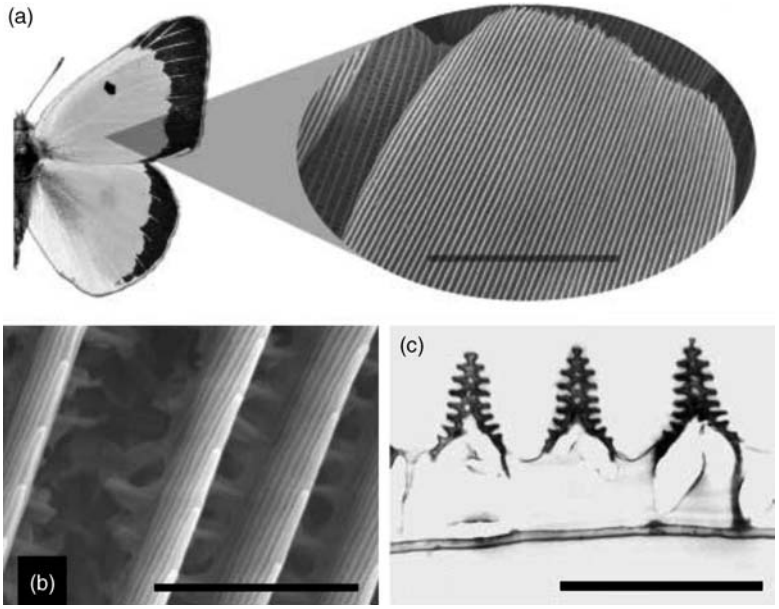


Figure 10.11 Microstructures and nanostructures on the wings of the butterfly *Colias eurytheme*. (a) Scanning electron microscopy (SEM) image of a single scale; (b) Magnified SEM image showing the ridges in close-up; (c) Transmission electron microscopy image of a cross-section through the scale, indicating the horizontal lamellae borne on the vertical scale ridges. Adapted from Ref. [58], with permission.

shapes and arrangement of these nanostructures provide butterflies with richer colors than peacocks, as well as other interesting optical properties on their wings. For example, the discrete multilayers of the cuticle and air on iridescent blue *Morpho rhetenor* butterflies [57] give rise to an ultralong-range visibility of up to 800 m, to the photonic structures of reduced dimensions that are present in certain *Colias* butterflies [58], and also effect an intense UV visibility. The orientational adjustments of such multilayers produces a highly angle-dependent iridescence that provides a high-contrast color flicker, with minimal wing movement.

Structural colors are also observed in plants, and normally mediated by the 1-D multilayer structure. These structures give rise to vivid colors and iridescence in vascular leaves, fruits and marine algae, which not only makes them beautiful but also greatly affects the development of the plant. For example, iridescence in leaves is believed to produce particular intensity ratios of incident radiation bands that can penetrate to phytochrome centers. In fruit skin, this can reduce post-maturation discoloration and ultimately improve dispersal.

10.3.3
The Directional Fluorescence Emission Property in *Papilio* Butterflies

Vukusic *et al*. also studied in detail the directionally controlled fluorescence emission properties in *Papilio* butterflies (*Princeps nireus* group) [59]. These butterflies have dark wings with bright blue or blue-green dorsal wing bands or patches. Interestingly, the studies of Vukusic *et al*. showed that nanostructures on the wings perfectly matched the design of high-efficiency light-emitting-diode (LED) devices that use 2-D photonic-crystal geometries to enhance the extraction efficiency of light, and also distributed Bragg reflectors (DBRs) to control emission direction. As shown in Figure 10.12a, the wing scales from their colored regions make up a nanostructure that is characterized by a ~2 μm-thick 2-D photonic crystal slab (PCS) of hollow air cylinders, with a mean diameter of about 240 nm and a spacing of about 240 nm in a medium of solid cuticle (Figure 10.12b). The PCS is infused exclusively with

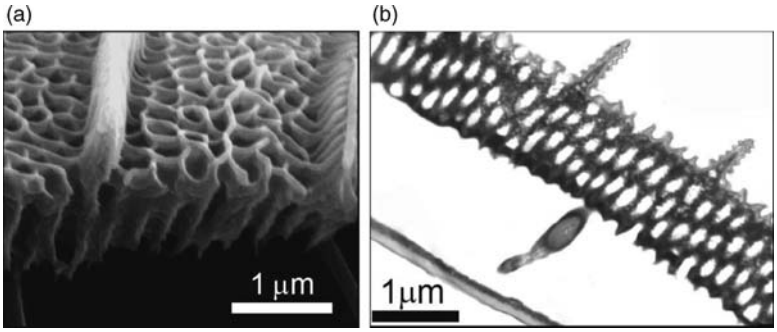


Figure 10.12 (a) Scanning electron microscopy image of the air cylinder nanostructure on the scale of butterfly *P. nireus*; (b) Transmission electron microscopy image of a section through the scale. Adapted from Ref. [59], with permission.

a highly fluorescent pigment, which rests parallel to and ~1.5 μm above a three-layer, cuticle-based DBR and forms the base of the scale. The DBRs reflect upwardly the downward-emitted fluorescence from the nanostructure, concurrent with nonabsorbed longer wavelengths that pass through the PCS. In this way it is possible to realize a directional emission of the bright blue-green color, which enhances signaling and is important in communication between different butterfly individuals.

10.3.4
Super Anti-Reflection Structures of Lepidopteran Eyes and Wings

Although most photonic structures in Nature are associated with bright colors or broad angle reflectivity, a specific type of nanostructure can minimize the reflectivity over broad angles or frequency ranges. In Nature, this effect is normally observed in the ommateum of moths or other lepidoptera [60], and shows that a super anti-reflection property can improve the light sensitivity of light-craving moths and help them not to be detected by natural enemies at night. Hence, this is also termed the "moth-eye effect." The effect is commonly achieved by the incorporation of arrays of tapered elements, also described as nipple arrays (Figure 10.13a–c). When light is

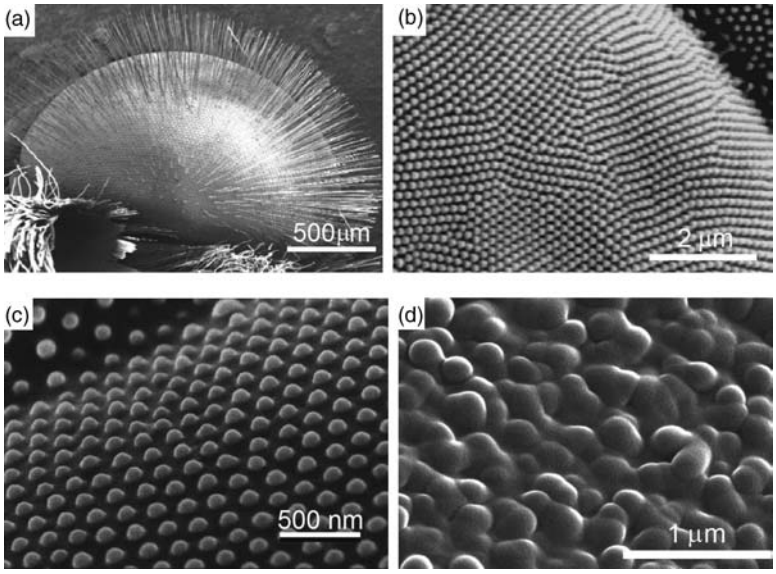


Figure 10.13 Scanning electron microscopy images of nipple arrays in the compound eye of butterflies and on the wings of a dragonfly. (a) The whole compound eye of butterfly *Inachis io*; (b) The detailed nanostructure in one facet lens of the compound eye; (c) Nipple array structure in compound eyes of another type of butterfly, *Polygonia c-aureum*; (d) Nipple array structure on wings of the dragonfly *Aeshna cyanea*. Panels (a–c) adapted from Ref. [60]; Panel (d) adapted from Ref. [61], with permission.

projected onto a transparent medium, the sharp change in refractive index at two sides of the interface will result in a partial reflection of the light. However, the feature size of the nipple arrays, and the distance between neighboring nipples on the ommatidial surface (which are normally within 250 nm) are less than the wavelength of the visible light. This induces a continuous change for the apparent refractive index along the depth direction, such that the structure will gradually match the optical impedance of one medium with its neighbor, and this can cause a significant reduction in reflection at the interface.

According to different materials and usages, the moth-eye effect can be used to achieve different superior properties. On an opaque material surface, the structure may result in a super-black color or cause a notable increase in the adsorption of light. This effect has already found broad applications in solar energy utilization and other fields, such as solar cells and solar water heaters.

Another important property related to such structure is the ultra-transparency that has been widely observed on the wings of many insects, including cicadas (see above; Figure 10.3a) and dragonflies [61]. The nipple array structure has also been identified on the wings of these insects; the well-aligned nanonipple arrays on a dragonfly's wings are shown in Figure 10.13d. These structures are composed of wax, which may help to achieve super water-repellent properties while efficiently preventing the wings from being wetted, when combined with the nanostructure (see Section 10.2).

10.3.5
Unusual Bright Whiteness in Ultrathin Beetle Scales

Whilst many things in life are white – snow, milk, paper, and so on – none of these can be compared to the whiteness of the *Cyphochilus* beetle (Figure 10.14a), a genus of beetle with an unusually brilliant white body that is found in Southeast Asia. According to Vukusic *et al.* [62], the exceptional whiteness and brilliance of the *Cyphochilus* beetle's body was not augmented by either pigments or fluorescence, but rather resulted from a 3-D photonic solid in the scales.

The scales (Figure 10.14b) that imbricate the beetle's body are about 5 μm thick, 250 μm, and 100 μm wide, and their interiors are composed of a random network of interconnecting cuticular filaments with diameters of about 250 nm (Figure 10.14c and d). However, unlike conventional photonic crystal structures there is no well-defined periodicity of the nanostructure. According to Vukusic *et al.*, the 5 μm-thick scales can provide standard whiteness and brightness values of 60 and 65, respectively. In synthetic systems where whiteness is desirable, a far more substantial structure is necessary. For example, an ultrawhite paper to which optical brightening agents have been added can reach similar brightness and whiteness for a thickness which is about 25-fold greater than the beetle scales. Detailed studies have indicated that the cuticular filament network is the origin of the extraordinary whiteness and brightness properties, in a thickness as low as 5 μm. On the one hand, an intrascale cuticle occupation rate of about 70% can optimize scattering intensity by maximizing the scattering center number, but on the other hand the aperiodicity will efficiently assure that scattering occurs over the whole wavelength range.

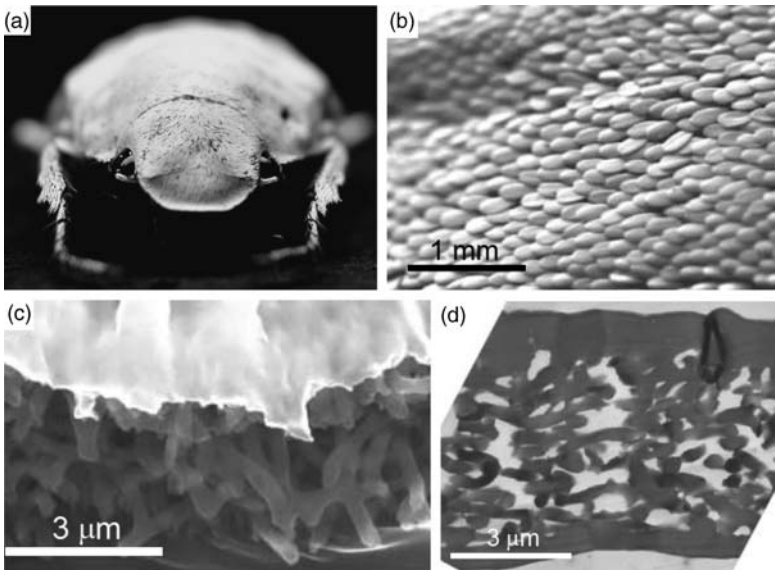


Figure 10.14 *Cyphochilus* beetle (a) with super whiteness and the nanostructures of its scales; (b) Large-scale SEM image of the scales; (c) Magnified SEM image of the scale nanostructure; (d) TEM image of the nanostructure. Adapted from Ref. [62], with permission.

10.3.6
Photosensitivity in Brittlestar (*Ophiuroidea*)

Photosensitivity is normally considered to be a chemical process that is induced by specific chemical-based photoreceptors. This phenomenon is frequently observed in Nature – chameleons and frogs can change their skin colors according to the color and light of their environment. However, in some cases – and especially in the case of echinoderm animals – the special arrangement of microstructures and nanostructures of the skeleton may also act as a component of the specialized photosensory organs, conceivably with the function of compound eyes.

Echinoderms (starfish), especially the brittlestars (*Ophiuroidea*) [63], generally exhibit a wide range of responses to light intensity. Whilst some show almost no response to environmental light (e.g., *Ophiocoma pumila*), others – such as *Ophiocoma wendtii* (Figure 10.15a) – can change their colors markedly when the light intensity of their environment changes. Another interesting behavior of *O. wendtii* is that it can detect shadows and rapidly escape from predators; this is unexpected for such animals because they have no photosensory organs such as eyes. This sensitivity to light appears to be contributed by to the specialized skeletal structure of the dorsal arm plates, which protect the upper part of each joint in brittlestar arms (Figure 10.15b). Recent SEM analyses have disclosed elaborate regular arrays of the spherical microstructures (Figure 10.15c and d) for the skeletal structure of the dorsal arm plates. The skeletal elements of echinoderms are each composed of a single

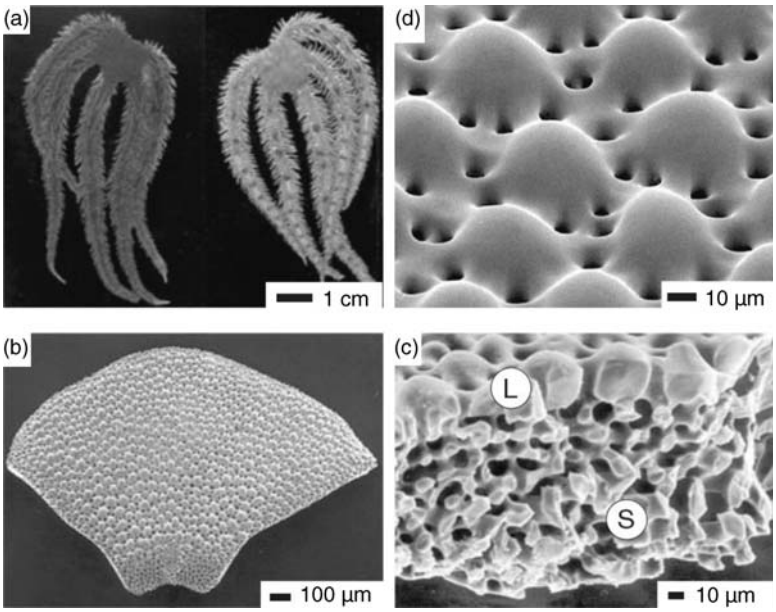


Figure 10.15 Photosensitive brittlestar *Ophiocoma wendtii* (a) and the microstructures of its bones; (b) A dorsal arm plate (DAP) of *O. wendtii*; (c) Cross-section of a fractured DAP, showing the typical calcitic stereom (S) and the enlarged lens structures (L); (d) Peripheral layer of a DAP enlarged lens structures. Adapted from Ref. [63] with permission.

crystal of oriented calcite shaped into a unique, 3-D mesh, although in the case of *O. wendtii* the structure has a remarkably regular double-lens design. Aizenberg *et al.* have shown that such structures can guide and focus the light inside the tissue, and this coincides with the location of the nerve bundles, which act as the primary photoreceptors. A special design of the lens array was also found to minimize spherical aberration and birefringence, and to allow the detection of light from a specific direction. These structures represent examples of biomaterials that perform simultaneous mechanical and optical functions, and may shed light on the design of multifunctional artificial materials.

10.4
The High Adhesive Force of Gecko Foot-Hairs

In biological systems, microstructures and nanostructures also contribute to superior mechanical and dynamic properties. Typical examples include porous structures in bones, wood, and pith, which bestow the materials with maximum strength at the lowest density [64–66]. In tooth materials [67], the dense arrangement of the nanostructures that is combined intimately with proteins provides sufficient strength and toughness at the same time. For surface materials, the nanostructure brings some unique properties that differ from those in the bulk materials.

The gecko is capable of climbing rapidly and freely along vertical walls, or even on the ceiling. Since this phenomenon was first noted almost 100 years ago, much effort has been expended to determine the origin of the high adhesive force between the gecko foot and the underlying surface. Experimental analyses [68, 69] have indicated that each gecko foot (Figure 10.16a) has about 5000 setae (Figure 10.16b) per mm^2, and can produce 10 N with approximately 100 mm^2 pad area; in other words, each seta should produce an average force of about 20 μN, and an average stress of 0.1 N mm^2 (1 atm). However, this force might be greatly underestimated as only a small number of setae would contact the surface simultaneously.

Full *et al.* [70] measured the exact force of a single seta using a micro-electro-mechanical system (MEMS) cantilever that attached with the seta. This showed that one seta could provide a force that, when parallel to the surface, might be as large

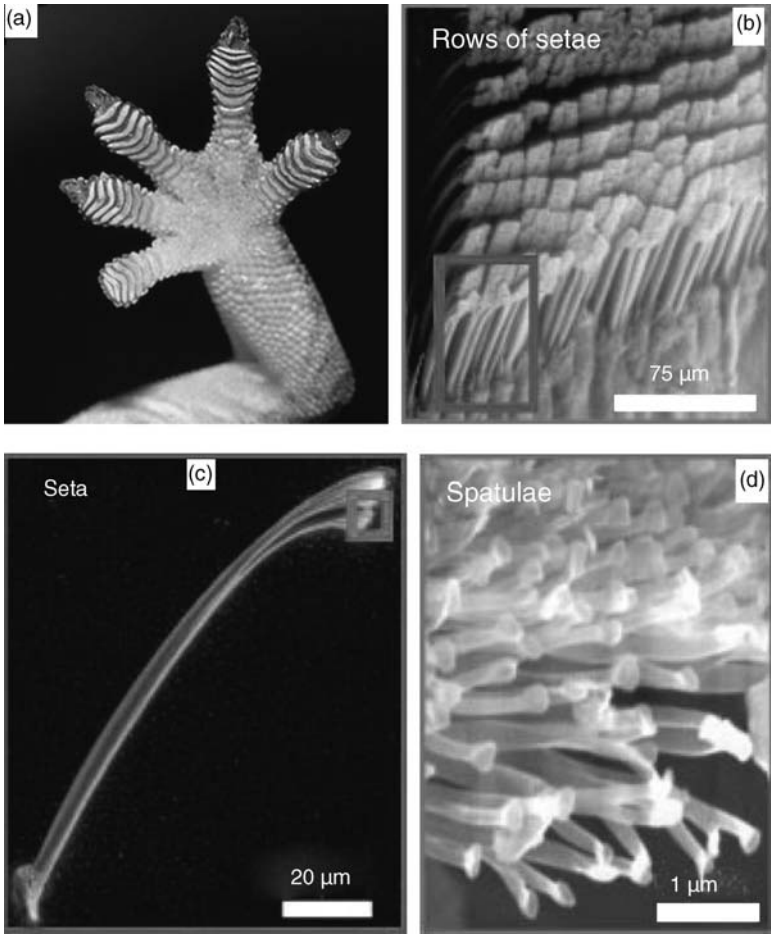


Figure 10.16 Nanostructures on a high-adhesive gecko's foot (a); (b) Rows of setae on the foot; (c) A single seta; (c) Nanospatulae structure of a single seta. Adapted from Ref. [70], with permission.

as $194 \pm 25\,\mu N$ – almost tenfold greater than the estimated value. This indicates that, if all the setae were to be attached simultaneously to the surface, a single gecko foot could provide a 100 N adhesive force (10 atm), which is several hundred-fold the gecko's own body weight. Subsequent SEM studies showed that the terminal of each seta (Figure 10.16c) was further composed of smaller branches (Figure 10.16d) termed *spatulae*, each about 200 nm in size. Results reported by Full *et al*. revealed that the intermolecular forces (e.g., van der Waals forces, etc.) between the spatulae and the surface were the origin of the high adhesive force. Whilst the gecko's feet contain about one billion spatulae, providing a large contact area with the surface, the spatulae are soft and thin and can deform easily so as to fit the complicated local surface topology, thus guaranteeing sufficient surface contact.

More interestingly, the gecko's feet possess good anti-adhesive properties to dust [71], while simultaneously exhibiting a high adhesive force to the substrate. In other words, as the gecko walks about, its sticky feet will always be clean and free from dust contamination, but will not require grooming as they will retain their stickiness for months. In fact, geckos with dirty feet have been shown to recover their ability to cling to vertical surfaces after only a few steps. By using an array of setae isolated from geckos, it was possible to demonstrate the self-cleaning process and to show, using contact mechanical models, that the self-cleaning occurred via an energetic disequilibrium between the adhesive forces that attract dirt to the substrate and those that attract the same particle to the setae. In this process, the setal nanostructure plays a crucial role.

Similar to the gecko, spiders and some beetles can also creep in inverted fashion along almost any type of surface. The study of Kesel *et al*. [72] showed the feet of spiders to be covered with setae nanostructures, very similar to the gecko. Figure 10.17a–c are SEM images of the setae structure and the further densely packed setule structures on each seta, at different magnifications. According to Kesel's experiments and calculations, each nanosetule can provide an adhesive force in excess of 40 nN, with all of the setae on a spider's feet providing a total adhesive force of about 170-fold its own body weight. A similar nanostructure was observed in the case of the beetle *Hemisphaerota cynea* [73], as shown in Figure 10.17d–f.

On the basis of this recognition, much effort has been made to mimic the gecko foot artificially, by using elastic polymers or other materials [74, 75]. In a typical study conducted at the University of Dayton and Georgia Institute of Technology [76], ACNT arrays were used (Figure 10.18b and c) to mimic the binding-on and lifting-off behaviors of gecko foot setae as the gecko walked. The devices fabricated could provide adhesive forces (Figure 10.18a) of about 100 N per cm^2, which was about 10-fold that of the gecko foot, and almost equal to the theoretical value that all setae attach close to the surface. The shear adhesive force was shown to be much stronger than the normal adhesive direction, which ensured a strong binding along the shear direction. This effect was found to be caused by a shear-induced alignment of the nonaligned nanotube top layer, which caused a dramatic enhancement of the line contact with the surface, similar to the operating function of the gecko setae.

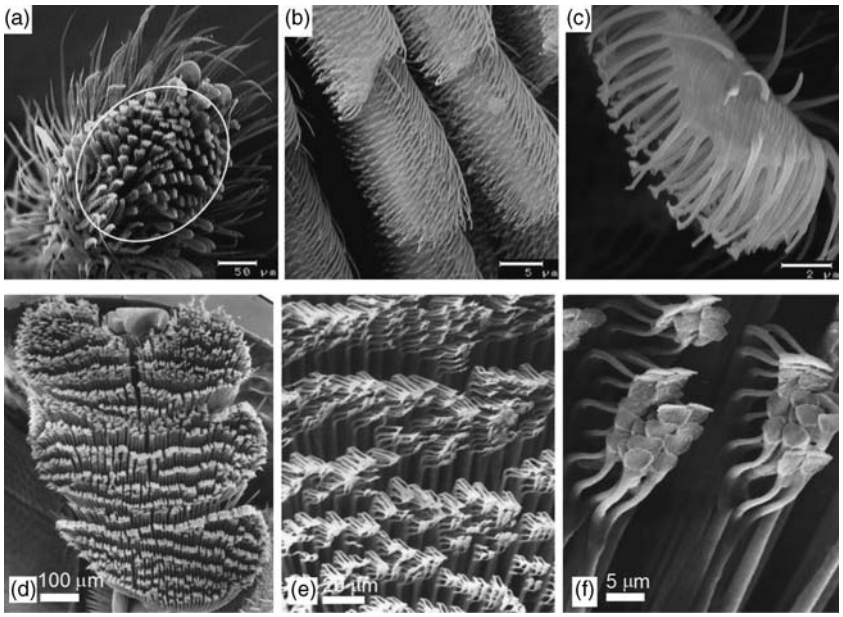


Figure 10.17 Microsetae structure and further nanosetule structure on the high-adhesive feet of a spider (a–c) and a beetle *Hemisphaerota cynea* (d–f), with different magnifications (see scale bars). Adapted from Refs [72] and [73], with permission.

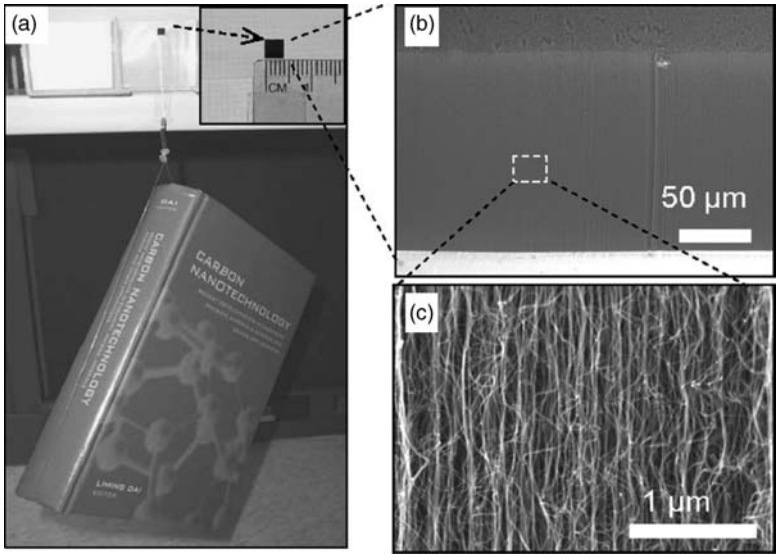


Figure 10.18 High-adhesive carbon nanotube film. (a) A book of 1480 g in weight suspended from a glass surface with use of ACNT supported on a silicon wafer; (b, c) Scanning electron microscopy images of ACNT arrays at different magnifications (see scale bars). Adapted from Ref. [76], with permission.

10.5
Summary and Outlook

During billions of years of evolution, biosystems in Nature have developed diverse and elegant microstructures and nanostructures on their surfaces that play important roles in the special functions and properties of organisms, including superhydrophobicity, anti-contamination, optical properties, and dynamic and mechanical properties. Yet, learning from Nature provides much inspiration to combine these structural effects with conventional artificial materials, so as produce superior properties, the essence of which is the cooperative effect on the microscale and nanoscale. Indeed, in recent years the deep and broad investigation of this effect has become one of the most important aspects of biomimetics, and will surely provide tremendous insights into the design of novel artificial materials and devices for use in a wide variety of domains that includes industry, medicine, agriculture, and general lifestyle.

References

1 Sun, T., Feng, L., Gao, X., and Jiang, L. (2005) *Acc. Chem. Res.*, **38**, 644.
2 Wang, R., Hashimoto, K., Fujishima, A., Chikuni, M., Kojima, E., Kitamura, A., Shimohigoshi, M., and Watanabe, T. (1997) *Nature*, **388**, 431.
3 Chen, W., Fadeev, A.Y., Hsieh, M.C., Öner, D., Youngblood, J., and McCarthy, T.J. (1999) *Langmuir*, **15**, 3395.
4 Feng, L., Li, S., Li, Y., Li, H., Zhang, L., Zhai, J., Song, Y., Liu, B., Jiang, J., and Zhu, D. (2002) *Adv. Mater.*, **14**, 1857.
5 Blossey, R. (2003) *Nature Mater.*, **2**, 301.
6 Lafuma, A. and Quéré, D. (2003) *Nature Mater.*, **2**, 457.
7 Nakajima, A., Fujishima, A., Hashimoto, K., and Watanabe, T. (1999) *Adv. Mater.*, **11**, 1365.
8 Woodward, J.T., Gwin, H., and Schwartz, D.K. (2000) *Langmuir*, **16**, 2957.
9 Sun, T., Song, W., and Jiang, L. (2005) *Chem. Commun.*, 1723.
10 Nishino, T., Meguro, M., Nakamae, K., Matsushita, M., and Ueda, Y. (1999) *Langmuir*, **15**, 4321.
11 Wenzel, R.N. (1936) *Ind. Eng. Chem.*, **28**, 988.
12 Cassie, A.B.D. and Baxter, S. (1944) *Trans. Faraday Soc.*, **40**, 546.
13 Barthlott, W. and Neinhuis, C. (1997) *Planta*, **202**, 1.
14 Adamson, A.W. and Gast, A.P. (1997) *Physical Chemistry of Surfaces*, John Wiley & Sons, New York.
15 Mandelbrot, B.B. (1982) *The Fractal Geometry of Nature*, Freeman, San Francisco, CA.
16 Feng, L., Li, S., Li, Y., Li, H., Zhang, L., Zhai, J., Song, Y., Liu, B., Jiang, L., and Zhu, D. (2002) *Adv. Mater.*, **14**, 1857.
17 Li, H., Wang, X., Song, Y., Liu, Y., Li, Q., Jiang, L., and Zhu, D. (2001) *Angew. Chem., Int. Ed.*, **40**, 1743.
18 Liu, H., Li, S., Zhai, J., Li, H., Zheng, Q., Jiang, L., and Zhu, D. (2004) *Angew. Chem., Int. Ed.*, **43**, 1146.
19 Erbil, H.Y., Demirel, A.L., Avci, Y., and Mert, O. (2003) *Science*, **299**, 1377.
20 Li, S., Li, H., Wang, X., Song, Y., Liu, Y., Jiang, L., and Zhu, D. (2002) *J. Phys. Chem. B*, **106**, 9274.
21 Li, H., Wang, X., Song, Y., Liu, Y., Li, Q., Jiang, L., and Zhu, D. (2001) *Angew. Chem., Int. Ed.*, **40**, 1743.
22 Marmur, A. (2004) *Langmuir*, **20**, 3517.
23 Feng, L., Li, S., Li, H., Zhai, J., Song, Y., Jiang, L., and Zhu, D. (2002) *Angew. Chem., Int. Ed.*, **41**, 1221.
24 Jiang, L., Zhao, Y., and Zhai, J. (2004) *Angew. Chem., Int. Ed.*, **43**, 4338.
25 Sun, T., Wang, G., Feng, L., Liu, B., Ma, Y., Jiang, L., and Zhu, D. (2004) *Angew. Chem., Int. Ed.*, **43**, 357.

26 Feng, X., Feng, L., Zhai, J., Jiang, L., and Zhu, D. (2004) *J. Am. Chem. Soc.*, **126**, 62.

27 Liu, H., Feng, L., Zhai, J., Jiang, L., and Zhu, D. (2004) *Langmuir*, **20**, 5659.

28 Minko, S., Müller, M., Motornov, M., Nitschke, M., Grundke, K., and Stamm, M. (2003) *J. Am. Chem. Soc.*, **125**, 3896.

29 Guo, C., Feng, L., Zhai, J., Wang, G., Song, Y., Jiang, L., and Zhu, D. (2004) *ChemPhysChem*, **5**, 750.

30 Zhang, G., Zhang, J., Xie, G., Liu, Z., and Shao, H. (2006) *Small*, **2**, 1440.

31 Stavega, D.G., Foletti, S., Palasantzas, G., and Arikawa, K. (2006) *Proc. R. Soc. B*, **273**, 661.

32 Gao, X., Yan, X., Yao, X., Xu, L., Zhang, K., Zhang, J., Yang, B., and Jiang, L. (2007) *Adv. Mater.*, **19**, 2213.

33 Gu, Z., Uetsuka, H., Takahashi, K., Nakajima, R., Onishi, H., Fujishima, A., and Sato, O. (2003) *Angew. Chem., Int. Ed.*, **42**, 894.

34 Kennedy, R.J. (1970) *Nature*, **227**, 736.

35 Zheng, Y., Gao, X., and Jiang, L. (2007) *Soft Matter*, **3**, 178.

36 Keller, J.B. (1998) *Phys. Fluids*, **10**, 3009.

37 Sun, S.M. and Keller, J.B. (2001) *Phys. Fluids*, **13**, 2146.

38 Hu, D.L., Chan, B., and Bush, J.W.M. (2003) *Nature*, **424**, 663.

39 Gao, X. and Jiang, L. (2004) *Nature*, **432**, 36.

40 Parker, A.W. and Lawrence, C.R. (2001) *Nature*, **414**, 33.

41 Zhai, L., Berg, M.C., Cebeci, F., Kim, Y., Milwid, J.M., Rubner, M.F., and Cohen, R.E. (2006) *Nano Lett.*, **6**, 1213.

42 Kinoshita, S., Yoshioka, S., and Miyazaki, J. (2008) *Rep. Prog. Phys.*, **71**, 076401.

43 Rüdiger, W. and Thümmler, F. (1991) *Angew. Chem., Int. Ed.*, **30**, 1216.

44 Britton, G. (1995) *FASEB J.*, **9**, 1551.

45 Vukusic, P. and Sambles, J.R. (2003) *Nature*, **424**, 852.

46 Istrate, E. and Sargent, E.H. (2006) *Rev. Modern Phys.*, **78**, 455.

47 Fritsch, E., Gaillou, E., Rondeau, B., Barreau, A., Albertini, D., and Ostroumov, M. (2006) *J. Non-Cryst. Solids*, **352**, 3957.

48 Blanco, A., Chomski, E., Grabtchak, S., Ibisate, M., John, S., Leonard, S.W., Lopez, C., Meseguer, F., Migues, H., Mondia, J.P., Ozin, G.A., Toader, O., and van Driel, H.M. (2000) *Nature*, **405**, 437.

49 Parker, A.R., Welch, V.L., Driver, D., and Martini, N. (2003) *Nature*, **426**, 786.

50 Parker, A.R., McPhedran, R.C., McKenzie, D.R., Botten, L.C., and Nicorovici, N.-A.P. (2001) *Nature*, **409**, 36.

51 Newton, I. (1730) *Optics*, 4th edn, reprinted by Dover Publications, New York, p. 252.

52 Zi, J., Yu, X., Li, Y., Hu, X., Xu, C., Wang, X., Liu, X., and Fu, R. (2003) *Proc. Natl Acad. Sci. USA*, **100**, 12576.

53 Srnivasarao, M. (1999) *Chem. Rev.*, **99**, 1935.

54 Vukusic, P. (2006) *Curr. Biol.*, **16**, R621.

55 Vukusic, P., Sambles, J.R., and Lawrence, C.R. (2000) *Nature*, **404**, 457.

56 Vukusic, P., Sambles, J.R., Lawrence, C.R., and Wootton, R.J. (2001) *Nature*, **410**, 36.

57 Plattner, L. (2004) *J. R. Soc. Interface*, **1**, 49.

58 Kemp, D.J., Vukusic, P., and Rutowski, R.L. (2006) *Funct. Ecol.*, **20**, 282.

59 Vukusic, P. and Hooper, I. (2005) *Science*, **310**, 1151.

60 Stavenga, D.G., Foletti, S., Palasantzas, G., and Arikawa, K. (2006) *Proc. R. Soc. B*, **273**, 661.

61 Hooper, I.R., Vukusic, P., and Wootton, R.J. (2006) *Opt. Express*, **14**, 4891.

62 Vukusic, P., Hallam, B., and Noyes, J. (2007) *Science*, **315**, 348.

63 Aizenberg, J., Tkachenko, A., Weiner, S., Addadi, L., and Hendler, G. (2001) *Nature*, **412**, 819.

64 Gibson, L.J. and Ashby, M.F. (1982) Mechanics of 3-dimensional cellular materials. *Proc. R. Soc. London A*, **382**, 43.

65 Fan, H., Hartshorn, C., Buchheit, T., Tallant, D., Assink, R., Simpson, R., Kissel, D.J., Lacks, D.J., Torquato, S., and Brinker, C.J. (2007) *Nat. Mater.*, **6**, 418.

66 Tai, K., Dao, M., Suresh, S., Palazoglu, A., and Ortiz, C. (2007) *Nat. Mater.*, **6**, 454.

67 Meyers, M.A., Chen, P.-Y., Lin, A.Y.-M., and Seki, Y. (2008) *Prog. Mater. Sci.*, **53**, 1.

68 Ruibal, R. and Ernst, V. (1965) *J. Morphol.*, **117**, 271.

69 Irschick, D.J., Austin, C.C., Petren, K., Fisher, R.N., Losos, J.B., and Ellers, O. (1996) *J. Linnaen Soc.*, **59**, 21.

70 Autumn, K., Liang, Y.A., Hsieh, S.T., Zesch, W., Chan, W.P., Kenny, T.W., Fearing, R., and Full, R.J. (2000) *Nature*, **405**, 681.
71 Hansen, W.R. and Autumn, K. (2005) *Proc. Natl Acad. Sci. USA*, **102**, 385.
72 Kesel, A.B., Martin, A., and Seidl, T. (2004) *Smart Mater. Struct.*, **13**, 512.
73 Eisner, T. and Aneshansley, D.J. (2000) *Proc. Natl Acad. Sci. USA*, **97**, 6568.
74 Sitti, M., Fearing, R.S., and Adhes, J. (2003) *Sci. Technol.*, **17**, 1055.
75 Geim, A.K., Dubonos, S.V., Grigorieva, I.V., Novoselov, K.S., Zhukov, A.A., and Shapoval, S.Yu. (2003) *Nat. Mater.*, **2**, 461.
76 Qu, L., Dai, L., Stone, M., Xia, Z., and Wang, Z.-L. (2008) *Science*, **322**, 238.